UNIVERSITÉ DE GENÈVE

Archive ouverte UNIGE
https://archive-ouverte.unige.ch

Article scientifique | Article | 2024

Published version | Public access

This is the published version of the publication, made available in accordance with the publisher's policy.

DNA methylation enables recurrent endogenization of giant viruses in an animal relative

Sarre, Luke A; Kim, Iana V; Ovchinnikov, Vladimir; Olivetta, Marine; Suga, Hiroshi; Dudin, Omaya; Sebé-Pedrós, Arnau; de Mendoza, Alex

How to cite

SARRE, Luke A et al. DNA methylation enables recurrent endogenization of giant viruses in an animal relative. In: Science advances, 2024, vol. 10, n° 28, p. eado6406. doi: 10.1126/sciadv.ado6406

This publication URL: https://archive-ouverte.unige.ch/unige:179458
Publication DOI: 10.1126/sciadv.ado6406

© The author(s). This work is licensed under a Creative Commons Attribution (CC BY 4.0)
https://creativecommons.org/licenses/by/4.0

Last deposit update in Archive ouverte UNIGE on 27.08.2024 14:13

GENETICS

DNA methylation enables recurrent endogenization of giant viruses in an animal relative

Luke A. Sarre[1], Iana V. Kim[2], Vladimir Ovchinnikov[1], Marine Olivetta[3], Hiroshi Suga[4], Omaya Dudin[3], Arnau Sebé-Pedrós[2,5,6], Alex de Mendoza[1]*

Copyright © 2024 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution License 4.0 (CC BY).

5-Methylcytosine (5mC) is a widespread silencing mechanism that controls genomic parasites. In eukaryotes, 5mC has gained complex roles in gene regulation beyond parasite control, yet 5mC has also been lost in many lineages. The causes for 5mC retention and its genomic consequences are still poorly understood. Here, we show that the protist closely related to animals *Amoebidium appalachense* features both transposon and gene body methylation, a pattern reminiscent of invertebrates and plants. Unexpectedly, hypermethylated genomic regions in *Amoebidium* derive from viral insertions, including hundreds of endogenized giant viruses, contributing 14% of the proteome. Using a combination of inhibitors and genomic assays, we demonstrate that 5mC silences these giant virus insertions. Moreover, alternative *Amoebidium* isolates show polymorphic giant virus insertions, highlighting a dynamic process of infection, endogenization, and purging. Our results indicate that 5mC is critical for the controlled coexistence of newly acquired viral DNA into eukaryotic genomes, making *Amoebidium* a unique model to understand the hybrid origins of eukaryotic DNA.

INTRODUCTION

5-Methylcytosine (5mC) is a common base modification among eukaryotes (*1*–*3*). 5mC is deposited by DNA methyltransferases (DNMTs), a family of enzymes with ancestral families conserved throughout eukaryotes (*4*, *5*). Some DNMTs are maintenance type enzymes, perpetuating 5mC patterns, including DNMT1 and DNMT5, while other DNMTs have de novo activity, such as DNMT3 (*6*, *7*). However, the DNMT repertoire of an organism is not predictive of 5mC function. In some eukaryotes, including plants and animals, 5mC is associated with gene regulation, exemplified by gene body methylation, where 5mC positively correlates with gene transcriptional levels (*1*, *3*, *8*). However, the most widespread role of 5mC is in transposable element (TE) silencing, which is the assumed ancestral role in eukaryotes (*9*, *10*).

Despite most attention being devoted to controlling endogenous parasitic elements, one of the first described functions of 5mC in eukaryotes was to silence retroviral insertions in mammals (*11*). Similarly, in bacteria, the main role of 5mC is to combat viruses (*12*). Therefore, controlling exogenous viral invasions is arguably as important as TE control for epigenetic silencing. It is increasingly recognized that many eukaryotic genes have viral origins, co-opted repeatedly throughout evolution (*13*). One of the most common sources for these acquisitions are giant viruses (*Nucleocytoviricota*). Giant viruses have a wide range of eukaryotic hosts and are present in almost all ecosystems, posing a widespread threat to eukaryotic cells (*14*, *15*). Giant viruses are exceptional among viruses as they have enormous genomes (100 kb to 2.5 Mb) encoding many proteins thought to be eukaryotic hallmarks such as histones (*14*, *15*). Giant viruses originated before modern eukaryotes, and they have been proposed to have contributed essential genes to eukaryogenesis (*16*–*18*). Furthermore, recent reports indicate that giant viruses can endogenize into extant eukaryotes (*19*–*21*). However, how this potentially lethal DNA is incorporated into eukaryotic genomes is currently not understood.

Finding a link between viral control and epigenetic regulation, however, is hampered by the scarcity of reported recent giant virus endogenizations (*19*, *22*). Moreover, 5mC is evolutionarily very plastic, and many eukaryotic lineages have lost this epigenetic modification (*1*, *2*), possibly because of its mutagenic potential and cytotoxic off-target effects of DNMTs (*23*). Furthermore, 5mC function varies across lineages. In fungi, 5mC is restricted to silencing TEs (*24*), whereas in invertebrates 5mC is usually restricted to gene bodies, and most TEs remain unmethylated (*1*, *2*, *25*). To expand our knowledge of 5mC systems and to unravel how a potentially ancestral fungal-like methylation pattern gave rise to the animal 5mC system, we focused on protists of the holozoan clade. These close animal relatives form four major lineages: choanoflagellates, filastereans, ichthyosporeans, and pluriformeans (Fig. 1A) (*26*, *27*). In recent years, unicellular holozoan genomes have been shown to encode many genes previously thought to be unique to animals, informing the complex genomic nature of the unicellular ancestors of animals (*26*–*29*). However, none of these genomes encode DNMTs, suggesting an evolutionary loss of 5mC capacity (*30*). Here, we fill this gap by describing a unicellular relative of animals that has maintained 5mC, and unexpectedly find an unappreciated and potentially ancestral use of 5mC in regulating giant virus endogenizations.

RESULTS

The *Amoebidium* genome presents both gene body and TE methylation

To reconstruct the pre-animal roots of 5mC, we searched the available genomes and transcriptomes of unicellular holozoans for DNMT1 orthologs (*29*, *31*), the maintenance DNMT in animals. We discovered that DNMT1 is expressed by *Amoebidium appalachense* (Fig. 1A), an ichthysoporean originally isolated from the cuticle of freshwater

[1]School of Biological and Behavioural Sciences, Queen Mary University of London, London, UK. [2]CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. [3]Swiss Institute for Experimental Cancer Research, School of Life Sciences, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland. [4]Faculty of Life and Environmental Sciences, Prefectural University of Hiroshima, Shobara, Japan. [5]ICREA, Barcelona, Spain. [6]Universitat Pompeu Fabra (UPF), Barcelona, Spain.
*Corresponding author. Email: a.demendozasoler@qmul.ac.uk

Downloaded from https://www.science.org at Universite de Geneve on August 27, 2024

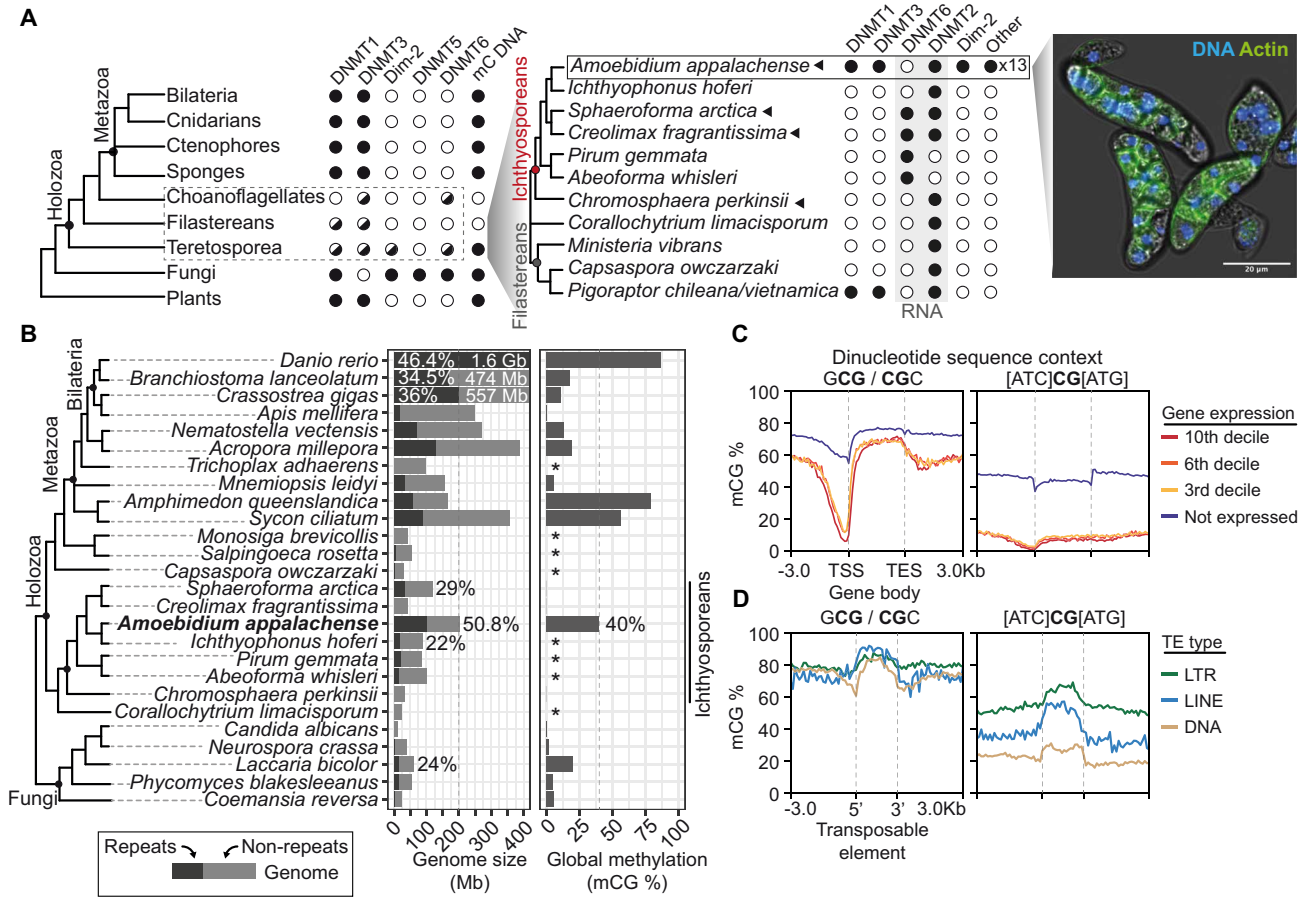

Fig. 1. *A. appalachense* displays a methylome with gene body and transposon methylation. (**A**) Distribution of DNMTs in eukaryotes and teretosporeans. Black dot indicates enzyme presence, white dot indicates enzyme absence, and black triangles indicate species for which Enzymatic Methyl-seq has been performed. *Amoebidium* cells are stained with phalloidin and Hoechst. Phylogenetic relationships based on previous studies (*29*). (**B**) Distribution of genome sizes, repetitive content, and global methylation levels across opisthokonts. Asterisks indicate species for which the lack of 5mC is inferred from their absence of DNMTs. Genome sizes for species with genomes above 400 Mb are displayed in white, and repeat content (%) is also highlighted for clarity. (**C**) Gene body methylation and (**D**) TE methylation averages split by the extended CG sequence context. Not expressed genes display <1 TPM. TSS, transcriptional start site; TES, transcriptional end site. TEs only include copies spanning at least 70% of the RepeatModeler2 consensus model.

arthropods that can be grown axenically (*32*). Like other Ichthyosporea (*33*), *Amoebidium* are parasites or commensals of animal hosts, and show a coenocytic life cycle, with cells dividing their nuclei in a shared cytoplasm until forming a mature colony, which gives rise to unicellular and uninucleated cells (*34*). Yet, *Amoebidium* branches quite deep within the ichthyosporean lineage (Fig. 1A). *Amoebidium* DNMT1 sequence presents the same domain architecture of animal DNMT1 orthologs, including a zinc finger CXXC absent in non-holozoan sequences (fig. S1A), which suggests that this domain architecture was an innovation of holozoans despite DNMT1 being lost repeatedly. Additionally, we found DNMT1 orthologs in the metagenome-derived assemblies of the filastereans of the genus *Pigoraptor* (*35*), with the same domain architectures as *Amoebidium* (fig. S1A). Additionally, both *Amoebidium* and *Pigoraptor* encode UHRF1 orthologs with highly similar domain architectures to animal counterparts (fig. S1A), suggesting that DNMT1 and UHRF1 heterodimerization is conserved. Unlike *Amoebidium*, *Pigoraptor* species can only be grown at low densities in association with diverse bacteria and eukaryotic prey, the kinetoplastid *Parabodo caudatus*, making functional approaches and reliable methylation profiling challenging (*36*).

To fully characterize the gene repertoire and investigate *Amoebidium* 5mC patterns, we sequenced the genome of this species using a combination of nanopore long reads, Illumina short reads, and micro-C. We obtained a chromosome-scale assembly spanning 202 Mb, with 96.6% of the sequence contained in 18 chromosomes and a 95% BUSCO score (Fig. 1B and fig. S2, A and B). When examining *Amoebidium*'s DNMT repertoire, we identified a total of 18 DNMTs, comprising orthologs of DNMT1, DNMT3, DNMT2, Dim-2, and various lineage-specific clades (Fig. 1A and fig. S1, B and C). Notably, unlike DNMT1, none of the other DNMTs exhibit additional domains. Specifically, DNMT3 lacks the protein domains found in animal DNMT3, such as PWWP or ADD (fig. S1A) (*7*, *30*). In contrast, none of the ichthyosporeans with genomic data available show DNMTs other than the RNA-specific DNMT2 and DNMT6 (Fig. 1A) (*3*, *7*, *37*). Among other holozoans, we could only find a putative DNMT3 in the transcriptome of the choanoflagellate *Achanthoeca spectabilis* (*31*) and both

Downloaded from https://www.science.org at Universite de Geneve on August 27, 2024

Pigoraptor species (*35*), with the same domain architecture lacking additional protein domains as in *Amoebidium* (fig. S1A). Thus, *Amoebidium* and *Pigoraptor* are the only sequenced unicellular holozoan species that retain the ancestral eukaryotic complement of DNMTs, highlighting the pervasive tendency of eukaryotes to lose 5mC.

Next, we performed whole-genome DNA methylation profiling to analyze the 5mC patterns in *Amoebidium*, as well as in three other ichthyosporean species lacking DNA DNMTs as negative controls. In *Amoebidium*, global methylation levels soar to 40%, exclusively within the CG dinucleotide context, setting it apart from most invertebrates and fungi and the other ichthyosporean species, which exhibit negligible levels of 5mC (Fig. 1B) (*1*, *24*). Notably, not all CG dinucleotides exhibit uniform methylation levels. Specifically, the symmetrical mCGC and GmCG trinucleotides stand out with hypermethylation levels at around 70%, whereas the remaining CG dinucleotides maintain lower levels at approximately 20% (fig. S3A). This suggests that *Amoebidium* boasts elevated methylation levels with a wider sequence specificity beyond the CG dinucleotide, a context-dependent regionalization of 5mC reminiscent of heterochromatin methylation in mammals (*38*), likely reflecting the sequence preferences of the diverse *Amoebidium* DNMTs.

Considering the high global methylation levels in *Amoebidium*, we proceeded to investigate which genomic regions exhibit enriched 5mC. Protein-coding genes displayed a gene body methylation pattern reminiscent of plants and animals, with relatively low levels of promoter methylation (Fig. 1C) (*39*, *40*). However, *Amoebidium*'s gene body methylation is not positively correlated with transcription as in plants or animals, as all active genes have similar methylation levels irrespectively of transcriptional level, whereas silent genes show higher methylation, including the promoter (Fig. 1C and fig. S3B). Therefore, gene body methylation appears not exclusive to animals in the holozoan clade, yet its positive association with transcription is an animal-specific feature potentially linked to the domain acquisitions of animal DNMT3s (fig. S1A).

In contrast to most invertebrates (*1*, *2*), *Amoebidium* exhibits targeted methylation of TEs (Fig. 1D and fig. S3, B and C). Notably, methylation levels are highest in recent TE insertions and on transcriptionally silent genes (Fig. 1D and fig. S3C), regardless of the adjacent CG sequence context. In contrast, gene body methylation of actively transcribed genes primarily occurs within the CGC/GCG trinucleotide context (Fig. 1C). This indicates that in *Amoebidium*, 5mC of CGs in non-CGC/GCG trinucleotide context correlates with silencing, whereas CGC/GCG methylation is widespread. Further supporting the link between 5mC and TE silencing, approximately 50% of *Amoebidium*'s genome is composed of TEs, a level unmatched in any unicellular holozoans, yet similar to vertebrates such as humans (50%) or zebrafish (Fig. 1B and fig. S2, C and D). Therefore, the genome of *Amoebidium* is possibly permissive to TE expansions because 5mC can silence these novel insertions by reducing their potential deleterious effects, similar to what has been proposed for vertebrates (*41*),

Large hypermethylated regions uncover hundreds of viral insertions

To characterize the chromosome-level distribution of 5mC, we took advantage of the relative depletion of non-CGC/GCG methylation to locate regions of hypermethylation across the genome. We found many islands of hypermethylation spread across the chromosomes (Fig. 2A), many of which were consistent with regions of high TE content. However, many presented highly gene-rich areas spanning up to 200 kb, with most genes showing few to no introns, in clear contrast to the intron-rich *Amoebidium* genes (average 7.2 exon/gene; fig. S4A). Further characterization of these areas revealed core giant virus genes, including poxvirus late transcription factor (VLTF3), A32-like packaging adenosine triphosphatase (ATPase), D5 DNA primase, or nucleocytoplasmic large DNA viruses (NCLDV) major capsid proteins (fig. S4A) (*14*, *42*). Using these core genes, we searched the National Center for Biotechnology Information (NCBI) database and performed phylogenetic analyses using curated databases of giant virus marker genes (*42*). We found that these insertions could be classified as belonging to a lineage belonging to the order pandoravirales, closely related to the *Mamonoviridae* family of Medusavirus and Clandestinovirus (infecting amoebozoans; fig. S5, A and B) (*43*). Yet, not all the giant endogenous viral elements (GEVEs) in the *Amoebidium* genome originate from a single catastrophic insertion event or even a single viral lineage, as they show high levels of sequence divergence among them (fig. S5C). Furthermore, there are insertions in almost all chromosomes (90 chromosomal insertions, with 42 sequences in unplaced contigs), some having accumulated secondary TE insertions (Fig. 2A). The disparity of insertion lengths, and the observation that none of them encode a full repertoire of core giant virus genes (Fig. 2B), suggests that complete viral genome integrations are rare or that gene loss occurs rapidly after insertion. Using ViralRecall to detect giant viruses solely based on sequence (*44*), we confirmed the presence of GEVEs in *Amoebidium*, yet we failed to recover any hits from other holozoans other than in *Pigoraptor*, which encodes for few giant virus markers, including a capsid protein or a VLTF3, yet these branch far from *Amoebidium* hits (fig. S5, A and B), suggesting that distant classes of giant viruses might endogenize into *Pigoraptor* genomes. However, the fragmented status and metagenomic source of *Pigoraptor* genome assemblies render confident GEVE identification problematic, as some might belong to viruses or other species found in the complex cultures.

In addition to the giant viruses, other compact hypermethylated regions in the *Amoebidium* genome were characterized by genes encoding VLTF3, Dam methyltransferase, minor and major capsid proteins, and a DNA polymerase family B (fig. S4B). DNA polymerase sequences produced closest matches to adintovirus (family *Adintoviridae*), a group of recently described double-stranded DNA polinton-related viruses thought to exclusively infect animals (fig. S5D) (*45*). It is worth noting that many polinton-related viruses are virophages or descendants of these (*46*, *47*), known to parasitize giant viruses, which could explain the abundance of these sequences in the *Amoebidium* genome. Similarly to the giant viruses, not all adintoviruses were closely related among each other (fig. S5E), suggesting multiple independent insertion events. In contrast to GEVEs, some insertions kept long terminal repeats and were complete (~30 kb; Fig. 2B), yet others were truncated and in the process of degeneration.

Then, we identified a third type of giant repeat in *Amoebidium*, consisting of tandem clusters of repetitive intron-poor genes up to 50 kb long, usually flanked by a Plavaka transposase (fig. S4C) (*48*). Many of these genes encode for tyrosine recombinases, one of the major type of transposon integrases in eukaryotes (*49*), and interestingly their only hits in the NCBI NR database belong to very distant eukaryotic lineages including dinoflagellates or red algae, thus

Downloaded from https://www.science.org at Universite de Geneve on August 27, 2024

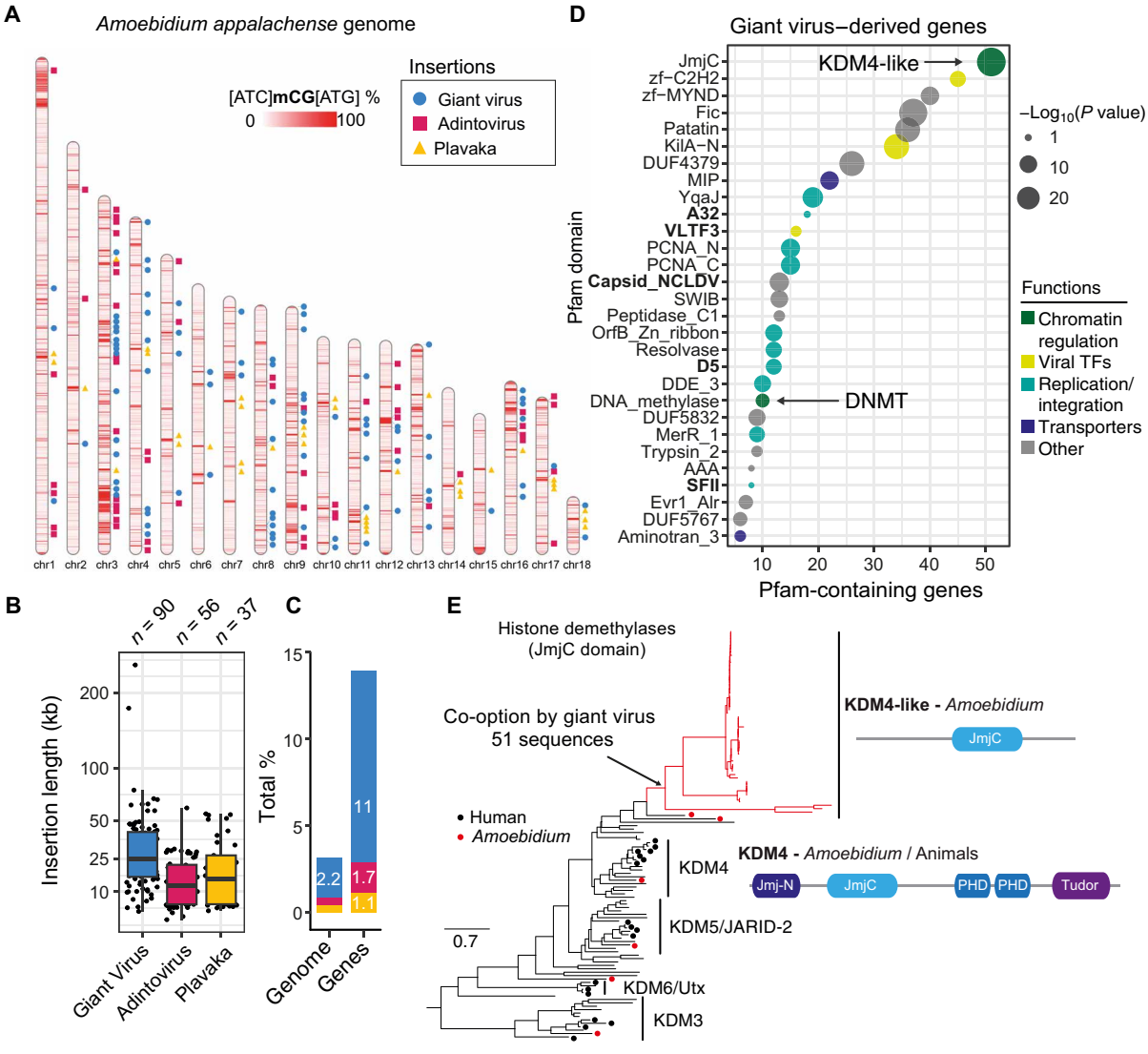

Fig. 2. The *Amoebidium* genome harbors hundreds of viral insertions. (**A**) Location of GEVEs, adintovirus, and Plavaka giant repeats across the *Amoebidium* genome. Windows of 10 kb are colored according to their methylation level in non-CGC/GCG trinucleotides. (**B**) Distribution of insertion sizes of the giant repeats within chromosomes. Center lines in boxplots are the median, box is the interquartile range (IQR), and whiskers are the first or third quartile ± 1.5× IQR. (**C**) Contribution of giant repeats to genome size and gene counts. (**D**) Pfam domains enriched in genes encoded in GEVE regions. In bold, marker giant virus domains. The displayed *P* values correspond to a two-sided Fisher exact test. (**E**) Maximum likelihood phylogeny of JmjC in eukaryotes, highlighting the expansion of KDM4-like enzymes in GEVE regions. Black dots indicate human sequences, red dots indicate *Amoebidium* sequences, and red branches indicate genes within endogenized viral regions. Domain architectures defined with PFAM domains.

suggesting some form of lateral gene transfer as their source (fig. S4C). When we combine the three types of highly methylated giant repeats, they make up 3.1% of *Amoebidium*'s total DNA. Their contribution to the protein-coding genes constitutes 14% of the entire proteome, with the majority originating from viruses. The amount of giant virus insertions in *Amoebidium* is among the largest reported in eukaryotes, at par with the moss *Physcomitrium patens* (Fig. 2C) (*50*).

Endogenized giant virus co-opted eukaryotic histone demethylases

To understand the potential contribution of endogenized genes to the *Amoebidium* gene repertoire, and also to better understand the gene complement of the original giant virus genomes that infect *Amoebidium*, we characterized the functional enrichment of genes encoded in these endogenized regions. An enrichment test of Pfam domains revealed many domain categories involved in the viral replication and integration process [recombinases, integrases, proliferating cell nuclear antigen (PCNA)], viral gene regulation (transcription factors), or some transporters [e.g., aquaporins/Major Intrinsic Protein (MIP)], which are likely critical to taking control of the host during infection (Fig. 2D) (*51*–*53*). Gene ontologies also suggested that these genes were enriched in membrane fission or tubulin depolymerization (fig. S6A). Notably, some of the most enriched categories were involved in chromatin regulation. Among these, 10 of the 18 DNMTs encoded in the *Amoebidium* genome

Downloaded from https://www.science.org at Universite de Geneve on August 27, 2024

reside in GEVEs, which suggests that these could be used by the virus to modify its own DNA. Consistently, giant viruses, and members of the pandoravirales in particular, are known to use various forms of DNA methylation (N^6-methyladenine and N^4-methylcytosines) to methylate their own genomes (*54*), which might play a role in infection. However, the *Amoebidium* GEVE DNMTs form a sister group to other giant virus uncharacterized DNMTs (fig. S1B); thus, they were not recently acquired from the host and their sequence-substrate preferences remain unknown.

The most enriched endogenized domain is the Jumonji C (JmjC) domain. Although JmjC domains can perform many enzymatic functions, our phylogenetic analysis revealed that these are divergent paralogs of the histone lysine demethylase subfamily 4 (KDM4). Notably, despite JmjC-containing proteins having been identified in giant viruses (*55*), we could not find any KDM4-like JmjC homologs in publicly available giant virus genomes. *Amoebidium* encodes a canonical KDM4 ortholog like those of other eukaryotes, including its characteristic histone-interacting domains (PHD, Tudor; Fig. 2E). However, the endogenized KDM4-like enzymes only contain the enzymatic JmjC domain (Fig. 2E). KDM4 enzymes are known to demethylate histone 3 tail lysines, most commonly lysine 9 (H3K9me2/3) or lysine 36 (H3K36me2/3) residues. Although many giant viruses encode all four eukaryotic nucleosome histones (H2A/B,H3,H4) (*42*, *56*), we did not find any in the viral insertions. Furthermore, viral histones present very divergent histone tails (*57*); thus, it is unlikely that KDM4-likes are used to control potential giant virus histones. Instead, given the conserved role of H3K9me3 in heterochromatin formation in eukaryotes, KDM4-like enzymes could be used by the virus to avoid silencing by the host chromatin. In KDM4-overexpressing cancer cells, depletion of H3K9me3 promotes DNA breaks and genome instability (*58*), a process that could serve the virus to integrate into the host genome, or explain the amount of endogenization events.

The KDM4-like enzymes stand out among the endogenized genes as they have preserved the multi-exon domain structure of eukaryotic genes (fig. S6C), unlike the vast majority of GEVE genes that lack introns. While most of the endogenized genes remain silent in culture conditions, four of these KDM4-like genes are transcribed [Transcripts Per Million (TPM) > 1] (fig. S6C). Moreover, JmjC genes are found in 39% (*52*) of the insertions, which could reflect a lower chance of purging those genes after the insertion event. A couple of KDM4-like genes are found outside hypermethylated giant virus regions and are flanked by normal host genes, showing almost exclusively mCGC/GmCG methylation (the default state for transcribed host genes; fig. S6C). Given their basal position in the phylogeny of KDM4-likes (Fig. 2E), these genes could be *Amoebidium*-specific KDM4 divergent paralogs that already lost the chromatin interaction domains compared to the canonical KDM4 copy, and were later co-opted by the giant viruses. Alternatively, the giant virus might have originally acquired a canonical KDM4 gene from the host, which then lost some of its companion domains to perform virus-associated functions. Then, these basally branching KDM4-like genes are the remnants of past GEVE insertions, where most other viral genes have been purged and only JmjC loci are kept, being domesticated to become part of the host repertoire. Thus, the intricate interaction between the host chromatin and the giant viruses is likely critical to explain the gene flow between the host and parasite.

DNA methylation removal is sufficient for viral transcriptional reactivation

Since dense 5mC demarcates the viral insertions and these are transcriptionally silent, we wanted to directly investigate the causal relationship between 5mC and gene expression in *Amoebidium*. We tested the effect of cytidine analogs 5-azacytidine, zebularine, and decitabine [which block DNMTs and lead to passive dilution of 5mC (*59*, *60*)] to investigate the impact of 5mC on gene expression. A 3-day cytidine analog treatment spans at least two generations of *Amoebidium* colonies, covering two rounds of coenocytic development starting from an uninucleate cell to colony maturation and cell release in ~30 hours (fig. S7A). Therefore, several rounds of nuclear division maximize the potential of obtaining sufficient passive 5mC loss. 5mCG remains constant across development, thus minimizing the potential confounding staging effects across treatments (fig. S7, B and C). We then used Enzymatic Methyl-seq to quantify 5mC of the treated cells and found that only 5-azacytidine showed a decrease in global methylation levels (from ~40 to 15%; Fig. 3A). Consistently, only the *Amoebidium* cells treated with 5-azacytidine showed growth defects and increased mortality (fig. S7D). However, 5-azacytidine can potentially be incorporated into RNA and be cytotoxic (*61*, *62*). To control for those off-target effects, we also treated two ichthyosporean species lacking genomic 5mC with 5-azacytidine, observing mild growth defects (fig. S7E).

We then used RNA sequencing (RNA-seq) to characterize the transcriptional response to 5-azacytidine in *Amoebidium* and *Sphaeroforma arctica*. *Sphaeroforma* is a closely related ichthyosporean that also has a relatively large amount of TEs and few instances of polinton-type viruses (*63*), yet lacks genomic 5mC (Fig. 1B). Both species showed hundreds of differentially expressed genes upon treatment (5630 in *Amoebidium* and 1807 in *Sphaeroforma*; false discovery rate < 0.01), but very few of these showed consistent dynamics across species (fig. S8A), thus not suggesting generic stress response shared across species. Nevertheless, genes that were up-regulated upon 5-azacytidine treatment in *Amoebidium* have stress-associated gene ontologies, while a wide range of metabolic processes are down-regulated (fig. S8D). As observed in 5-azacytidine–treated cancer cells, the stress response might be driven by TE reactivation (*64*). Focusing on TEs, only *Amoebidium* showed a drastic expression increase in almost all TE types after 5-azacytidine (fig. S8B), whereas *Sphaeroforma* did not show any particular enrichment in TE or viral up-regulation (most remaining transcriptionally silent/unchanged; fig. S8B), suggesting that the TE response to 5-azacytidine is a direct consequence of 5mC loss. We further validated this observation by dividing *Amoebidium* genes according to their promoter methylation level in untreated conditions. Genes that normally present unmethylated promoters had a mixed transcriptional response to DNA methylation removal suggestive of indirect effects, whereas genes with hypermethylated promoters were almost exclusively up-regulated upon 5-azacytidine treatment (Fig. 3B). Thus, 5mC is a silencing mark in *Amoebidium* sufficient to repress methylated genes.

When inspecting giant virus and adintovirus endogenized genes, we saw a consistent transcriptional reactivation upon methylation removal. Seven hundred thirty-seven genes encoded in GEVEs (26%) were transcriptionally reactivated (Fig. 3C), with the majority of them being the JmjC genes, but also many genes involved in gene regulation (fig. S8C). Similarly, 144 adintovirus genes were reactivated upon demethylation (32%). However, we did not observe

Downloaded from https://www.science.org at Universite de Geneve on August 27, 2024

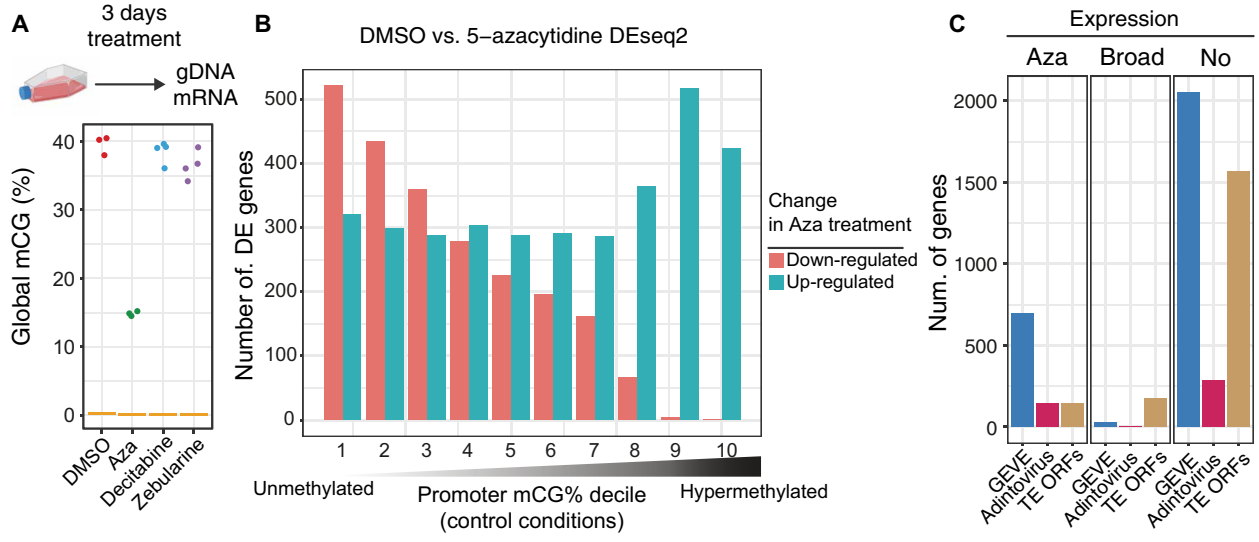

Fig. 3. 5mC removal leads to viral transcriptional reactivation. (**A**) Global methylation levels measured with Enzymatic Methyl-seq for *Amoebidium* cultures treated for 3 days with DMSO (three biological replicates), 5-azacytidine (Aza; three biological replicates), decitabine (four biological replicates), and zebularine (four biological replicates). (**B**) Distribution of differentially expressed genes classified according to the promoter methylation status in untreated conditions (divided in deciles). The bar color depicts the direction of change upon 5-azacytidine treatment. (**C**) Number of genes encoded in GEVEs, adintoviruses, or TE open reading frames according to their transcriptional response to 5-azacytidine treatment. "Aza" are genes that are only transcribed upon treatment, "Broad" are genes that are expressed in any moment of *Amoebidium* development/control conditions, and "No" are genes that are not expressed in any condition (TPM < 1).

formation of viral particles through microscopy, and consistently, we did not see transcriptional reactivation of capsid proteins. This suggests that viral formation would require extra genes that might have been purged or have accumulated critical mutations since the insertion occurred. Alternatively, posttranscriptional silencing mechanisms could stop the formation of mature viral particles. In sum, direct manipulation of the host methylome demonstrates that 5mC is instrumental for silencing and minimizing the consequence of viral DNA acquisition.

Giant virus endogenization is polymorphic and highly dynamic in *Amoebidium*

Maintaining a substantial quantity of potentially harmful viral DNA in the *Amoebidium* genome may serve as an adaptive mechanism with important roles. Conversely, it could also represent a passive outcome facilitated by epigenetic silencing. To assess these hypotheses, we set out to compare genetically distinct *Amoebidium* isolates from our reference genome. We first obtained the transcriptome of six isolates, four belonging to *A. appalachense* and two to *A. parasiticum*. Whereas the isolate's 18*S* sequences were identical at the species level (fig. S9A), the rapidly divergent mitochondrial 16*S* revealed four clades, including a slightly divergent *A. appalachense* lineage (Fig. 4A). We selected a member of the divergent *A. appalachense* lineage (isolate 9181) and one *A. parasiticum* (isolate 9257) for genome sequencing using nanopore long reads. Genome assembly size varied across isolates (Fig. 4A), yet annotation qualities and nanopore-assessed 5mC levels were consistent (fig. S9B).

Annotation of the viral endogenizations in these alternative genotypes revealed a dynamic and diverse history for GEVEs and adintoviruses associated with the *Amoebidium* lineage. Phylogenetic markers such as the major capsid proteins or VLTF3 revealed that at least six separate clades of giant viruses infect these protists, with some clades unique to one isolate (clade II) and others shared by the isolates (clade IV) but absent in the reference genome (Fig. 4B and fig. S9C). Similarly, four adintovirus clades are found across the isolates, with some being shared across all three genomes (fig. S9D). Notably, isolate 9257 shows only 4 adintoviruses compared to the 44 present in the reference genome. This reveals that viral diversity infecting *Amoebidium* is not limited to a single lineage and is often endogenized in an isolate-specific manner.

Since sequence similarity and gene synteny between isolate 9181 and the reference genome remains highly conserved (fig. S9E), we used this to assess the ancestral nature of endogenization events. Despite their close phylogenetic relationship, only a minority of endogenization events were shared across the isolates, with most featuring polymorphic insertions amidst synteny blocks (Fig. 4C). Giant viruses are not known to require integration into the host genome during their infectious cycles (*14*); this process appears to be stochastic, potentially occurring during unsuccessful infections and at various chromosome positions. Additionally, it underscores the dynamic nature of integration, which the host tolerates through multiple cycles, with most endogenized elements being quickly eliminated after insertion.

DISCUSSION

Here, we show how a unicellular eukaryote closely related to animals undergoes a recurrent process of mixing its genome with that of its giant virus predators. This foreign DNA is curbed by 5mC silencing, allowing for survival after these potentially lethal events. We propose that epigenetic silencing greatly reduces the lethality of these endogenization events. Supporting this general hypothesis, many of the previously described large-scale giant virus endogenizations in eukaryotes, including early land plant lineages, the fungus *Rhizophagus irregularis*, green algae, or the amoebozoan *Acanthamoeba castellanii*, coincide with species that have retained 5mC as a silencing mechanism (Fig. 5)

Downloaded from https://www.science.org at Universite de Geneve on August 27, 2024

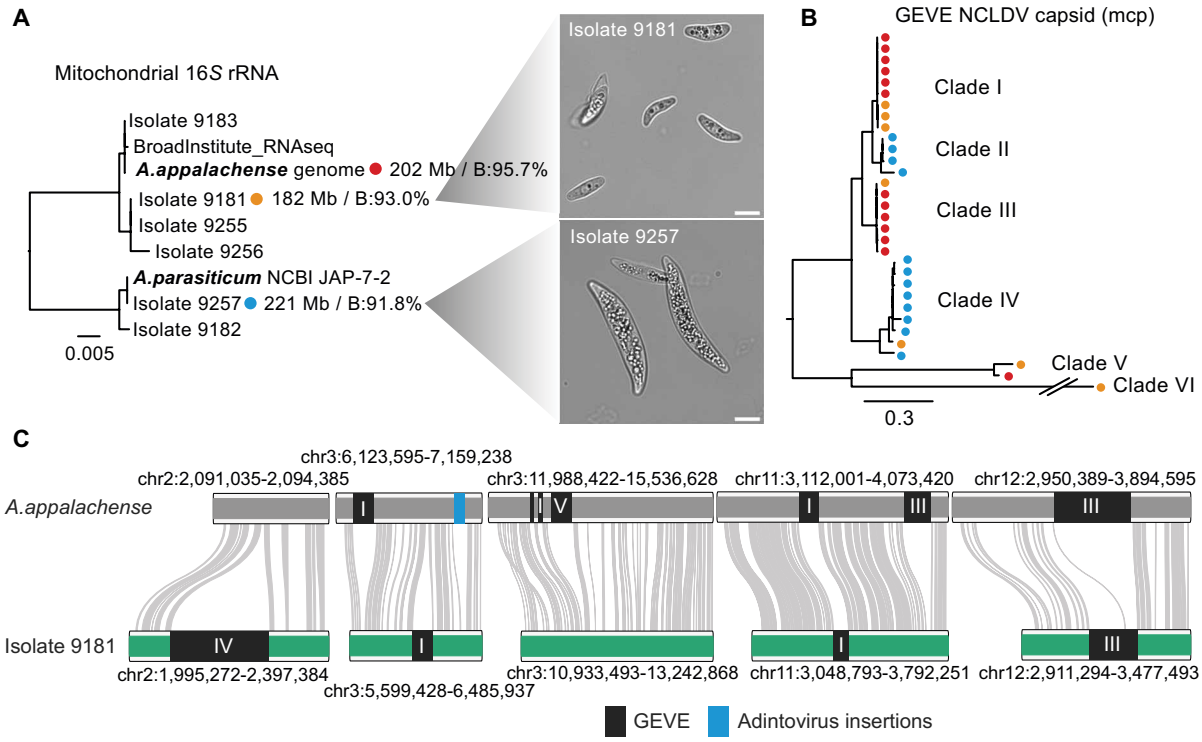

Fig. 4. *Amoebidium* isolates display rapid turnover of viral endogenization events. (**A**) Maximum likelihood phylogenetic tree displaying the mitochondrial 16*S* phylogeny of *Amoebidium* isolates. The reference genome (red), isolate 9181 (orange), and isolate 9257 (blue) display the genome assembly characteristics: size (megabases) and BUSCO completeness (B %). Micrographs display *Amoebidium* isolate cells in culture, with the white bar spanning 10 μm. (**B**) Maximum likelihood phylogeny of NCLDV Major capsid proteins encoded in GEVEs. Dots represent the genome they come from following panel (A) color code. Double slash indicates that the branch has been shortened for display purposes. (**C**) Conserved and polymorphic viral insertions across reference genome and isolate 9181. Gray lines indicate the presence of one-to-one orthologs, whereas dark rectangles indicate viral insertions. Roman numerals indicate the clade of the GEVE according to panel (B) phylogeny.

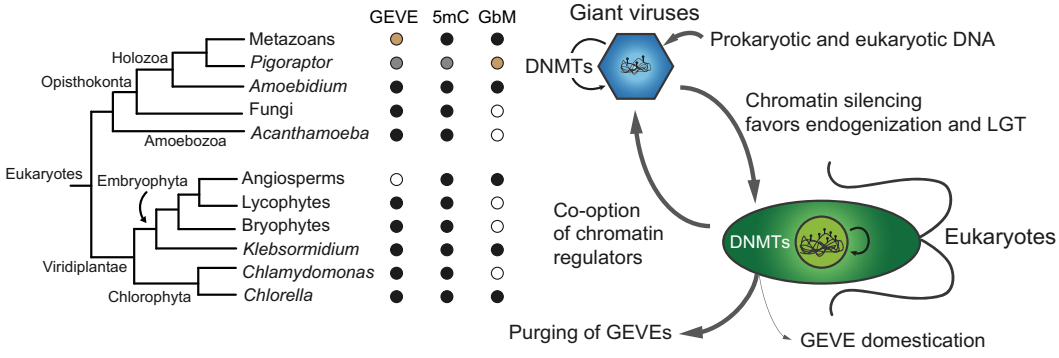

Fig. 5. Giant virus endogenization events correlate with the presence of 5mC in eukaryotes. Cladogram representing the lineages where large GEVEs have been described in the literature, and for which 5mC and gene body methylation pattern (GbM) presence has been reported. Dark dots indicate presence, white dots indicate absence, gray dots indicate that it is highly likely (due to the presence of DNMT1/3), and brown dots indicate lack of data/studies. Schematic of the lateral gene transfers (LGT) from giant viruses to eukaryotes, and vice versa, mediated by chromatin regulatory mechanisms. DNMTs mediate DNA methyl marks shown as lollipops.

(*19*–*21*, *65*, *66*). Plants that are not dependent on water for reproduction (Spermatophyta) or germline-segregating animals are likely protected from giant virus endogenization events despite carrying silencing mechanisms (*20*), yet chromosome-scale genomes and directed searches might reveal exceptions to this rule. It is also possible that some eukaryotic lifestyles might make some species less likely to be infected by giant viruses, such as that of internal parasites, or that extreme genome compaction requirements make endogenizations unlikely to be fixed in a population, such as in prasinophytes. Giant viruses infect all kinds of eukaryotic groups, including species with and without DNMTs. However, eukaryotes that have secondarily lost 5mC silencing, as exemplified by most of the available unicellular holozoan genomes, rarely present large giant viral DNA insertions. Notably, most eukaryotic clades have genes derived from

Downloaded from https://www.science.org at Universite de Geneve on August 27, 2024

giant viruses (*13*), or insertions of double-stranded medium size DNA viruses like polinton-like/virophages (*63*), suggesting that infection and endogenization are widespread, but the retention of these insertions is uneven across lineages. It is likely that other silencing mechanisms other than 5mC, such as histone modifications (e.g., H3K27me3 or H3K9me3) or small interfering RNAs (siRNAs), can be used for silencing GEVEs, as exemplified by H3K79me2 in the brown algae *Ectocarpus siliculosus* GEVE (*67*–*69*), a species that lacks DNMTs and 5mC. Yet, possibly the cost of integrating large amounts of viral DNA in epigenetically unprotected species is coped with by extremely rapid purging or takeover by uninfected conspecific cells.

The 5mC patterns in *Amoebidium* also suggest that gene body methylation predates animal origins. Although this would potentially support the hypothesis that gene body methylation was present in the ancestor of eukaryotes (*39*), this pattern is very sparsely distributed (Fig. 5). Among chlorophytes, only *Chlorella variabilis* has a pattern similar to that of *Amoebidium* (*39*), while absent in *Chlamydomonas* and Prasinophytes (*5*, *40*). Spermatophytes [angiosperms (*70*), conifers (*71*), and ferns (*72*)] show gene body methylation, whereas liverworts and mosses generally lack it (*39*, *73*, *74*). In contrast, the more basally branching streptophyte *Klebsormidium nitens* shows gene body methylation (*65*), albeit in a pattern quite divergent to land plants, which could suggest independent origins of gene body methylation in these lineages. *Amoebidium*, *Chlorella* (*75*), and *Klebsormidium* (*65*) present giant virus endogenization events, which suggests that gene body methylation might arise as a convergent response or by-product to recurrent infections and expansion of parasitic DNA (*70*), perhaps avoiding intra-genic elements hijacking transcription from the host genes (*76*). In particular, *Amoebidium* encodes DNMT3 and an animal-like DNMT1, which could support that gene body methylation across holozoans is homologous and deposited by orthologous DNMTs. In the future, if 5mC data can be obtained from *Pigoraptor* species, it will help to elucidate if *Amoebidium* represents a case of convergent evolution of gene body methylation or this is ancestral to holozoans. With available data, the link with gene body and transcription appears an animal innovation, enabled through the acquisition of PWWP and ADD domains in animal DNMT3 orthologs, starting a feedback loop with histone modifications such as H3K36me2/3 (*77*–*79*). Regulation of host gene transcription in multicellular animals might have restricted and weakened the role of 5mC in TE silencing, suggested by its absence across many invertebrate genomes (*1*, *2*).

Giant viruses emerged before the origins of modern eukaryotes (*16*), and chromatin silencing mechanisms such as 5mC or histone modifications were present in the Last Eukaryotic Common Ancestor (*1*, *9*, *55*). Thus, these patterns of frequent giant virus endogenization that we observe in modern eukaryotes must have been constant during the whole history of the lineage. Although domestication of giant virus–derived genes might be rare, we can see examples of this occurring throughout the tree of life (*13*, *80*, *81*). It is worth highlighting that despite giant virus–derived genes being widespread, their domestication potential is harder to assess, given the difficulty to test their roles and expression across divergent protist species. Thus, giant viruses, whose genetic material is itself a composite of various origins (*82*–*84*), serve as a source of genetic novelty via lateral gene transfer across eukaryotes (Fig. 5). Unlike plasmids or other forms of bacterial lateral gene transfer mechanisms, giant viruses are a dangerous vessel for genetic interchange; thus, chromatin silencing mechanisms are probably required for a stepwise acquisition of foreign DNA. In turn, the host chromatin protection is likely counteracted by giant viruses, as exemplified by the histone demethylases present in *Amoebidium* GEVEs, or other examples of chromatin modifiers reported in giant virus genomes (*55*). Similarly, the presence of DNMTs in GEVEs, and the capacity of giant viruses to modify their own DNA (*54*), could be a protective response against eukaryotic chromatin, avoiding viral DNA to be recognized as a threat. Chromatin hijacking by giant viruses is a process reminiscent of cases in which TEs have co-opted host chromatin regulators (*48*, *65*, *85*, *86*), highlighting the age-long conflict between eukaryotic chromatin and parasitic DNA. In summary, *Amoebidium* exemplifies the intricate network-like origins of eukaryotic DNA, challenging traditional notions of strict vertical inheritance within the clade.

MATERIALS AND METHODS

Cell culture, treatment, and nucleic acid extraction

Amoebidium isolates were grown on Brain Heart Infusion (10% BHI, Thermo Fisher Scientific CM1135) liquid medium at 25°C in 25-ml culture flasks. *S. arctica, Creolimax fragrantissima*, and *Chromosphaera perkinsii* were grown in liquid Marine Broth (Difco Marine Broth 2216) at 17°C. Six *Amoebidium* alternative isolates were obtained from the ARS Collection of Entomopathogenic Fungal Cultures.

DNA methylation drugs 5-azacytidine (ab142744), decitabine (ab120842), and zebularine (ab141264) were dissolved in dimethyl sulfoxide (DMSO). *A. appalachense* was grown with 0 M, 0.1 μM, 1 μM, 10 μM, 100 μM, and 1 mM final concentration of each drug in 2 ml of 10% BHI with 10% DMSO in a 12-well plate, and effects were tracked daily for 5 days. Only 100 μM and 1 mM 5-azacytidine showed a growth phenotype. *A. appalachense* DNA and RNA were extracted from cultures grown for 3 days in 10 ml of 10% BHI with 1% DMSO, and 1% DMSO with 100 μM of their respective drug, in triplicate. 5-Azacytidine (12 nmol) in 120 μl of DMSO was spread over 12-ml agar plates of BHI (*A. appalachense*) and Marine Broth (*S. arctica*, *C. fragrantissima*, *C. perkinsii*), and dilution assays for growth were done for all four ichthyosporean species using 1×, 10×, 100×, 1000×, and 10,000× serial dilutions of saturated culture.

The developmental cell cycle of *A. appalachense* was determined using a combination of live and fixed-cell microscopy using a fully motorized Nikon Ti2-E epifluorescence inverted microscope equipped with a hardware autofocus PFS4 system, a Lumencor SOLA SMII illumination system, and a Hamamatsu ORCA-spark Digital CMOS camera. CFI Plan Fluor 20×, 0.50 NA (numerical aperture), CFI Plan Fluor 40× Air, and CFI Plan Fluor 60× Oil, 0.5 to 1.25 NA objectives were used for imaging. For live-cell microscopy, a 25-day-old culture was diluted 1:250 and imaged with bright field every 15 min for 72 hours at a controlled temperature of 23°C in 600-μl wells using a cooling/heating P Lab-Tek S1 insert (Pecon GmbH) with Lauda Loop 100 circulating water bath. We examined 120 videos counting events of spontaneous cell death, cellularization, and cell release, and the number of released spores per colony (total 703 cells tracked; table S1). For fluorescent microscopy, samples were fixed in 4% formaldehyde, washed with phosphate-buffered saline (PBS), and stained with phalloidin and Hoechst to visualize and count actin and nuclei, respectively, every 4 to 5 hours over 72 hours. RNA and DNA were obtained for representative

Downloaded from https://www.science.org at Universite de Geneve on August 27, 2024

stages of the life cycle: 5 hours after inoculation (unicell—uninucleated), 14 hours (coenocyte), 20 hours (cellularization), 33 hours (cell release).

DNA for *A. appalachense* genome sequencing was extracted using liquid nitrogen grinding and Qiagen MagAttract HMW DNA Kit & QIAGEN Genomic-tip 20/G (10223), and for *A. appalachense*, *S. arctica*, *C. perkinsii*, and *C. fragrantissima* Enzymatic Methyl-seq samples, we used NEB Monarch Genomic DNA Purification Kit. DNA for *Amoebidium* isolates 9181 and 9257 was extracted with phenol chloroform extraction and further purification with NEB Monarch Genomic DNA Purification Kit. RNA for all samples was extracted using nitrogen grinding and Monarch Total RNA Miniprep Kit.

Micro-C library preparation

A. appalachense cells grown for 7 days were crosslinked for 10 min with 1% formaldehyde under vacuum conditions in a desiccator. The reaction was quenched with 128 mM glycine for 5 min under vacuum, followed by an additional incubation on ice for 15 min. Crosslinked cells were washed twice and subsequently resuspended in a $^{1}/_{10}$ PBS solution. Coenocytic cell walls were disrupted by glass bead beating for 5 min followed by a second crosslinking step with 3 mM DSG (disuccinimidyl glutarate) for 40 min at room temperature.

Micro-C libraries were prepared as described (*87*) with the following modifications. In-nuclei chromatin digestion to achieve 80% monomer/20% oligomer nucleosome ratio was performed with 100 U of MNase (Takara Bio, 2910a) per 4 M nuclei for 10 min. The digested chromatin ends were repaired and labeled with biotinylated nucleotides. Before proximity ligation, the digested chromatin was released from nuclei and permeabilized coenocytes by glass bead beating for 10 min. Next, proximal nucleosomes were ligated together, and unligated ends were treated with Exonuclease III (NEB, M0206) to remove biotin-dNTPs (Deoxynucleotide Triphosphates). The chromatin was then decrosslinked and deproteinized, and ligated DNA fragments were captured with Dynabeads MyOne Streptavidin (Life Technologies, 65602). Libraries were barcoded using the NEBNext End repair/dA-tailing mix (NEB, E7546) and NEBNext Ultra II Ligation Module (NEB, E7595S). The final amplified libraries, comprising three biological replicates, were sequenced with NextSeq500 in paired-end format with a read length 42 bases per mate, obtaining a total of 131,547,803 sequenced reads.

Genome sequencing and assembly

High molecular weight genomic DNA from *A. appalachense* was ligated with the Nanopore SQK-LSK110 ligation kit and sequenced in Promethion R9 flowcells. Since pore clogging occurred quickly, we performed short sequencing runs followed by flowcell cleanup steps, and reloading of fresh library in intervals, requiring three flowcells. In parallel, a library of paired-end short reads was generated with the TruSeq kit and sequenced with an Illumina HiSeq2500. Nanopore reads were basecalled using the "sup" model with Guppy (v6.2.1) and assembled with Flye (v2.9-b1768) with the "--nanopore_hq" parameter and two rounds of polishing (*88*). The resulting genome was further polished with the short reads with Pilon (*89*) for two rounds, using BUSCO score (-m genome, v5) (*90*) to validate improvements, obtaining a contig level N50 of 1.8 Mb. Micro-C data were mapped on the genome using Juicer (v1.6) (*91*) with the -p assembly option. The 3D-DNA pipeline (*92*), using the proximity ligation data, was used to scaffold the genome with -r3 -editor-repeat-coverage 10. Final manual curation in the Juicebox Assembly Tool (*93*) resulted in 18 chromosomes. The genome was then polished using Medaka with the original nanopore reads.

For the isolates 9181 and 9257, we ligated the DNA using Nanopore SQK-LSK114 ligation kit and sequenced following the same strategy but using PromethIon R10 flowcells (table S2). Contig-level assembly was obtained using Flye with Guppy "sup" base called reads as above. Medaka polishing was discarded as it decreased BUSCO score. Then, D-GENIES was used to visualize the synteny with the reference genome (*94*). RagTag scaffolding using the reference genome was performed for both isolate contigs (*95*), yet only 9181 was kept as 98% of the sequence were placed into chromosomes, whereas 9257 only got 54%, rendering the scaffolding unreliable. Extra scaffolding using P_RNA_scaffolder (*96*) was performed for 9257 using its transcriptomic data to further increase contiguity, and validated through BUSCO improvement criteria.

Genome annotation

We generated a de novo RepeatModeler2 (*97*) annotation with the LTR module to characterize *A. appalachense* repeat landscape. This was then mapped to the genome using RepeatMasker. In parallel, publicly available deep coverage RNA-seq from *A. appalachense* (SRR545192) was mapped to the genome using HISAT2 with the–dta parameter, and Stringtie for reference based transcriptome assembly (*98*). The resulting bam was processed with Portcullis to generate a list of high-quality intron junctions (*99*). In parallel, de novo Trinity assembly of the SRR545192 reads was mapped using gmap to the genome (*100*). The combination of introns, Stringtie, and Trinity mappings was fed to Mikado to choose the best collection of transcripts based on the UniProt Sprot database. The best transcripts were used to train Augustus model for *Amoebidium* (*101*). To inform Augustus annotation, we mapped protein alignments against the genome using MetaEuk (*102*), using closely related high-quality ichthyosporean genomes as query, obtaining coding sequence hints. Portcullis introns and Mikado exons were also introduced as hints for Augustus genome annotation. The resulting Augustus annotation was then updated using PASA with the Mikado transcripts, fixing broken gene models and adding untranslated regions. Annotation was visually inspected in the IGV genome browser. Functional annotation was obtained using hmmscan with Pfam-A database (*103*) and the eggNOG-mapper server (*104*).

To annotate the alternative isolate genomes, the same process was followed, using the reference annotation for the MetaEuk CDS hints, and the pre-trained *Amoebidium* Augustus model. All annotations were evaluated using BUSCO v5 with eukaryota_odb10 database.

Transcriptome sequencing, assembly, and analysis

We used 50 to 1000 ng of RNA from treated samples, developmental time points, and isolates to build mRNA-seq libraries (see table S3 for details), first enriching for poly-A transcripts with the NEB Magnetic mRNA Isolation Kit S1550S, and then building the libraries with the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (E7760L) according to manufacturer's instructions. Short-read Illumina reads were obtained with NovaSeq6000. De novo transcriptome assemblies were obtained with Trinity (strand-specific) for the isolates. The Trinity assemblies were searched

Downloaded from https://www.science.org at Universite de Geneve on August 27, 2024

for 18*S* and 16*S* sequences using BLASTn with NCBI query sequences.

Drug treatment and developmental samples were mapped against the annotation using Kallisto to obtain TPMs (*105*). To perform differential expression analysis of TEs and protein-coding genes, we used HISAT2 with the TElocal pipeline (*106*), obtaining gene counts that were then analyzed in DEseq2 (*107*). Only intergenic TEs above 500 base pairs (bp) were kept for the analysis. *Sphaeroforma* treatment samples were done in the same way and mapped to the latest version of the genome (*108*).

Methylome sequencing and analysis

We sonicated genomic DNA from *A. appalachense* (control, developmental time points, DMSO/5-azacytidine treated), *S. arctica*, *C. perkinsii*, and *C. fragrantissima*, spiked with phage lambda DNA and methylated pUC19 controls, to obtain 300-bp fragments with Covaris M220. Then, we used the NEB Enzymatic Methyl-Kit to convert all the unmethylated Cs into Ts as described in the manufacturer's instructions (*109*). These libraries were then sequenced in Illumina NovaSeq6000 to various coverages (table S4). The reads were then mapped with fastp and mapped to the reference genomes (*29*, *108*, *110*) using BS-Seeker2 backed with bowtie2 (*111*). Sambamba was used to remove polymerase chain reaction (PCR) duplicates, and CGmapTools was used to obtain the methyl calls (*112*). These files were processed in R using the bsseq package, and bigwig tracks using the BedGraphToBigWig UCSC utility.

In parallel, nanopore reads were basecalled and mapped for base modifications using the Guppy dna_r9.4.1_450bps_modbases_5mc_cg_sup_prom.cfg and dna_r10.4.1_e8.2_400bps_modbases_5mc_cg_sup_prom.cfg models. The resulting read alignments were processed with modbam2bed the --cpg -e -m 5mC parameters. These bed files were also processed in R using the bsseq package.

Giant virus identification and phylogenetic analysis

Visual inspection of hypermethylated blocks revealed core giant virus genes in unusual gene architecture patterns. To validate these potential claims, we used ViralRecall (*44*) that flagged just a few of these sequences as potential giant virus endogenization events. However, we observed that many events were not captured by that software, so we manually inspected the genome to obtain the longest potential inserts, filtering out TEs inserted within the viral region. We searched those consensus sequences against the genome using BLASTn to obtain all potential regions of homology to giant viruses. Another round of manual inspection of all chromosomes using non-CGC/GCG methylation blocks as boundary demarcation was used to delimit integration sites. The same process was used for adintoviruses and Plavaka giant repeats. We ran ViralRecall on the genomes of other ichthyosporeans (*C. fragrantissima*, *S. arctica*, *Ichthyophonus hoferi*, *C. perkinsii*, *Abeoforma whisleri*, *Pirum gemmata*) (*29*, *110*), *Corallochytrium limacisporum*, the filasterean *Capsaspora owczarzaki* (*113*), and the choanoflagellates *Monosiga brevicollis* and *Salpingoeca rosetta* (*114*, *115*), and we did not obtain any reliable hit on this collection of holozoan genomes (table S5).

Hmmsearch was used to identify core viral genes, DNMTs (PF00145), and JmjC (PF02373)–containing proteins. DNMTs and core NCLDV genes were searched in a large collection of holozoan genomes and transcriptomes, including 22 choanoflagellates (*31*, *116*), 4 filastereans (*35*), 7 ichthyosporeans, and *C. limacisporum* (see table S5). The obtained genes were included to sequences from reference databases (*42*, *45*, *55*) and aligned using MAFFT in lins-i mode (*117*). Alignments were trimmed using TrimAL with the -gappyout mode (*118*). The resulting alignments were fed into IQ-TREE 2 with automatic model testing to build maximum likelihood phylogenetic trees using altr and uboot as nodal support measures (*119*). Adintovirus minor and major capsid proteins were annotated with HHpred against PDB_mmCIF70 database.

Comparative genomics among giant virus insertions (used as independent taxa) or across the isolate genomes was performed using OrthoFinder with DIAMOND as a search engine (*120*).

Supplementary Materials

This PDF file includes:

Figs. S1 to S9

Legends for tables S1 to S5

Other Supplementary Material for this manuscript includes the following:

Tables S1 to S5

REFERENCES AND NOTES

1. A. de Mendoza, R. Lister, O. Bogdanovic, Evolution of DNA Methylome diversity in eukaryotes. *J. Mol. Biol.* **432**, 1687–1705 (2020).
2. P. Sarkies, Encyclopaedia of eukaryotic DNA methylation: From patterns to mechanisms and functions. *Biochem. Soc. Trans.* **50**, 1179–1190 (2022).
3. R. J. Schmitz, Z. A. Lewis, M. G. Goll, DNA methylation: Shared and divergent features across eukaryotes. *Trends Genet.* **35**, 818–827 (2019).
4. L. Ponger, W.-H. Li, Evolutionary diversification of DNA methyltransferases in eukaryotic genomes. *Mol. Biol. Evol.* **22**, 1119–1128 (2005).
5. J. T. Huff, D. Zilberman, Dnmt1-independent CG methylation contributes to nucleosome positioning in diverse eukaryotes. *Cell* **156**, 1286–1297 (2014).
6. S. Catania, P. A. Dumesic, H. Pimentel, A. Nasif, C. I. Stoddard, J. E. Burke, J. K. Diedrich, S. Cooke, T. Shea, E. Gienger, R. Lintner, J. R. Yates III, P. Hajkova, G. J. Narlikar, C. A. Cuomo, J. K. Pritchard, H. D. Madhani, Evolutionary persistence of DNA methylation for millions of years after ancient loss of a *de novo* methyltransferase. *Cell* **180**, 263–277.e20 (2020).
7. F. Lyko, The DNA methyltransferase family: A versatile toolkit for epigenetic regulation. *Nat. Rev. Genet.* **19**, 81–92 (2018).
8. M. M. Suzuki, A. Bird, DNA methylation landscapes: Provocative insights from epigenomics. *Nat. Rev. Genet.* **9**, 465–476 (2008).
9. A. Zemach, D. Zilberman, Evolution of eukaryotic DNA methylation and the pursuit of safer sex. *Curr. Biol.* **20**, R780–R785 (2010).
10. Ö. Deniz, J. M. Frost, M. R. Branco, Regulation of transposable elements by DNA modifications. *Nat. Rev. Genet.* **20**, 417–431 (2019).
11. K. Harbers, A. Schnieke, H. Stuhlmann, D. Jähner, R. Jaenisch, DNA methylation and gene expression: Endogenous retroviral genome becomes infectious after molecular cloning. *Proc. Natl. Acad. Sci. U.S.A.* **78**, 7609–7613 (1981).
12. B. P. Anton, R. J. Roberts, Beyond restriction modification: Epigenomic roles of DNA methylation in prokaryotes. *Annu. Rev. Microbiol.* **75**, 129–149 (2021).
13. N. A. T. Irwin, A. A. Pittis, T. A. Richards, P. J. Keeling, Systematic evaluation of horizontal gene transfer between eukaryotes and viruses. *Nat. Microbiol.* **7**, 327–336 (2022).
14. F. Schulz, C. Abergel, T. Woyke, Giant virus biology and diversity in the era of genome-resolved metagenomics. *Nat. Rev. Microbiol.* **20**, 721–736 (2022).
15. M. Krupovic, V. V. Dolja, E. V. Koonin, The virome of the last eukaryotic common ancestor and eukaryogenesis. *Nat. Microbiol.* **8**, 1008–1017 (2023).
16. J. Guglielmini, A. C. Woo, M. Krupovic, P. Forterre, M. Gaia, Diversification of giant and large eukaryotic dsDNA viruses predated the origin of modern eukaryotes. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 19585–19592 (2019).
17. S. Karki, Z. K. Barth, F. O. Aylward, Chimeric origin of eukaryotes from Asgard archaea and ancestral giant viruses. *bioRxiv* 2024.04.22.590592 (2024).
18. P. J. L. Bell, Eukaryogenesis: The rise of an emergent superorganism. *Front. Microbiol.* **13**, 858064 (2022).
19. M. Moniruzzaman, A. R. Weinheimer, C. A. Martinez-Gutierrez, F. O. Aylward, Widespread endogenization of giant viruses shapes genomes of green algae. *Nature* **588**, 141–145 (2020).
20. F. Maumus, A. Epert, F. Nogué, G. Blanc, Plant genomes enclose footprints of past infections by giant virus relatives. *Nat. Commun.* **5**, 4268 (2014).
21. J. Filée, Multiple occurrences of giant virus core genes acquired by eukaryotic genomes: The visible part of the iceberg? *Virology* **466–467**, 53–59 (2014).

Downloaded from https://www.science.org at Universite de Geneve on August 27, 2024

22. M. Moniruzzaman, M. P. Erazo-Garcia, F. O. Aylward, Endogenous giant viruses contribute to intraspecies genomic variability in the model green alga *Chlamydomonas reinhardtii*. *Virus Evol.* **8**, veac102 (2022).
23. S. Rošić, R. Amouroux, C. E. Requena, A. Gomes, M. Emperle, T. Beltran, J. K. Rane, S. Linnett, M. E. Selkirk, P. H. Schiffer, A. J. Bancroft, R. K. Grencis, A. Jeltsch, P. Hajkova, P. Sarkies, Evolutionary analysis indicates that DNA alkylation damage is a byproduct of cytosine DNA methyltransferase activity. *Nat. Genet.* **50**, 452–459 (2018).
24. A. J. Bewick, B. T. Hofmeister, R. A. Powers, S. J. Mondo, I. V. Grigoriev, T. Y. James, J. E. Stajich, R. J. Schmitz, Diversity of cytosine methylation across the fungal tree of life. *Nat. Ecol. Evol.* **3**, 479–490 (2019).
25. D. Schübeler, Function and information content of DNA methylation. *Nature* **517**, 321–326 (2015).
26. N. Ros-Rocher, A. Pérez-Posada, M. M. Leger, I. Ruiz-Trillo, The origin of animals: An ancestral reconstruction of the unicellular-to-multicellular transition. *Open Biol.* **11**, 200359 (2021).
27. A. Sebé-Pedrós, B. M. Degnan, I. Ruiz-Trillo, The origin of Metazoa: A unicellular perspective. *Nat. Rev. Genet.* **18**, 498–512 (2017).
28. T. Brunet, N. King, The origin of animal multicellularity and cell differentiation. *Dev. Cell* **43**, 124–140 (2017).
29. X. Grau-Bové, G. Torruella, S. Donachie, H. Suga, G. Leonard, T. A. Richards, I. Ruiz-Trillo, Dynamics of genomic innovation in the unicellular ancestry of animals. *eLife* **6**, e26036 (2017).
30. A. de Mendoza, W. L. Hatleberg, K. Pang, S. Leininger, O. Bogdanovic, J. Pflueger, S. Buckberry, U. Technau, A. Hejnol, M. Adamska, B. M. Degnan, S. M. Degnan, R. Lister, Convergent evolution of a vertebrate-like methylome in a marine sponge. *Nat. Ecol. Evol.* **3**, 1464–1473 (2019).
31. D. J. Richter, P. Fozouni, M. B. Eisen, N. King, Gene family innovation, conservation and loss on the animal stem lineage. *eLife* **7**, e34226 (2018).
32. M. M. White, A. Siri, R. W. Lichtwardt, Trichomycete insect symbionts in Great Smoky Mountains National Park and vicinity. *Mycologia* **98**, 333–352 (2006).
33. S. L. Glockling, W. L. Marshall, F. H. Gleason, Phylogenetic interpretations and ecological potentials of the Mesomycetozoea (Ichthyosporea). *Fungal Ecol.* **6**, 237–247 (2013).
34. H. C. Whisler, Culture and nutrition of Amoebidium parasiticum. *Am. J. Bot.* **49**, 193–199 (1962).
35. E. Ocaña-Pallarès, T. A. Williams, D. López-Escardó, A. S. Arroyo, J. S. Pathmanathan, E. Bapteste, D. V. Tikhonenkov, P. J. Keeling, G. J. Szöllősi, I. Ruiz-Trillo, Divergent genomic trajectories predate the origin of animals and fungi. *Nature* **609**, 747–753 (2022).
36. E. Hehenberger, D. V. Tikhonenkov, M. Kolisko, J. Del Campo, A. S. Esaulov, A. P. Mylnikov, P. J. Keeling, Novel predators reshape holozoan phylogeny and reveal the presence of a two-component signaling system in the ancestor of animals. *Curr. Biol.* **27**, 2043–2050.e6 (2017).
37. B. Cuypers, F. Dumetz, P. Meysman, K. Laukens, G. De Muylder, J.-C. Dujardin, M. A. Domagalska, The absence of C-5 DNA Methylation in *Leishmania donovani* allows DNA enrichment from complex samples. *Microorganisms* **8**, 1252 (2020).
38. W. Zhou, H. Q. Dinh, Z. Ramjan, D. J. Weisenberger, C. M. Nicolet, H. Shen, P. W. Laird, B. P. Berman, DNA methylation loss in late-replicating domains is linked to mitotic cell division. *Nat. Genet.* **50**, 591–602 (2018).
39. A. Zemach, I. E. McDaniel, P. Silva, D. Zilberman, Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* **328**, 916–919 (2010).
40. S. Feng, S. J. Cokus, X. Zhang, P.-Y. Chen, M. Bostick, M. G. Goll, J. Hetzel, J. Jain, S. H. Strauss, M. E. Halpern, C. Ukomadu, K. C. Sadler, S. Pradhan, M. Pellegrini, S. E. Jacobsen, Conservation and divergence of methylation patterning in plants and animals. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 8689–8694 (2010).
41. W. Zhou, G. Liang, P. L. Molloy, P. A. Jones, DNA methylation enables transposable element-driven genome expansion. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 19359–19366 (2020).
42. F. O. Aylward, M. Moniruzzaman, A. D. Ha, E. V. Koonin, A phylogenomic framework for charting the diversity and evolution of giant viruses. *PLOS Biol.* **19**, e3001430 (2021).
43. R. Zhang, M. Takemura, K. Murata, H. Ogata, "Mamonoviridae", a proposed new family of the phylum Nucleocytoviricota. *Arch. Virol.* **168**, 80 (2023).
44. F. O. Aylward, M. Moniruzzaman, ViralRecall—A flexible command-line tool for the detection of giant virus signatures in 'omic data. *Viruses* **13**, 150 (2021).
45. G. J. Starrett, M. J. Tisza, N. L. Welch, A. K. Belford, A. Peretti, D. V. Pastrana, C. B. Buck, Adintoviruses: A proposed animal-tropic family of midsize eukaryotic linear dsDNA (MELD) viruses. *Virus Evol.* **7**, veaa055 (2021).
46. M. Krupovic, E. V. Koonin, Polintons: A hotbed of eukaryotic virus, transposon and plasmid evolution. *Nat. Rev. Microbiol.* **13**, 105–115 (2015).
47. M. G. Fischer, C. A. Suttle, A virophage at the origin of large DNA transposons. *Science* **332**, 231–234 (2011).
48. L. M. Iyer, D. Zhang, R. F. de Souza, P. J. Pukkila, A. Rao, L. Aravind, Lineage-specific expansions of TET/JBP genes and a new class of DNA transposons shape fungal genomic and epigenetic landscapes. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 1676–1683 (2014).
49. R. T. M. Poulter, M. I. Butler, Tyrosine recombinase retrotransposons and transposons. *Microbiol. Spectr.* **3**, MDNA3–0036–2014 (2015).
50. D. Lang, K. K. Ullrich, F. Murat, J. Fuchs, J. Jenkins, F. B. Haas, M. Piednoel, H. Gundlach, M. Van Bel, R. Meyberg, C. Vives, J. Morata, A. Symeonidi, M. Hiss, W. Muchero, Y. Kamisugi, O. Saleh, G. Blanc, E. L. Decker, N. van Gessel, J. Grimwood, R. D. Hayes, S. W. Graham, L. E. Gunter, S. F. McDaniel, S. N. W. Hoernstein, A. Larsson, F.-W. Li, P.-F. Perroud, J. Phillips, P. Ranjan, D. S. Rokshar, C. J. Rothfels, L. Schneider, S. Shu, D. W. Stevenson, F. Thümmler, M. Tillich, J. C. Villarreal Aguilar, T. Widiez, G. K.-S. Wong, A. Wymore, Y. Zhang, A. D. Zimmer, R. S. Quatrano, K. F. X. Mayer, D. Goodstein, J. M. Casacuberta, K. Vandepoele, R. Reski, A. C. Cuming, G. A. Tuskan, F. Maumus, J. Salse, J. Schmutz, S. A. Rensing, The Physcomitrella patens chromosome-scale assembly reveals moss genome structure and evolution. *Plant J.* **93**, 515–533 (2018).
51. F. Schulz, S. Roux, D. Paez-Espino, S. Jungbluth, D. A. Walsh, V. J. Denef, K. D. McMahon, K. T. Konstantinidis, E. A. Eloe-Fadrosh, N. C. Kyrpides, T. Woyke, Giant virus diversity and host interactions through global metagenomics. *Nature* **578**, 432–436 (2020).
52. G. Thiel, T. Greiner, D. D. Dunigan, A. Moroni, J. L. Van Etten, Large dsDNA chloroviruses encode diverse membrane transport proteins. *Virology* **479–480**, 38–45 (2015).
53. M. Moniruzzaman, M. P. Erazo Garcia, R. Farzad, A. D. Ha, A. Jivaji, S. Karki, U. Sheyn, J. Stanton, B. Minch, D. Stephens, D. C. Hancks, R. A. L. Rodrigues, J. S. Abrahao, A. Vardi, F. O. Aylward, Virologs, viral mimicry, and virocell metabolism: The expanding scale of cellular functions encoded in the complex genomes of giant viruses. *FEMS Microbiol. Rev.* **47**, fuad053 (2023).
54. S. Jeudy, S. Rigou, J.-M. Alempic, J.-M. Claverie, C. Abergel, M. Legendre, The DNA methylation landscape of giant viruses. *Nat. Commun.* **11**, 2657 (2020).
55. X. Grau-Bové, C. Navarrete, C. Chiva, T. Pribasnig, M. Antó, G. Torruella, L. J. Galindo, B. F. Lang, D. Moreira, P. López-Garcia, I. Ruiz-Trillo, C. Schleper, E. Sabidó, A. Sebé-Pedrós, A phylogenetic and proteomic reconstruction of eukaryotic chromatin evolution. *Nat. Ecol. Evol.* **6**, 1007–1023 (2022).
56. A. J. Erives, Phylogenetic analysis of the core histone doublet and DNA topo II genes of Marseilleviridae: Evidence of proto-eukaryotic provenance. *Epigenet. Chromatin* **10**, 55 (2017).
57. M. I. Valencia-Sánchez, S. Abini-Agbomson, M. Wang, R. Lee, N. Vasilyev, J. Zhang, P. De Ioannes, B. La Scola, P. Talbert, S. Henikoff, E. Nudler, A. Erives, K.-J. Armache, The structure of a virus-encoded nucleosome. *Nat. Struct. Mol. Biol.* **28**, 413–417 (2021).
58. D. H. Lee, G. W. Kim, Y. H. Jeon, J. Yoo, S. W. Lee, S. H. Kwon, Advances in histone demethylase KDM4 as cancer therapeutic targets. *FASEB J.* **34**, 3461–3484 (2020).
59. D. V. Santi, A. Norment, C. E. Garrett, Covalent bond formation between a DNA-cytosine methyltransferase and DNA containing 5-azacytosine. *Proc. Natl. Acad. Sci. U.S.A.* **81**, 6993–6997 (1984).
60. L. Zhou, X. Cheng, B. A. Connolly, M. J. Dickman, P. J. Hurd, D. P. Hornby, Zebularine: A novel DNA methylation inhibitor that forms a covalent complex with DNA methyltransferases. *J. Mol. Biol.* **321**, 591–599 (2002).
61. J. Diesch, M.-M. Le Pannérer, R. Winkler, R. Casquero, M. Muhar, M. van der Garde, M. Maher, C. M. Herráez, J. J. Bech-Serra, M. Fellner, P. Rathert, N. Brooks, L. Zamora, A. Gentilella, C. de la Torre, J. Zuber, K. S. Götze, M. Buschbeck, Inhibition of CBP synergizes with the RNA-dependent mechanisms of azacitidine by limiting protein synthesis. *Nat. Commun.* **12**, 6060 (2021).
62. L. H. Li, E. J. Olin, H. H. Buskirk, L. M. Reineke, Cytotoxicity and mode of action of 5-azacytidine on L1210 leukemia. *Cancer Res.* **30**, 2760–2769 (1970).
63. C. Bellas, T. Hackl, M.-S. Plakolb, A. Koslová, M. G. Fischer, R. Sommaruga, Large-scale invasion of unicellular eukaryotic genomes by integrating DNA viruses. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2300465120 (2023).
64. D. Roulois, H. Loo Yau, R. Singhania, Y. Wang, A. Danesh, S. Y. Shen, H. Han, G. Liang, P. A. Jones, T. J. Pugh, C. O'Brien, D. D. De Carvalho, DNA-demethylating agents target colorectal cancer cells by inducing viral mimicry by endogenous transcripts. *Cell* **162**, 961–973 (2015).
65. A. de Mendoza, A. Bonnet, D. B. Vargas-Landin, N. Ji, H. Li, F. Yang, L. Li, K. Hori, J. Pflueger, S. Buckberry, H. Ohta, N. Rosic, P. Lesage, S. Lin, R. Lister, Recurrent acquisition of cytosine methyltransferases into eukaryotic retrotransposons. *Nat. Commun.* **9**, 1341 (2018).
66. F. Maumus, G. Blanc, Study of gene trafficking between *Acanthamoeba* and giant viruses suggests an undiscovered family of amoeba-infecting viruses. *Genome Biol. Evol.* **8**, 3351–3363 (2016).
67. N. Kobayashi, T. A. Dang, K. T. M. Pham, L. B. Gómez Luciano, B. Van Vu, K. Izumitsu, M. Shimizu, K.-I. Ikeda, W.-H. Li, H. Nakayashiki, Horizontally transferred DNA in the genome of the fungus *Pyricularia oryzae* is associated with repressive histone modifications. *Mol. Biol. Evol.* **40**, msad186 (2023).
68. T. Hisanaga, F. Romani, S. Wu, T. Kowar, Y. Wu, R. Lintermann, A. Fridrich, C. H. Cho, T. Chaumier, B. Jamge, S. A. Montgomery, E. Axelsson, S. Akimcheva, T. Dierschke, J. L. Bowman, T. Fujiwara, S. Hirooka, S.-Y. Miyagishima, L. Dolan, L. Tirichine, D. Schubert, F. Berger, The Polycomb repressive complex 2 deposits H3K27me3 and represses transposable elements in a broad range of eukaryotes. *Curr. Biol.* **33**, 4367–4380.e9 (2023).

Downloaded from https://www.science.org at Universite de Geneve on August 27, 2024

69. S. Bourdareau, L. Tirichine, B. Lombard, D. Loew, D. Scornet, Y. Wu, S. M. Coelho, J. M. Cock, Histone modifications during the life cycle of the brown alga Ectocarpus. *Genome Biol.* **22**, 12 (2021).
70. A. J. Bewick, R. J. Schmitz, Gene body DNA methylation in plants. *Curr. Opin. Plant Biol.* **36**, 103–110 (2017).
71. J. Li, F. Han, T. Yuan, W. Li, Y. Li, H. X. Wu, H. Wei, S. Niu, The methylation landscape of giga-genome and the epigenetic timer of age in Chinese pine. *Nat. Commun.* **14**, 1947 (2023).
72. D. B. Marchant, G. Chen, S. Cai, F. Chen, P. Schafran, J. Jenkins, S. Shu, C. Plott, J. Webber, J. T. Lovell, G. He, L. Sandor, M. Williams, S. Rajasekar, A. Healey, K. Barry, Y. Zhang, E. Sessa, R. R. Dhakal, P. G. Wolf, A. Harkess, F.-W. Li, C. Rössner, A. Becker, L. Gramzow, D. Xue, Y. Wu, T. Tong, Y. Wang, F. Dai, S. Hua, H. Wang, S. Xu, F. Xu, H. Duan, G. Theißen, M. R. McKain, Z. Li, M. T. W. McKibben, M. S. Barker, R. J. Schmitz, D. W. Stevenson, C. Zumajo-Cardona, B. A. Ambrose, J. H. Leebens-Mack, J. Grimwood, J. Schmutz, P. S. Soltis, D. E. Soltis, Z.-H. Chen, Dynamic genome evolution in a model fern. *Nat. Plants* **8**, 1038–1051 (2022).
73. S. Takuno, J.-H. Ran, B. S. Gaut, Evolutionary patterns of genic DNA methylation vary across land plants. *Nat. Plants* **2**, 15222 (2016).
74. A. J. Bewick, C. E. Niederhuth, L. Ji, N. A. Rohr, P. T. Griffin, J. Leebens-Mack, R. J. Schmitz, The evolution of CHROMOMETHYLASES and gene body DNA methylation in plants. *Genome Biol.* **18**, 65 (2017).
75. G. Blanc, G. Duncan, I. Agarkova, M. Borodovsky, J. Gurnon, A. Kuo, E. Lindquist, S. Lucas, J. Pangilinan, J. Polle, A. Salamov, A. Terry, T. Yamada, D. D. Dunigan, I. V. Grigoriev, J.-M. Claverie, J. L. Van Etten, The Chlorella variabilis NC64A genome reveals adaptation to photosymbiosis, coevolution with viruses, and cryptic sex. *Plant Cell* **22**, 2943–2955 (2010).
76. D. Zilberman, An evolutionary case for functional gene body methylation in plants and animals. *Genome Biol.* **18**, 87 (2017).
77. A. Dhayalan, A. Rajavelu, P. Rathert, R. Tamas, R. Z. Jurkowska, S. Ragozin, A. Jeltsch, The Dnmt3a PWWP domain reads histone 3 lysine 36 trimethylation and guides DNA methylation. *J. Biol. Chem.* **285**, 26114–26120 (2010).
78. D. N. Weinberg, S. Papillon-Cavanagh, H. Chen, Y. Yue, X. Chen, K. N. Rajagopalan, C. Horth, J. T. McGuire, X. Xu, H. Nikbakht, A. E. Lemiesz, D. M. Marchione, M. R. Marunde, M. J. Meiners, M. A. Cheek, M.-C. Keogh, E. Bareke, A. Djedid, A. S. Harutyunyan, N. Jabado, B. A. Garcia, H. Li, C. D. Allis, J. Majewski, C. Lu, The histone mark H3K36me2 recruits DNMT3A and shapes the intergenic DNA methylation landscape. *Nature* **573**, 281–286 (2019).
79. F. Neri, S. Rapelli, A. Krepelova, D. Incarnato, C. Parlato, G. Basile, M. Maldotti, F. Anselmi, S. Oliviero, Intragenic DNA methylation prevents spurious transcription initiation. *Nature* **543**, 72–77 (2017).
80. A. Monier, A. Pagarete, C. de Vargas, M. J. Allen, B. Read, J.-M. Claverie, H. Ogata, Horizontal gene transfer of an entire metabolic pathway between a eukaryotic alga and its DNA virus. *Genome Res.* **19**, 1441–1449 (2009).
81. S. G. Gornik, K. L. Ford, T. D. Mulhern, A. Bacic, G. I. McFadden, R. F. Waller, Loss of nucleosomal DNA condensation coincides with appearance of a novel nuclear protein in dinoflagellates. *Curr. Biol.* **22**, 2303–2312 (2012).
82. L. M. Iyer, S. Balaji, E. V. Koonin, L. Aravind, Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. *Virus Res.* **117**, 156–184 (2006).
83. M. Boyer, N. Yutin, I. Pagnier, L. Barrassi, G. Fournous, L. Espinosa, C. Robert, S. Azza, S. Sun, M. G. Rossmann, M. Suzan-Monti, B. La Scola, E. V. Koonin, D. Raoult, Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 21848–21853 (2009).
84. D. Moreira, C. Brochier-Armanet, Giant viruses, giant chimeras: The multiple evolutionary histories of Mimivirus genes. *BMC Evol. Biol.* **8**, 12 (2008).
85. A. de Mendoza, J. Pflueger, R. Lister, Capture of a functionally active methyl-CpG binding domain by an arthropod retrotransposon family. *Genome Res.* **29**, 1277–1286 (2019).
86. X. Gao, Y. Hou, H. Ebina, H. L. Levin, D. F. Voytas, Chromodomains direct integration of retrotransposons to heterochromatin. *Genome Res.* **18**, 359–369 (2008).
87. T.-H. S. Hsieh, C. Cattoglio, E. Slobodyanyuk, A. S. Hansen, O. J. Rando, R. Tjian, X. Darzacq, Resolving the 3D landscape of transcription-linked mammalian chromatin folding. *Mol. Cell* **78**, 539–553.e8 (2020).
88. M. Kolmogorov, J. Yuan, Y. Lin, P. A. Pevzner, Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
89. B. J. Walker, T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman, S. K. Young, A. M. Earl, Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLOS ONE* **9**, e112963 (2014).
90. M. Manni, M. R. Berkeley, M. Seppey, F. A. Simão, E. M. Zdobnov, 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 (2021).
91. N. C. Durand, M. S. Shamim, I. Machol, S. S. P. Rao, M. H. Huntley, E. S. Lander, E. L. Aiden, Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
92. O. Dudchenko, S. S. Batra, A. D. Omer, S. K. Nyquist, M. Hoeger, N. C. Durand, M. S. Shamim, I. Machol, E. S. Lander, A. P. Aiden, E. L. Aiden, De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
93. N. C. Durand, J. T. Robinson, M. S. Shamim, I. Machol, J. P. Mesirov, E. S. Lander, E. L. Aiden, Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).
94. F. Cabanettes, C. Klopp, D-GENIES: Dot plot large genomes in an interactive, efficient and simple way. *PeerJ* **6**, e4958 (2018).
95. M. Alonge, L. Lebeigle, M. Kirsche, K. Jenike, S. Ou, S. Aganezov, X. Wang, Z. B. Lippman, M. C. Schatz, S. Soyk, Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol.* **23**, 258 (2022).
96. B.-H. Zhu, J. Xiao, W. Xue, G.-C. Xu, M.-Y. Sun, J.-T. Li, P_RNA_scaffolder: A fast and accurate genome scaffolder using paired-end RNA-sequencing reads. *BMC Genomics* **19**, 175 (2018).
97. J. M. Flynn, R. Hubley, C. Goubert, J. Rosen, A. G. Clark, C. Feschotte, A. F. Smit, RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 9451–9457 (2020).
98. J. Sirén, N. Välimäki, V. Mäkinen, HISAT2—Fast and sensitive alignment against general human population. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **11**, 375–388 (2014).
99. D. Mapleson, L. Venturini, G. Kaithakottil, D. Swarbreck, Efficient and accurate detection of splice junctions from RNA-seq with Portcullis. *Gigascience* **7**, giy131 (2018).
100. B. J. Haas, A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, M. B. Couger, D. Eccles, B. Li, M. Lieber, M. D. Macmanes, M. Ott, J. Orvis, N. Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C. N. Dewey, R. Henschel, R. D. Leduc, N. Friedman, A. Regev, De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
101. M. Stanke, M. Diekhans, R. Baertsch, D. Haussler, Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008).
102. E. Levy Karin, M. Mirdita, J. Söding, MetaEuk-sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. *Microbiome* **8**, 48 (2020).
103. J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G. A. Salazar, E. L. L. Sonnhammer, S. C. E. Tosatto, L. Paladin, S. Raj, L. J. Richardson, R. D. Finn, A. Bateman, Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
104. C. P. Cantalapiedra, A. Hernández-Plaza, I. Letunic, P. Bork, J. Huerta-Cepas, eggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
105. N. L. Bray, H. Pimentel, P. Melsted, L. Pachter, Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
106. Y. Jin, O. H. Tam, E. Paniagua, M. Hammell, TEtranscripts: A package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics* **31**, 3593–3599 (2015).
107. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
108. O. Dudin, A. Ondracka, X. Grau-Bové, A. A. B. Haraldsen, A. Toyoda, H. Suga, J. Bråte, I. Ruiz-Trillo, A unicellular relative of animals generates a layer of polarized cells by actomyosin-dependent cellularization. *eLife* **8**, e49801 (2019).
109. R. Vaisvila, V. K. C. Ponnaluri, Z. Sun, B. W. Langhorst, L. Saleh, S. Guan, N. Dai, M. A. Campbell, B. S. Sexton, K. Marks, M. Samaranayake, J. C. Samuelson, H. E. Church, E. Tamanaha, I. R. Corrêa Jr., S. Pradhan, E. T. Dimalanta, T. C. Evans Jr., L. Williams, T. B. Davis, Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA. *Genome Res.* **31**, 1280–1289 (2021).
110. A. de Mendoza, H. Suga, J. Permanyer, M. Irimia, I. Ruiz-Trillo, Complex transcriptional regulation and independent evolution of fungal-like traits in a relative of animals. *eLife* **4**, 1–5 (2015).
111. W. Guo, P. Fiziev, W. Yan, S. Cokus, X. Sun, M. Q. Zhang, P.-Y. Chen, M. Pellegrini, BS-Seeker2: A versatile aligning pipeline for bisulfite sequencing data. *BMC Genomics* **14**, 774 (2013).
112. W. Guo, P. Zhu, M. Pellegrini, M. Q. Zhang, X. Wang, Z. Ni, CGmapTools improves the precision of heterozygous SNV calls and supports allele-specific methylation detection and visualization in bisulfite-sequencing data. *Bioinformatics* **34**, 381–387 (2018).
113. H. Suga, Z. Chen, A. de Mendoza, A. Sebé-Pedrós, M. W. Brown, E. Kramer, M. Carr, P. Kerner, M. Vervoort, N. Sánchez-Pons, G. Torruella, R. Derelle, G. Manning, B. F. Lang, C. Russ, B. J. Haas, A. J. Roger, C. Nusbaum, I. Ruiz-Trillo, The Capsaspora genome reveals a complex unicellular prehistory of animals. *Nat. Commun.* **4**, 1–9 (2013).
114. N. King, M. J. Westbrook, S. L. Young, A. Kuo, M. Abedin, J. Chapman, S. Fairclough, U. Hellsten, Y. Isogai, I. Letunic, M. Marr, D. Pincus, N. Putnam, A. Rokas, K. J. Wright, R. Zuzow, W. Dirks, M. Good, D. Goodstein, D. Lemons, W. Li, J. B. Lyons, A. Morris, S. Nichols, D. J. Richter, A. Salamov, J. G. I. Sequencing, P. Bork, W. A. Lim, G. Manning, W. T. Miller, W. M. Ginnis, H. Shapiro, R. Tjian, I. V. Grigoriev, D. Rokhsar, The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *Nature* **451**, 783–788 (2008).

Downloaded from https://www.science.org at Universite de Geneve on August 27, 2024

115. S. R. Fairclough, Z. Chen, E. Kramer, Q. Zeng, S. Young, H. M. Robertson, E. Begovic, D. J. Richter, C. Russ, M. J. Westbrook, G. Manning, B. F. Lang, B. Haas, C. Nusbaum, N. King, Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate Salpingoeca rosetta. *Genome Biol.* **14**, R15 (2013).
116. T. Brunet, B. T. Larson, T. A. Linden, M. J. A. Vermeij, K. McDonald, N. King, Light-regulated collective contractility in a multicellular choanoflagellate. *Science* **366**, 326–334 (2019).
117. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
118. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
119. B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler, R. Lanfear, IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
120. D. M. Emms, S. Kelly, OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).

Acknowledgments: We thank C. Martín-Durán, S. Buckberry, M. Greenberg, and F. Marletaz for reading and giving feedback on this manuscript. We thank M. Antó and I. Ruiz-Trillo for sharing the culture and DNA samples, and A. Toyoda at National Institute of Genetics in Japan for technical support for NGS sequencing. We thank B. E. Davies for help with nanopore library construction and the technical staff at the Department of Biology at Queen Mary University of London for their support, especially W. Tyne and G. Mastroianni. This work used computing resources from Queen Mary University of London's Apocrita HPC facilities. **Funding:** This work was supported by European Research Council Starting Grant 950230 (A.d.M., L.A.S., and V.O.), European Research Council Starting Grant 851647 (A.S.-P.), Juan de la Cierva postdoctoral fellowship FJC2020-043131-I (I.V.K.), JSPS KAKENHI Grant Number JP16H06279 (H.S.), JSPS KAKENHI Grant Number JP26891021 (H.S.), Swiss National Science Foundation Ambizione fellowship PZ00P3_185859 (M.O. and O.D.), Spanish Ministry of Science and Innovation grant PID2021-124757NB-I00 (A.S.-P.), and Spanish Ministry's support to the EMBL partnership, the Centro de Excelencia Severo Ochoa and the CERCA Programme (Generalitat de Catalunya) (A.S.-P.). **Author contributions:** A.S.-P.: Resources, funding acquisition, data curation, validation, and supervision. O.D.: Investigation, writing—review and editing, methodology, resources, funding acquisition, data curation, validation, supervision, project administration, and visualization. M.O.: Investigation, methodology, and resources. I.V.K.: Investigation, writing—review and editing, validation, formal analysis, and visualization. L.A.S.: Writing—original draft, conceptualization, investigation, writing—review and editing, methodology, resources, data curation, formal analysis, software, and visualization. A.d.M.: Writing—original draft, conceptualization, writing—review and editing, methodology, resources, funding acquisition, validation, supervision, formal analysis, software, project administration, and visualization. V.O.: Formal analysis. H.S.: Investigation, writing—review and editing, methodology, resources, funding acquisition, and data curation. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** Genome sequencing data have been uploaded to ENA under the project number PRJEB68378, and the assembly accession is GCA_963693365. Functional genomics data can be found in the GEO submission GSE249241. Annotation and other analysis files associated with this article can be found at: https://github.com/AlexdeMendoza/AmoebidiumGenomes, whereas fixed code version is here: https://doi.org/10.5281/zenodo.11208735. A genome browser session with methylation data is available at: https://tinyurl.com/yw7fpe4w. All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials.

Submitted 13 February 2024
Accepted 7 June 2024
Published 12 July 2024
10.1126/sciadv.ado6406

Downloaded from https://www.science.org at Universite de Geneve on August 27, 2024