A cnidarian parasite of salmon (Myxozoa: Henneguya) lacks a mitochondrial genome

Dayana Yahalomi, Stephen D Atkinson, Moran Neuhof, E. Sally Chang, Herve Philippe, Paulyn Cartwright, Jerri L Bartholomew, Dorothée Huchon

Classification: BIOLOGICAL SCIENCES: EVOLUTION

A cnidarian parasite of salmon (Myxozoa: Henneguya) lacks a mitochondrial genome

Dayana Yahalomi[a,1], Stephen D. Atkinson[b,1], Moran Neuhof[c], E. Sally Chang[d,e], Hervé Philippe[f,g], Paulyn Cartwright[d], Jerri L. Bartholomew[b], and Dorothée Huchon[a,h,2]

[a]School of Zoology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv, 6997801 Israel
[b]Department of Microbiology, Oregon State University, Corvallis, OR, 97331
[c]Department of Neurobiology, Biochemistry & Biophysics George S. Wise Faculty of Life Sciences, Tel-Aviv University, Tel-Aviv 6997801 Israel
[d]Department of Ecology and Evolutionary Biology, University of Kansas, 1200 Sunnyside Avenue, Haworth Hall, Lawrence, KS, 66045
[e]Computational and Statistical Genomics Branch, Division of Intramural Research, National Human Genome Research, National Institutes of Health, Bethesda, 20892, MD, USA
[f]CNRS, Station d'Ecologie Expérimentale du CNRS, Moulis, 09200, France
[g]Département de Biochimie, Centre Robert-Cedergren, Université de Montréal, Montréal, QC, Canada H3C 3J7
[h]The Steinhardt Museum of Natural History and National Research Center, Tel Aviv University, Tel Aviv, 6997801 Israel

[1] D.Y. and S.D.A. contributed equally to this work.

[2] Corresponding Author:

Dorothée Huchon
School of Zoology, George S. Wise Faculty of Life Sciences, Tel Aviv University,
Tel Aviv 6997801, Israel
Phone: (+972) 036409817
Email: huchond@tauex.tau.ac.il

Keywords : Myxozoa, Cnidaria, Mitochondrial evolution

Abstract

All known animals share a mitochondrial genome, and perform aerobic respiration. Aerobic respiration is the hallmark of eukaryotes, with only few unicellular lineages having lost this ability. More than 20% of the Cnidaria (corals, jellyfish) are Myxozoa – a parasitic class of ~2,400 species that have a simple body organization. Using deep sequencing approaches, we discovered that the myxozoan *Henneguya salminicola* has no mitochondrial genome, and thus has lost the ability to perform aerobic cellular respiration. This is the first animal found without these core eukaryotic features. We did not find any mitochondria-encoding genes in the genome or transcriptome assemblies of *H. salminicola*, and determined that it has lost almost all nuclear genes involved in transcription and replication of the mitochondrial genome. In contrast, we identified many genes that encode proteins involved in mitochondrial metabolism (e.g., Fe-S cluster synthesis) and determined that genes involved in aerobic respiration or mitochondrial DNA replication were either absent, or present only as pseudogenes. As control, we used the same sequencing and annotation methods to show that the closely related myxozoan, *Myxobolus squamalis*, has a mitochondrial genome. The molecular results are supported by fluorescence micrographs, which show the presence of mitochondrial DNA in *M. squamalis* but not in *H. salminicola*. Our discovery confirms that adaptation to an anaerobic environment is not unique to single-celled eukaryotes but has evolved also in a multicellular, parasitic animal. Hence, *H. salminicola* provides a new opportunity for understanding the evolutionary transition from an aerobic to an exclusive anaerobic metabolism.

Significance

Mitochondrial respiration is an ancient characteristic of eukaryotes. However, it was lost independently in multiple eukaryotic lineages as part of adaptations to an anaerobic lifestyles. We show, that a similar adaptation occurred in a member of the Myxozoa, a large group of microscopic parasitic animals that are closely related to jellyfish and hydras. Using deep sequencing approaches supported by microscopic observations, we present evidence that an animal has lost its mitochondrial genome. The myxozoan cells retain structures deemed mitochondrion-related organelles (MROs), but have lost genes related to aerobic respiration and mitochondrial replication. Our discovery shows that aerobic respiration, one of the most important metabolic pathways, is not omnipresent among animals.

"\body"

Introduction

The acquisition of the mitochondrion was a fundamental event in the evolution of eukaryotes, and most extant eukaryotes cannot survive without oxygen. Interestingly, the loss of aerobic respiration has occurred independently in several eukaryotic lineages that adapted to low-oxygen environments, and replaced the standard mitochondrial (mt) oxidative phosphorylation pathway with novel anaerobic metabolic mechanisms (Fig. 1) (1). Such anaerobic metabolism occurs within mitochondria-related organelles (MROs), which have often lost their cristae, and include hydrogenosomes and mitosomes (1). There is debate regarding the existence of exclusively anaerobic animals and accompanying MROs (2). Although it was reported that some loriciferans found in anoxic conditions possess hydrogenosomes (3, 4), genomic data were not obtained for these samples and alternative explanations have been proposed (2). Herein, we show that a myxozoan parasite (Cnidaria) has lost its mt genome and aerobic metabolic pathways but retains mitosomes.

Myxozoans are a large group of enigmatic, parasitic cnidarians with complex life cycles that require two hosts, usually a fish and an annelid (5). They have substantial negative economic impact on fisheries and aquaculture (6). Myxozoan mitochondria have highly divergent genome structures, with large multipartite circular mt chromosomes and unusually high evolutionary rates (7, 8). To gain further insight into the evolution of the myxozoan mt genome, we studied two closely-related fresh water species, *Henneguya salminicola* and *Myxobolus squamalis* (Fig. S1), both of which are parasites of salmonid fish (9-11).

Results

We assembled transcriptomes and genomes from both species using identical protocols and computational pipelines. Our phylogenetic analyses based on 78 nuclear ribosomal protein-encoding genes from taxa representative of eukaryotic diversity confirmed that the organisms we sequenced are closely related myxozoans, and not contaminants (Fig. 1 and Fig. S2). The genome assembly statistics revealed that *H. salminicola* has a more complete assembly with higher coverage and more predicted protein sequences than *M. squamalis* (Table 1, Fig. S3 and Fig. S4). Targeted searches in the genomes identified 75/78 ribosomal nuclear genes, which suggested that the completeness is >90% for both species. However, estimates of genome completeness using the Core Eukaryotic Genes Mapping Approach (CEGMA) (12) recovered only 53.6% of core eukaryotic genes for *H. salminicola* and 37.5% for *M. squamalis*. We hypothesize that the fast evolutionary rates of myxozoans (13) reduced our ability to detect many common eukaryotic genes – a challenge also known with other fast evolving eukaryotic lineages (14). This view is supported by calculations using only the most conserved CEGMA genes, which have higher recovery in both *H. salminicola* and *M. squamalis* (76.9% and 56.9%, respectively).

Assembly of the mt genomes revealed striking differences between the two parasites. For *M. squamalis*, we successfully recovered a circular mt genome composed of a single chromosome, which phylogenetic analyses confirmed was myxozoan (*SI Results*, Figs. S5 and Fig. S6). Similar to other myxozoans, the *M. squamalis* mt genome lacked tRNAs, and had fast evolutionary rates (*SI Results*, Figs. S5 and Fig. S6). In stark contrast, we could not identify any mt sequence among the contigs of *H. salminicola*,

despite the higher quality of that assembly compared to that of *M*. *squamalis*. To identify if DNA was present in the myxozoan mitochondria, we stained living multi-cellular developing stages of *M*. *squamalis* and *H*. *salminicola* with DAPI (4',6-diamidino-2-phenylindole, blue) (Fig. 2). Cells of *M*. *squamalis* showed the characteristic eukaryotic staining of both nuclei and mitochondria (as much smaller blue dots) *(*Fig. 2A), whereas *H*. *salminicola* showed only nuclear staining (Fig. 2B). This microscopy together with the lack of mt contigs in the genome and transcriptome assemblies supported our central hypothesis that this animal has lost its mt genome. Electron microscopy (EM) images, however, showed mt-like double membrane organelles with cristae in *H*. *salminicola* (Fig. 2C). Together these results confirm that MROs without mt genomes are present in this species.

In animals, most of the mt proteome is encoded in the nucleus. Accordingly, we identified 51 and 57 genes involved in key mt metabolic pathways (e.g. amino acid, carbohydrate or nucleotide metabolism) in *H*. *salminicola* and *M*. *squamalis*, respectively (Fig 3, Table 1, Data table S1). This suggests that the morphologically standard mitochondria of *H*. *salminicola* perform diverse metabolisms, similar to those of *M*. *squamalis*. In contrast, almost all nuclear-encoded proteins involved in mt genome replication and translation were absent from the *H*. *salminicola* genome. Using a database of 118 such nuclear-encoded genes in *Drosophila*, we identified 41-58 homologous mt genes in *M*. *squamalis* and among published myxozoan data (13, 15), but only six of these genes in *H*. *salminicola* (Table 1, Data table S2). In addition, we calculated that *H*. *salminicola* does not have a faster evolutionary rate than other myxozoans, which might otherwise have precluded gene discovery (Fig. 1 and Fig. S2).

Interestingly, in *H. salminicola*, we found that the mt DNA polymerase subunit gamma-1 (16) gene is a pseudogene, as it contains three point mutations that create premature stop codons (Fig. S7). Furthermore, this gene was not found to be expressed in *H. salminicola*. It was absent from *H. salminicola* transcriptome assembly, despite identifying homologous contigs in all other myxozoan transcriptomes (Data table S2). The presence of a pseudogene copy of the DNA polymerase subunit gamma-1 has several implications. First, it supports our central conclusion that *H. salminicola* has lost its mtDNA, as it has no mtDNA replication machinery. Second, it shows that the absence of protein homologues in this species is the result of pseudogeneization and not an assembly artefact. Finally, it indicates that the loss of mtDNA is a recent event, since the pseudogene can still be identified in the genome.

The loss of the mt genome should impact aerobic respiration, since animal mt genomes code for essential proteins of the electron-transport chain (17). To verify whether the loss of the mt genome meant loss of aerobic respiration in *H. salminicola*, we searched for homologues of known *Drosophila* nuclear genes, which are typically encoded by ~100 proteins from the mt electron-transport chain complexes (Fig. 3, *SI Methods*). Our searches of all myxozoan genomes available revealed that nuclear genes for only 7 of these mt proteins remain in *H. salminicola*, whereas 18-25 are present in other myxozoans (Fig. 3, Table 1; Data table S1). Specifically, all complex I, III, and IV genes that we identified in myxozoans are absent in *H. salminicola* (Fig 3, *SI Results*, Fig. S8, Data table S1) or present as pseudogenes (Fig. S7). Since complex IV interacts with O_2 molecules, we conclude that *H. salminicola* might not be capable of standard cellular aerobic respiration. In contrast, for both complex II, which is part of the Krebs

cycle, and the ATP synthase complex, *H. salminicola* encode the same number of protein coding genes as other myxozoans. This suggests that a proton gradient is still present across the inner organelle membrane in *H. salminicola* that we observed in EM.

Discussion

Structurally, *H. salminicola* has lost its mt genome but has retained an organelle that resembles a mitochondrion. However, as mitochondria are defined based on the presence of a mt genome (1), we conclude that *H. salminicola* possesses MROs rather than true mitochondria. Further, as our *H. salminicola* assemblies did not contain any hydrogenase or genes of prokaryotic origin (Data table S3), as found in some anaerobic protists, we conclude that the MROs present in *H. salminicola*, are mitosomes not hydrogenosomes.

Similar to most Myxozoa, *H. salminicola* likely alternate between two hosts (5). In its fish host, it undergoes proliferation and sporogenesis in pseudocysts within the white muscle (10), a tissue known to have anaerobic metabolism (18). While the obligate invertebrate host of *H. salminicola* is unknown, it is probably an oligochaete worm from the family Naididae, based on known life cycles of myxozoans in the same family (19). Members of the Naididae can grow and reproduce in anoxic environments (20). As all protists that have lost their mt genomes live in anaerobic environments, we speculate that the loss of the mt genome in *H. salminicola* was driven by low-oxygen environments in both of its hosts.

Loss of superfluous genes likely conveys an evolutionary advantage, as it has been shown that the bioenergetic cost of a gene is higher in small genomes (21). Myxozoans have smaller genomes (22-180 Mb (13, 15)) than free-living Cnidaria (>250

Mb (22, 23)). Therefore, the loss of the mt genome and associated nuclear genes involved in its replication and electron pathways may be advantageous for a myxozoan living in anaerobic environments. However, the loss of useless genes by random drift cannot be excluded. Interestingly, our results also open the way to new treatment options against this pathogen since anaerobic protists are known to be sensitive to specific drugs (24).

Myxozoans have gone through outstanding morphological and genomic simplifications during their adaptation to parasitism from a free-living cnidarian ancestor (25). It is remarkable that these myxozoan simplifications do not appear to be ancestral, but rather the result of secondary losses (13). Here we show that at least one myxozoan species has lost a core animal feature: the genetic basis for aerobic respiration in its mitochondria. As a highly diverse group with over 2,400 species, which inhabit marine, freshwater, and even terrestrial environments (26), evolutionary loss and simplification has clearly been a successful strategy for Myxozoa, which shows that less is more (27).

Materials and Methods

Samples and sequencing

Samples of *H*. *salminicola* and *M*. *squamalis* were identified based on spore morphology (Fig. S1), tissue tropism, host, and 18S rDNA sequence similarity with published sequence available at the National Center for Biotechnology Information (NCBI) (*SI Methods*).

DNA and RNA of *H*. *salminicola* were extracted from a single cyst sampled from Chinook salmon (*Oncorhynchus tschawytscha*). Conversely, because of a small cyst size, DNA and RNA of *M*. *squamalis* were extracted from several cysts collected from a single

coho salmon (*Oncorhynchus kisutch*). The multi-isolate extract from *M. squamalis* may explain the differences in assembly quality between *H. salminicola* and *M. squamalis* since polymorphism is known to complicate assembly.

DNA and RNA were extracted with the DNeasy Blood & Tissue Kit (Qiagen, Germantown, MD) and the High Pure RNA extraction kit (Roche, Pleasanton, CA), respectively, following manufacturer instructions. The samples were sent for library construction and sequencing at the Center for Genome Research and Biocomputing (CGRB) at Oregon State University (Corvalis, OR). Paired-end sequencing of 150 bp reads derived from fragments of average length of ~350 bp was performed on a HiSeq3000 platform.

Light microscopy

Myxozoan cells were prefixed with 3:1 methanol:acetic acid, and 1-3 drops of cell suspension were put on slides. DNA staining was performed with VECTASHIELD (Vectorlabs, Burlingame, CA) antifade mounting medium, which contained DAPI (see *SI Methods*). Cells were visualized under a Leica DMR compound fluorescence microscope at 630x and 1000x magnification.

Electron microscopy

Fresh parasite pseudocysts were dissected from the tissue of a single host then fixed in a solution comprising 1% glutaraldehyde, 2% paraformaldehyde in 0.1 M phosphate buffer. Larger pseudocysts were sliced to permit penetration of the fixative. Fixed tissue was then stained with osmium tetroxide, then uranyl acetate, before being dehydrated in a

graded alcohol series and embedded in Epon resin. Ultra-thin sections were mounted on copper grids and examined using a Helios 650 FEG dual-beam SEM (Thermo Scientific) in transmission mode, at the Oregon State University Electron Microscope Facility.

Filtering and assembling the genomic data

Stringent filtering, involving multiple steps, was performed to eliminate host and bacterial contamination from *M*. *squamalis* and *H*. *salminicola* data. We first removed adaptor sequences in the raw reads using the FastQC (28) and CutAdapt (29) programs, and filtered fish contamination from the processed reads by mapping our raw reads to the corresponding host genomes downloaded from NCBI (see *SI Methods* for details) using bowtie2 (30). Reads that mapped to the fish genome assemblies were discarded, and the remaining reads of each myxozoan were assembled using IDBA (version 1.1.1) (31). Further filtering was then performed using BLAST searches. Specifically, the assembled contigs were used as query in blastn searches (local version 2.7.0) against the NCBI nucleotide database (last accessed: 2018/08) to detect bacterial and eukaryotic contamination (see *SI Methods* for details). Reads that mapped to contaminant contigs were identified using bowtie2 and discarded. After this second filtering step, each genome was then reassembled from the remaining reads with IDBA. From these assemblies, we discarded contigs <500bp and removed duplicates using cd-hit (32) (see *SI Methods* for details). Because fish contaminations were still detected in the *M*. *squamalis* DNA assembly, all contigs presenting significant blastn and blastx hits against NCBI teleost sequences were removed. Detailed information about assembly and filtering is provided in *SI Methods*.

Filtering and assembling the transcriptomic data

The overall quality- and contamination-filtering approach for the transcriptome data was similar to that used for the DNA assemblies. Adapter sequences were removed using FastQC (28) and CutAdapt (29). To filter reads from fish contamination we used bowtie2 (30) to map raw reads to representative fish transcriptome assemblies downloaded from NCBI (last accessed: 2018/08) (see *SI Methods* for details). Reads that mapped to these fish assemblies were discarded from further analyses. The remaining reads from the *Henneguya* and *Myxobolus* data sets were each assembled using Trinity (33). After assembly, we discarded contigs <300bp. Each assembly was further filtered using three steps: 1) Because RNA contigs were heavily host-contaminated, as observed in previous work on Myxozoa (13, 15), blastn searches were performed with the RNA contigs as queries against the corresponding DNA assembly to eliminated RNA contigs without any hit; 2) transcript abundance was determined with the RSEM script included in the Trinity package (33) and low abundance transcripts (--fpkm_cutoff=0.01 --isopct_cutoff=1.00) were discarded since most were found to be contaminants; and 3) a second blastn search was performed against the NCBI nucleotide database and contigs with >95% sequence identity to non-myxozoan sequences were discarded. The *Myxobolus* transcriptome was further filtered by discarding all contigs that shared >80% sequence identity with fish sequences present in NCBI (last accessed: 2018/08). Finally, for each transcript cluster reconstructed by Trinity, only the longest transcript was kept and used to build the unigene assemblies of each species, which were then used in downstream analyses (see *SI Methods* for details).

Assembly of the *M. squamalis* mt genome and absence of mt sequences in *H. salminicola*.

Local blastn and tblastn searches using cnidarian (including published myxozoan) mt genome and protein sequences, respectively, were performed against our myxozoan assemblies of *M*. *squamalis* and *H*. *salminicola*. After manual inspection of all sequences with E-values < $1e^{-1}$ no mitochondrial sequence could be identified for *H*. *salminicola*. In contrast, a putative mitochondrial contig was identified for *M*. *squamalis*. The coverage of the mt genome was 185, about twice the nuclear coverage. To further search the *H*. *salminicola* data, Hidden Markov Model (HMM) profiles were built based on alignments of myxozoan mt proteins using HMMer3.0 (34). These profiles were used to search protein predictions of *H*. *salminicola* by Maker2 v2.31.10 (35) (see *SI Methods* for details regarding Maker2 annotation), but no mt proteins were identified. Similarly, HMM profiles were built based on alignments of myxozoan RNA sequences using Infernal 1.1.1 (36), following the approach of Yahalomi et al. (7), and the profiles used to search genomic and transcriptomic assemblies. Again, no mt sequence was identified.

The Perl script Novoplasty v2.6.3 (37) was used to reconstruct a first draft of the mitochondrial sequence of *M*. *squamalis* based on the mt contig identified in the BLAST search. The draft sequence was then corrected using read mapping (see *SI Methods* for details regarding the assembly and annotation of the mt genome of *M*. *squamalis*).

Estimating completeness of genomic and transcriptomic assemblies

The CEGMA program (12) was used to estimate the completeness of our assemblies. Because myxozoans show extreme evolutionary rates (13) the completeness was

estimated based on the most conserved set of CEGMA genes (Group 4). We also used the program BUSCO V3 (38), but found that it performed poorly on Myxozoa, which was in concordance with other studies that show that BUSCO underestimates completeness of fast-evolving organisms (14).

Genome size estimation

Genome sizes were calculated based on K-mer frequency estimation using the GenomeScope web server (39) (last accessed 2018/02). For both species k-mer frequency histograms were generated for k=17 using the program Jellyfish 2.2.7 (40) on the filter reads with the following parameters: count -C -m 17 -s 1000000000 -t 10 (Fig. S3 and Fig. S4).

Characterization of myxozoan proteins interacting with mtDNA and mtRNA

To create an exhaustive database of nuclear-encoded proteins that interact with mt DNA and RNA, we downloaded three protein datasets: all mt ribosomal protein sequences from FlyBase (last accessed 2017/11) (41); all *Drosophila melanogaster* proteins with either the functional classifications “DNA and RNA” or “DNA and RNA/Protein synthesis/Others” from the MitoDrome database (last accessed 2018/01) (42); sequences of human proteins known to bind to mt RNA and described in Rackam et al. (43) from NCBI. We then performed reciprocal blastp searches against the *Drosophila* proteome to identify the corresponding homologues (see *SI Methods* for details). These sets of *Drosophila* proteins were used to perform reciprocal BLAST searches against the proteome of the cnidarian *Hydra vulgaris*. The *Hydra* and *Drosophila* sequences were

then used to identify homologous sequences in the myxozoan genome and transcriptome assemblies, after they had been filtered from contaminants. Detailed information about homologue identification is provided in *SI Methods*.

Characterization of myxozoan mitochondrial metabolic pathways

Drosophila proteins involved in the different mitochondrial metabolic pathways were downloaded from the MitoDrome database (42). As described above, reciprocal BLASTp searches were conducted using the *Drosophila* sequences as queries to identify homologous copies in *Hydra*. The *Drosophila* and *Hydra* sequences were then used as queries to identify nuclear encoded mitochondrial proteins in the myxozoans (see *SI methods* for details). It is worth noting that all *Kudoa* proteins previously identified by Muthye and Lavrov (44) using HMM profiles were identified also in our reciprocal BLAST searches, indicating that the use of HMM profiles did not improve protein identifications in our case.

Phylogenetic reconstructions

The phylogenetic analyses used a reference database of 78 ribosomal protein coding genes, which has been curated manually to avoid contamination and structural annotation errors. Two datasets were selected from this database, the first included 78 species representative of major eukaryote lineages (45); the second included 129 species that encompassed animal diversity and their closest outgroup (choanoflagellates, ichthyosporean and ministeriid). In both datasets, sequences were concatenated with SCaFoS (46). After removal of any ambiguously aligned positions using Gblocks Version

0.91b (47), with default parameter values, these Eukaryote and Metazoa datasets included 9,490 and 11,352 amino acid positions, respectively. Phylogenetic reconstructions were performed using the site-heterogeneous CAT model (48), which reduces the impact of long branch attraction (49), as implemented in Phylobayes MPI vs.1.5 (50). For both datasets, two independent chains were run for 10,000 cycles. The first 5,000 trees from each chain were discarded as burn-in. Chain convergence was assessed using the bpcomp and tracecomp scripts which are part of Phylobayes. Specifically, for both analyses the bpcomp maxdiff values were <0.3 and the tracecomp effsize values were >70 (except for eukaryotes where the tree length value was 21), indicating a proper convergence.

Acknowledgements: We thank Mark Dasenko and the Center for Genome Research and Biocomputing at OSU, and Teresa Sawyer of the OSU Electron Microscope Facility, for their assistance. This work was supported by the Binational Science Foundation (Grant No. 2015010 to D.H. and P.C.).

Authors Contributions: D.H., S.A., J.B. and P.C. conceived the study, participated in its design and coordination. S.A. identified the myxozoan samples, extracted the DNA and RNA, and performed LM and EM microscopy. D.Y. filtered the DNA and RNA reads against contaminant assemblies, performed the assemblies and conducted the searches of mitochondrial homologues. D.H. preformed preliminary analyses and the analysis of the mt genome of *M. squamalis*. M.N. performed the Trinotate and KEGG analyses and submitted all data to NCBI. S.C. performed the MAKER2 analyses. HP performed the phylogenetic analyses. D.H. and D.Y. drafted the

manuscript and draw the figures. All other authors assisted in revising the manuscript. All authors read and approved the final manuscript.

Data deposition: All sequence data have been deposited in the National Center for Biotechnology Information (NCBI) database. The *Henneguya salminicola* data are available under the BioProject accession number PRJNA485580. The complete 18S rRNA sequence was deposited under MK480607. The raw transcriptome and genome reads are available under accession numbers SRR7754566 and SRR7754567, respectively. The transcriptome and genome shotgun assembly projects were deposited at under the accession GHBP00000000 and SGJC00000000, respectively. The versions described in this paper are the first versions, GHBP01000000 and SGJC01000000, respectively. The *Myxobolus squamalis* data are available under the BioProject accession number PRJNA485581. The complete 18S rRNA sequence was deposited under MK480606. The raw transcriptome and genome reads are available under the accession numbers SRR7760054 and SRR7760053, respectively. The transcriptome and genome shotgun assembly projects were deposited at under the accession GHBR00000000 and QWKW00000000, respectively. The versions described in this paper are the first versions, GHBR01000000 and QWKW01000000, respectively. The mt genome of *M*. *squamalis* was deposited under the accession number MK 087050. The Bayesian trees and all alignments were deposited in the TreeBASE repository (http://purl.org/phylo/treebase/phylows/study/TB2:S23827?x-access-code=e9bba2fefab72915b2e4ee97923768d7&format=html). Finally, the uncropped pictures underlying Fig. 2 were deposited in the Figshare repository under the provisional link https://figshare.com/s/276f2bb28e0b77e15a98.

References

1. A. J. Roger, S. A. Muñoz-Gómez, R. Kamikawa, The origin and diversification of mitochondria. *Curr*. *Biol*. **27**, R1177-R1192 (2017).
2. J. M. Bernhard *et al*., Metazoans of redoxcline sediments in Mediterranean deep-sea hypersaline anoxic basins. *BMC Biol*. **13**, 105 (2015).
3. R. Danovaro *et al*., The first metazoa living in permanently anoxic conditions. *BMC Biol*. **8**, 30 (2010).
4. R. Danovaro *et al*., The challenge of proving the existence of metazoan life in permanently anoxic deep-sea sediments. *BMC Biol*. **14**, 43 (2016).
5. B. Okamura, A. Gruhl, J. L. Bartholomew, "An introduction to myxozoan evolution, ecology and development" in *Myxozoan evolution, ecology and development,* B. Okamura, A. Gruhl, J. L. Bartholomew, Eds. (Springer International Publishing, 2015), pp. 1-20.
6. I. Fontes, S. L. Hallett, T. A. Mo, "Comparative epidemiology of myxozoan diseases" in *Myxozoan evolution, ecology and development,* B. Okamura, A. Gruhl, J. L. Bartholomew, Eds. (Springer International Publishing, 2015), pp. 317-341.
7. D. Yahalomi *et al*., The multipartite mitochondrial genome of *Enteromyxum leei* (Myxozoa): eight fast-evolving megacircles. *Mol*. *Biol*. *Evol*. **34**, 1551-1556 (2017).
8. F. Takeuchi *et al*., The mitochondrial genomes of a myxozoan genus *Kudoa* are extremely divergent in Metazoa. *PLoS ONE* **10**, e0132030 (2015).
9. I. Fiala, P. Bartošová-Sojková, C. M. Whipps, "Classification and phylogenetics of Myxozoa" in *Myxozoan evolution, ecology and development,* B. Okamura, A. Gruhl, J. L. Bartholomew, Eds. (Springer International Publishing, 2015), pp. 85-110.
10. F. F. Fish, Observations on *Henneguya salminicola* Ward, a myxosporidian parasitic in Pacific salmon. *J*. *Parasitol*. **25**, 169-172 (1939).
11. T. M. Polley, S. D. Atkinson, G. R. Jones, J. L. Bartholomew, Supplemental description of *Myxobolus squamalis* (Myxozoa). *J*. *Parasitol*. **99**, 725-728 (2013).

12. G. Parra, K. Bradnam, I. Korf, CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061-1067 (2007).
13. S. E. Chang *et al*., Genomic insights into the evolutionary origin of Myxozoa within Cnidaria. *Proc*. *Natl*. *Acad*. *Sci*. *U.S.A*. **112**, 14912–14917 (2015).
14. A. Karnkowska *et al*., A eukaryote without a mitochondrial organelle. *Curr*. *Biol*. **26**, 1274-1284 (2016).
15. Y. Yang *et al*., The genome of the myxosporean *Thelohanellus kitauei* shows adaptations to nutrient acquisition within its fish host. *Genome Biol*. *Evol*. **6**, 3182-3198 (2014).
16. M. A. Graziewicz, M. J. Longley, W. C. Copeland, DNA polymerase γ in mitochondrial DNA replication and repair. *Chem*. *Rev*. **106**, 383-405 (2006).
17. D. V. Lavrov, W. Pett, Animal mitochondrial DNA as we do not know it: mt-genome organization and evolution in nonbilaterian lineages. *Genome Biol*. *Evol*. **8**, 2896-2913 (2016).
18. I. A. Johnston, Studies on the swimming musculature of the rainbow trout. *J*. *Fish Biol*. **7**, 459-467 (1975).
19. J. D. Alexander, B. L. Kerans, M. El-Matbouli, S. L. Hallett, L. Stevens, "Annelid-myxosporean interactions" in *Myxozoan evolution, ecology and development,* B. Okamura, A. Gruhl, J. L. Bartholomew, Eds. (Springer International Publishing, 2015), pp. 217-234.
20. P. Famme, J. Knudsen, Anoxic survival, growth and reproduction by the freshwater annelid, *Tubifex* sp., demonstrated using a new simple anoxic chemostat. *Comp*. *Biochem*. *Physiol*. *A* **81**, 251-253 (1985).
21. M. Lynch, G. K. Marinov, The bioenergetic costs of a gene. *Proc*. *Natl*. *Acad*. *Sci*. *U.S.A*. **112**, 15690-15695 (2015).
22. J. A. Chapman *et al*., The dynamic genome of *Hydra*. *Nature* **464**, 592-596 (2010).
23. N. H. Putnam *et al*., Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* **317**, 86-94 (2007).
24. C. Edlund, S. Löfmark, C. E. Nord, Metronidazole is still the drug of choice for treatment of anaerobic infections. *Clin*. *Infect*. *Dis*. **50**, S16-S23 (2010).

25. E. Kayal *et al*., Phylogenomics provides a robust topology of the major cnidarian lineages and insights on the origins of key organismal traits. *BMC Evol*. *Biol*. **18**, 68 (2018).
26. S. D. Atkinson, J. L. Bartholomew, T. Lotan, Myxozoans: Ancient metazoan parasites find a home in phylum Cnidaria. *Zoology* **129**, 66-68 (2018).
27. M. V. Olson, When less is more: gene loss as an engine of evolutionary change. *Am*. *J*. *Hum*. *Genet*. **64**, 18-23 (1999).
28. S. Andrews, *FastQC: a quality control tool for high throughput sequence data*. *Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc* (2010).
29. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10-12 (2011).
30. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat*. *Methods* **9**, 357-359 (2012).
31. Y. Peng, H. C. M. Leung, S. M. Yiu, F. Y. L. Chin, IDBA – A practical iterative de Bruijn graph *de novo* assembler. In: Berger B. (ed) Research in Computational Molecular Biology. RECOMB 2010. *Lect*. *Notes Comput*. *Sci*. **6044**, 426-440 (2010).
32. W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659 (2006).
33. B. J. Haas *et al*., De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat*. *Protocols* **8**, 1494-1512 (2013).
34. J. Mistry, R. D. Finn, S. R. Eddy, A. Bateman, M. Punta, Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res*. **41**, e121 (2013).
35. C. Holt, M. Yandell, MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
36. E. P. Nawrocki, S. R. Eddy, Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933-2935 (2013).

37. N. Dierckxsens, P. Mardulyn, G. Smits, NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res*. **45**, e18 (2017).
38. R. M. Waterhouse *et al*., BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol*. *Biol*. *Evol*. **35**, 543–548 (2018).
39. G. W. Vurture *et al*., GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202-2204 (2017).
40. C. Kingsford, G. Marçais, A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* **27**, 764-770 (2011).
41. L. S. Gramates *et al*., FlyBase at 25: looking to the future. *Nucleic Acids Res*. **45**, D663-D671 (2017).
42. D. D'Elia *et al*., The MitoDrome database annotates and compares the OXPHOS nuclear genes of *Drosophila melanogaster*, *Drosophila pseudoobscura* and *Anopheles gambiae*. *Mitochondrion* **6**, 252-257 (2006).
43. O. Rackham, T. R. Mercer, A. Filipovska, The human mitochondrial transcriptome and the RNA-binding proteins that regulate its expression. *Wiley Interdiscip Rev RNA* **3**, 675-695 (2012).
44. V. Muthye, D. V. Lavrov, Characterization of mitochondrial proteomes of nonbilaterian animals. *IUBMB Life* **70**, 1289-1301 (2018).
45. S. M. Adl *et al*., Revisions to the classification, nomenclature, and diversity of eukaryotes. *J*. *Eukaryot*. *Microbiol*. **26**, 12691 (2018).
46. B. Roure, N. Rodriguez-Ezpeleta, H. Philippe, SCaFoS: a tool for Selection, Concatenation and Fusion of Sequences for phylogenomics. *BMC Evol*. *Biol*. **7**, S2 (2007).
47. J. Castresana, Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol*. *Biol*. *Evol*. **17**, 540-552 (2000).
48. N. Lartillot, H. Philippe, A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol*. *Biol*. *Evol*. **21**, 1095-1109 (2004).
49. N. Lartillot, H. Brinkmann, H. Philippe, Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol*. *Biol*. **7**, S4 (2007).

50. N. Lartillot, N. Rodrigue, D. Stubbs, J. Richer, PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst*. *Biol*. **62**, 611-615 (2013).

Figure legends

Fig. 1: Eukaryote phylogenetic relationships inferred from a supermatrix of 9490 amino acid positions for 78 species. Bayesian majority-rule consensus tree reconstructed using the CAT + Γ model from two independent Markov-chain Monte Carlo chains. Branches with low node support (posterior probabilities PP<0.7) were collapsed. Most nodes were highly supported (PP>0.98) and PP are only indicated for nodes with PP<0.98. The Eukaryote classification is based on Adl et al. (45). Species known to have lost their mt genome are indicated in bold with an asterisk. Myxozoan species form a well-supported group (PP=1.0) and our reconstructions agree with previous studies (13), which show monophyly of the fresh-water/oligochaete host lineage (9).

Fig. 2: Microscopic evidence for the absence of mitochondria in *H. salminicola*. A, B, DAPI staining of normal 7-cell pre-sporogonic developmental stages of two myxozoan parasites of salmonid fish. (**A)** *Myxobolus squamalis*, showing large nuclei with many smaller mitochondrial nucleosomes (arrowed). (**B)** *Henneguya salminicola*, showing large nuclei but surprisingly no mitochondrial nucleosomes. **C**, TEM image of *H. salminicola* mitochondrion-related organelle with few cristae. Uncropped images are available in the Figshare repository.

Fig 3 – Comparison between the pathways present in the myxozoans (A) *Kudoa iwatai* mitochondrion and (B) *Henneguya salminicola* mitochondrion-related organelle. The presence and absence of organellar genomes are indicated. DNA pol. MtDNA polymerase; RNA pol; mtDNA-dependent RNA polymerase; CI-CV, respiratory complexes I-V; C, cytochrome c; PD: Pyruvate dehydrogenase; UQ, ubiquinone; e^-, electrons; H^+, protons; ψ indicates the presence of a pseudogene in the nuclear genome.

Table 1. Assembly statistics, presence of mt genome and number of nuclear-encoded mt genes identified for myxozoan genomes (gen.) and transcriptomes (trans.).

	H. salminicola	*M. squamalis*	*K. iwatai (13)*	*T. kitauei (15)*	*M. cerebralis (13)*
	Gen. / Trans.	Gen. / Trans.	Gen. / Trans.	Gen.	Trans.
Genome size MB	60.0 / -	53.1 / -	22.5 / -	188.5	-
Coverage	311 / -	86.1 / -	1,000 / -	37	-
DNA assembly size MB	61.4 / -	43.7 / -	23.7 / -	150.7	-
Number of contigs	18,330 / 31,825	37,919 / 11,236	22,174 / 6,528	5,610	52,821
N50	7,570 / 600	1,286 / 714	40,195 / 1,662	-	11,965
CEGs (complete)	53.6% / 26.6%	37.5% / 23.8%	73.0% / 76.6%	46.8%	39.1%
CEGs (complete group4)	76.9% / 33.9%	56.9% / 27.7%	96.9% / 95.4%	66.2%	55.4%
CEGs (partial group 4)	87.7% / 75.4%	76.9% / 70.8%	96.9% / 96.9%	73.9%	84.6%
%GC	29 / -	27 / -	23.6 / -	37.5	-
# predicted proteins	8,188 / -	5,725 / -	5,533 / -	16,638	-
Presence/absence of mt genome	**Absence**	Presence	Presence (7, 8)	Presence (7)	Unknown
# nuclear genes involved in mtDNA replication and translation	6	58	52	41	49
# nuclear genes involved in electron-transport chains	7	21	25	18	21
# genes involved in pyruvate metabolism	0*	3*	3*	3*	3*
# genes involved in other mt pathways	52	55	64	45	46

* There are two additional proteins, present in all Myxozoa, that appear also in other metabolic pathways.

Fig. 1

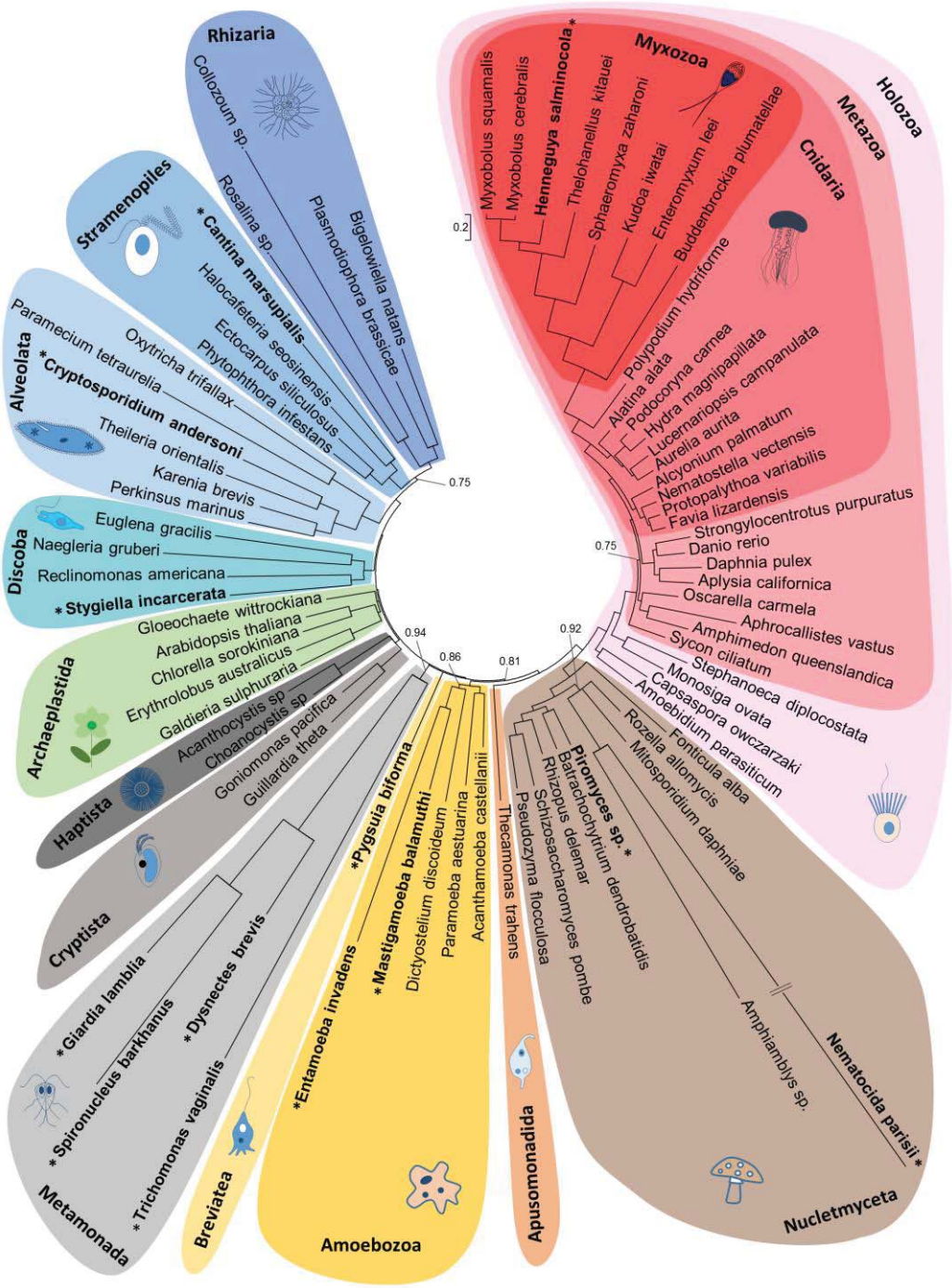

Rhizaria
Collozoum sp.
Rosalina sp.
Plasmodiophora brassicae
Bigelowiella natans
Stramenopiles
*Cantina marsupialis
Halocafeteria seosinensis
Ectocarpus siliculosus
Phytophthora infestans
Alveolata
Paramecium tetraurelia
Oxytricha trifallax
*Cryptosporidium andersoni
Theileria orientalis
Karenia brevis
Perkinsus marinus
Discoba
Euglena gracilis
Naegleria gruberi
Reclinomonas americana
*Stygiella incarcerata
Archaeplastida
Gloeochaete wittrockiana
Arabidopsis thaliana
Chlorella sorokiniana
Erythrolobus australicus
Galdieria sulphuraria
Haptista
Acanthocystis sp.
Choanocystis sp.
Goniomonas pacifica
Guillardia theta
Cryptista
*Giardia lamblia
*Spironucleus barkhanus
*Dysnectes brevis
*Trichomonas vaginalis
Metamonada
Breviatea
*Entamoeba invadens
*Pygsuia biforma
*Mastigamoeba balamuthi
Dictyostelium discoideum
Paramoeba aestuarina
Acanthamoeba castellanii
Amoebozoa
Thecamonas trahens
Apusomonadida
Myxozoa
Myxobolus squamalis
Myxobolus cerebralis
Henneguya salminocola*
Thelohanellus kitauei
Sphaeromyxa zaharoni
Kudoa iwatai
Enteromyxum leei
Buddenbrockia plumatellae
Holozoa
Metazoa
Cnidaria
Polypodium hydriforme
Alatina alata
Podocoryna carnea
Hydra magnipapillata
Lucernariopsis campanulata
Aurelia aurita
Alcyonium palmatum
Nematostella vectensis
Protopalythoa variabilis
Favia lizardensis
Strongylocentrotus purpuratus
Danio rerio
Daphnia pulex
Aplysia californica
Oscarella carmela
Aphrocallistes vastus
Amphimedon queenslandica
Sycon ciliatum
Stephanoeca diplocostata
Monosiga ovata
Capsaspora owczarzaki
Amoebidium parasiticum
Rozella allomycis
Fonticula alba
Mitosporidium daphniae
Piromyces sp.*
Batrachochytrium dendrobatidis
Rhizopus delemar
Schizosaccharomyces pombe
Pseudozyma flocculosa
Amphiamblys sp.
Nematocida parisii*
Nucletmyceta
0.2
0.75
0.75
0.92
0.94
0.86
0.81

Fig. 2.

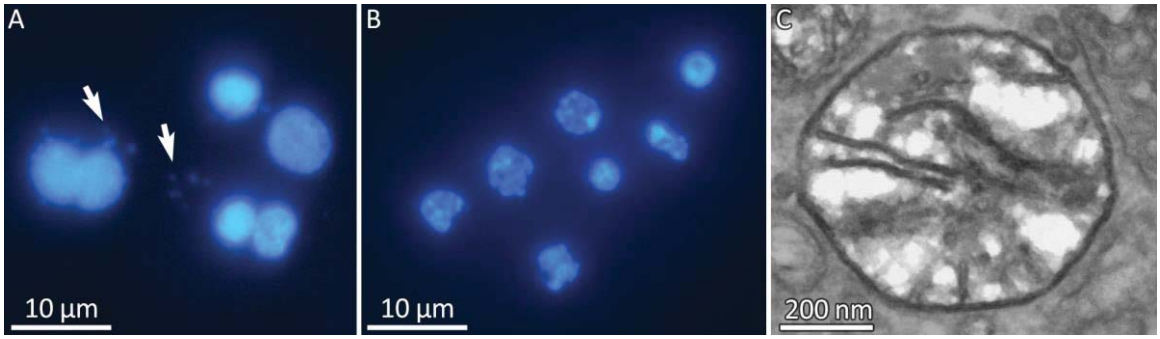
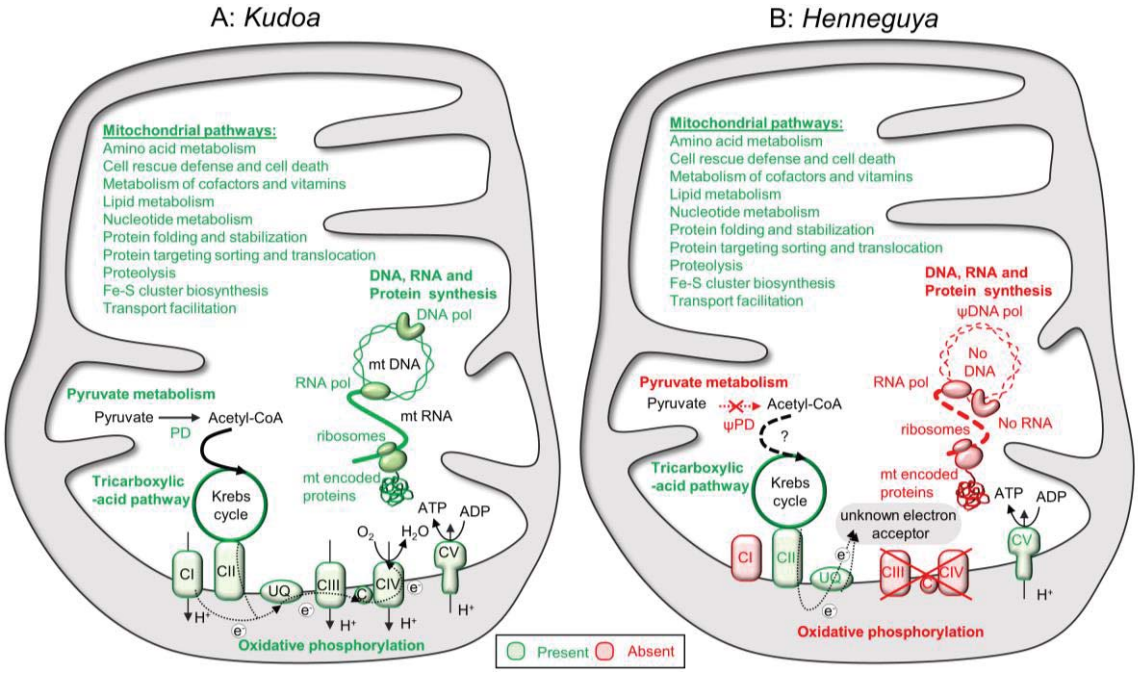


Fig 3.


A: Kudoa
Mitochondrial pathways:
Amino acid metabolism
Cell rescue defense and cell death
Metabolism of cofactors and vitamins
Lipid metabolism
Nucleotide metabolism
Protein folding and stabilization
Protein targeting sorting and translocation
Proteolysis
Fe-S cluster biosynthesis
Transport facilitation
DNA, RNA and Protein synthesis
DNA pol
mt DNA
RNA pol
mt RNA
ribosomes
mt encoded proteins
Pyruvate metabolism
Pyruvate
PD
Acetyl-CoA
Tricarboxylic -acid pathway
Krebs cycle
CI
CII
UQ
CIII
CIV
CV
ATP
ADP
O_2
H_2O
H^+
e
Oxidative phosphorylation
B: Henneguya
Mitochondrial pathways:
Amino acid metabolism
Cell rescue defense and cell death
Metabolism of cofactors and vitamins
Lipid metabolism
Nucleotide metabolism
Protein folding and stabilization
Protein targeting sorting and translocation
Proteolysis
Fe-S cluster biosynthesis
Transport facilitation
DNA, RNA and Protein synthesis
ψDNA pol
No DNA
RNA pol
No RNA
ribosomes
mt encoded proteins
Pyruvate metabolism
Pyruvate
ψPD
?
Acetyl-CoA
Tricarboxylic -acid pathway
Krebs cycle
unknown electron acceptor
CI
CII
UQ
CIII
CIV
CV
ATP
ADP
H^+
e
Oxidative phosphorylation
Present
Absent

PNAS
www.pnas.org

Supplementary Information for

An animal without aerobic cellular respiration

Dayana Yahalomi, Stephen D. Atkinson, Moran Neuhof, E. Sally Chang, Hervé Philippe, Paulyn Cartwright, Jerri L. Bartholomew, and Dorothée Huchon

Corresponding author: Dorothée Huchon
Email: huchond@tauex.tau.ac.il

This PDF file includes:

Supplementary text:

Supplementary Information Text

SUPPLEMENTARY METHODS

1- Sample identification

Parasites were identified initially by their fish host, tissue tropism of spore development and myxospore morphology (given that there are no other described myxozoans with these characters). For both species we initially Sanger-sequenced up to 2,000 bp of the taxonomically informative SSU rDNA (small subunit ribosomal RNA gene) region to characterize the isolates we used for genome/transcriptome and microscopy studies.

Host fish were killed by overdose of Tricaine methanesulfonate (MS222) then held on ice until necropsied within 24h. DNA and RNA were extracted from cysts with the DNeasy Blood & Tissue Kit (Qiagen, Germantown, MD) and the High Pure RNA extraction kit (Roche, Pleasanton, CA), respectively, following manufacturer instructions. The 5'-end of the SSU rDNA were amplified using primer 18e (1) with primer ACT1r (2). The 3'-end of the SSU rDNA were amplified using primers Myxgen4f (3) with primer 18r (4). PCRs were performed in 20uL reactions that comprised: 2 uL extracted DNA, 1.5 mM $MgCl_2$, 0.2 mM dNTPs, 2 u GoTaqFlexi polymerase (Promega, San Luis Obispo, CA, USA), 1X GoTaqFlexi clear buffer, 0.4 μM of each primer, 0.25 μM Bovine serum albumin and 0.5X Rediload loading dye (Invitrogen). PCR was performed on a PTC-200 (MJ Research) thermocycler with the following conditions: denaturation at 95 °C for 3 min, 30 cycles of 94 °C for 30s, 55 °C for 30s, 72 °C for 90s, before a terminal extension of 7 min at 72 °C. PCR products were purified amplicons using a QIAquick PCR Purification Kit (Qiagen), and sequenced using the PCR primers at the OSU's Center for Genome Research and Biocomputing. Sequence chromatograms were checked visually and assembled into contigs in BioEdit (5).

After completing the Illumina sequencing of the samples (see Material and Methods), the paired-end read for each genome and transcriptome were separately mapped to the rDNA PCR products obtained using Geneious 9.5.0. (Biomatters Ltd) with the medium-low sensitivity mapping setting and mapping paired-reads which "map

nearby". We manually verified that, in all cases, the coverage was homogeneous along the PCR sequence. This allowed us to confirm that the read mapping and the PCR products gave the same sequences. Blastn searches against the genomic and transcriptomic assemblies showed that identical contig sequences were also assembled for the transcriptome. However, the rRNA cluster was found to be badly assembled in the DNA assembly probably due to its high coverage. Complete SSU rDNA sequences were submitted to the National Center for Biotechnology Information (NCBI) under accession MK480607 and MK480606 for *Henneguya salminicola* and *Myxobolus squamalis*, respectively.

Henneguya salminicola (Ward 1919) forms conspicuous white pseudocysts in the skeletal muscle of multiple salmon (*Oncorhynchus*) species (6). We sampled the parasite from adult Chinook (*O. tshawytscha*) and Coho salmon (*O. kisutch*) from the Salmon and Nehalem rivers, respectively, in Oregon, USA. Partial parasite SSU rDNA sequences (~1,000 bp) from all material used in this project, regardless of fish host, were identical to each other and to the genome assembly. The SSU rDNA sequence was 98.8% (1628/1648 bp) similar to the GenBank reference sequence for *H. salminicola* (AF031411; from *Oncorhynchus nerka*) (7), and 99.4% similar to the GenBank reference sequence for *H. zschokkei* (AF378344; from *Prosopium williamsonii*) (8). Given its host range and occasional synonymization in the literature as *Henneguya zschokkei* (Gurley, 1894), a definitive re-description of the species is needed with morphological and molecular characterization from its different fish hosts. In the absence of better taxonomic clarity, we refer to the *Henneguya* species herein as *H. salminicola* given its host, clinical presentation and myxospore morphology.

Myxobolus squamalis (Iversen 1954) forms white pseudocysts under the scales of several species of salmon and trout (6, 9). We sampled the parasite from its type host rainbow trout (*O. mykiss*) and coho salmon (*O. kisutch*), from the Willamette and Trask river basins, respectively. The ssrDNA contig assembled from the genome reads was 97.1% (2053/2114bp) similar to the GenBank reference sequence for *M. squamalis* (JX910363) (9). We sequenced the parasite ssrDNA from multiple samples from different

fish of both species, and uncovered at least two genetic types: one that was 99.6% (1006/1010bp) similar to the genome contig, and one that was identical (1030/1030 bp) to the GenBank reference. In the absence of better taxonomic clarity, we refer to the *Myxobolus* species herein as *M. squamalis* given its host, clinical presentation and myxospore morphology.

Our ssrDNA sequencing of myxozoans from different hosts and river basins within Oregon, revealed genetic differences among isolates of both *Henneguya* and *Myxobolus* samples and sequences available in NCBI. Determining whether this level of genetic variation represents regional variation, or is a marker of cryptic species, will require additional characterization of these myxozoans.

2- Light microscopy

Myxozoan pseudocysts of *M. squamalis* (3-4 mm in diameter, found under the scales of the host fish), and *H. salminicola* (4 - 8 mm, in the skeletal muscle of the host fish) were ruptured, and the white parasite content aspirated by pipette. This material was then diluted in up to 100 µL PBS and examined by light microscope to confirm the presence of developing, multi-cellular parasite stages. Living parasite cells were incubated in a hypotonic solution to increase distances between cell components: 1.5 mL 0.56 % KCl was added to each parasite suspension, and the tubes incubated at room temperature (21C) up to 45 min. The parasite was then gently pelleted at 3000g for 3 min before 90% of the supernatant was discarded. The cells were re-suspended in the remaining buffer by flicking the tube gently, then, freshly prepared 3:1 methanol:acetic acid was added dropwise, with continuous flicking/agitation to prevent cells from aggregating. Again, the suspension was spun down, the supernatant discarded, cells re-suspended and 3:1 methanol:acetic acid added dropwise. The cells were once-again spun down and all but 100-200 µL fixative removed. Cells were re-suspended by flicking, then 1-3 drops of cell suspension were put along a clean microscope slide and the fixative allowed to evaporate on the bench for 1-2 min, before placing in a ventilated hood for a further 5-10 min. The dried slides were then incubated at 40C for 1-2h. For DNA staining, several small drop of VECTASHIELD (Vectorlabs, Burlingame, CA) mounting medium, which contained

DAPI (4',6-diamidino-2-phenylindole, blue), were added across the slide and a coverslip added. Cells were visualized under a Leica DMR compound fluorescence microscope at 630x and 1000x magnification under illumination conditions appropriate for DAPI. Digital images were acquired using SPOT camera and software (SPOT Imaging, Sterling Heights, MI).

3- Assembly filtering

Filtering and assembling the DNA data

We used FastQC (10) (version fastqc_v0.10.1) to detect adapters and over-expressed sequences in the raw reads from each species. Any such sequences were removed using CutAdapt (11) (version 1.5) with parameters –b –m 20. Only paired reads were kept for further use. To filter reads for fish contamination we used Bowtie2 (12) (version bowtie2-2.2.3) with parameters -p 10 --sensitive-local to map our raw reads to fish sequence reference. We then discarded all reads that mapped to fish sequences. *Oncorhynchus kisutch* and *O. tshawytscha* are the host fish for *M. squamalis* and *H. salminicola*, respectively. Thus, the genomes assemblies GCF_002021735.1 (*O. kisutch*) and GCF_002872995.1 (*O. tshawytscha*) (13) were used to filter *M. squamalis* and *H. salminicola* reads respectively. Next, we used the filtered reads to build the first assembly of each myxozoan using IDBA (14) (version idba-1.1.1) with the parameters idba_ud -r.

These first assemblies were used for identifying contaminations in our data by running blastn (local BLAST version 2.7.0) (15) against the NCBI non-redundant nucleotide database (nt) with parameters -evalue 1e-75 -num_descriptions 20 -num_alignments 5 -max_hsps 5. We considered as contaminant contigs (from the first assembly) that had identity scores > 98% with non-myxozoan species. Under these parameters, four species were found as contaminants in our data set *Brochothrix thermosphacta* (a bacteria); *S. salar* (a salmon); *O. mykiss* (a trout); and *Anisakis simplex* (a nematode parasite of fish). Corresponding genome sequences were thus downloaded (on Nov 14, 2017) from NCBI as followed: *Brochothrix thermosphacta* strain BI chromosome CP023483.1, *S. salar* GCF_000233375.1 (16), *O. mykiss* GCF_002163495.1 (17), and *Anisakis simplex* GCA_000951095.1_A_simplex. We then

filtered the reads a second time (referred to as Filter-2) by using the same Bowtie2 parameters as above and discarded reads that mapped to contaminants. We used the Filter-2 reads to reassemble each myxozoan dataset using IDBA again with the parameters idba_ud -r using the same settings as above. From each of these last assemblies, we discarded contigs < 500bp and removed duplicates from the assembly using cd-hit (18) (v4.6.8-2017-0621), with the following parameters: cd-hit-est -c 0.95 -n 10 -d 0 -M 16000 -T 8. To verify there was no more contamination left in each assembly we ran a blastn (15) search against the nucleotide NCBI database (nt) with the above parameters. The *H. salminicola* assembly showed no fish contamination, but *B. thermosphacta* contamination was still detected. Hence for *H. salminicola*, all the contigs whose best blastn hit was *B. thermosphacta* were filtered out. However, fish contamination was still present in the *M. squamalis* DNA assembly. Thus, to further filter this assembly, a blastn search was first performed against the NCBI non-redundant nucleotide database (nt) with no e-value threshold and all contigs having a fish hit > 80% pident were removed. Second, using a blastx (15) search against the NCBI non-redundant (nr) protein database (with the parameters -evalue 1e-5 -max_target_seqs 1 -max_hsps 1), we removed all *M. squamalis* contigs from the DNA assembly that were of potential fish origin with a pident score > 50%.

Filtering and assembling the RNA data

We used FastQC (10) (version fastqc_v0.10.1) to detect adapters and overexpressed sequences in the reads. Any such sequences were removed using CutAdapt (12) (version 1.5) with parameters -b -m 20 and only paired reads were kept for further use. To filter reads for fish contamination we used Bowtie2 (14) (version bowtie2-2.2.3) with parameters -p 10 --sensitive-local and discarded all reads that had mapped to fish sequences, in this case the transcriptome assembly of *Oncorhynchus kisutch* GDQG01.1.fsa_nt (downloaded from NCBI 28/12/17).

There was no transcriptome assembly available in public databases for the fish host of *H. salminicola*, *O. tshawytscha*. Hence, to filter *H. salminicola* reads we downloaded the RNA reads (on 30/11/2017) of *O. tshawytscha* from three experiments:

SRX3411741, SRX3379475, and SRX3379474 (13). Then we assembled all the RNA reads of *O. tshawytscha* using Trinity (19) (version Trinityrnaseq_r20131110) with min_kmer_cov 2 and --jaccard_clip and used this RNA assembly for filtering the *H. salminicola* RNA reads using Bowtie2 as described for *M. squamalis*. Finally, the filtered reads of *H. salminicola* and *M. squamalis* were each assembled using Trinity (19) with parameters --min_kmer_cov 2 --jaccard_clip --SS_lib_type FR. For each assembly, we discarded contigs < 300bp. We filtered each assembly for contamination in three steps: 1) Running blastn (version BLAST-2.6.0+) for each RNA assembly against each corresponding genome assembly and kept only RNA contigs that had a hit in the DNA assembly, 2) Running RSEM version1.2.8 (20) to determine transcript abundance and filter low abundance transcripts by using the filter_fasta_by_rsem_values.pl script supplied by Trinity on the above RNA assemblies with the parameters --fpkm_cutoff=0.01 --isopct_cutoff=1.00, and 3) Running a blastn search with our RNA assemblies as queries against the NCBI nr protein database with parameters -evalue 1e-75 -max_hsps 1 (21) and discarding contigs that have > 95% identity to species that are not myxozoans. Because fish contamination was still present in the *M. squamalis* RNA assembly after this third step, we then discarded all contigs that had > 80% identity to fish (Salmonidae, Cyprinidae, or Cichlidae). Last, for each RNA assembly, we kept only the longest transcript by using the Trinity script get_longest_isoform_seq_per_trinity_gene.pl.

4- Functional annotation of the transcriptomes

The *H. salminicola* and *M. squamalis* transcriptomes were annotated using the Trinotate pipeline (https://github.com/Trinotate/Trinotate.github.io/wiki) (22). The Trinotate output was converted to a table format, which is accepted by NCBI TSA database, using a custom python script (trinotate_table_to_tbl.py, available at https://github.com/neuhofmo/genome-annotation-tools/blob/master/trinotate_table_to_tbl.py). As a rule, blastp results of predicted CDS were preferred over blastx results. The annotated CDS features were modified manually to comply with NCBI requirements, and low quality CDS features, as well as partial sequences, were removed using the same script.

The *M*. *squamalis* annotated transcriptome shotgun assembly project has been deposited at DDBJ/EMBL/GenBank under the accession GHBR00000000. The version described in this paper is the first version, GHBR01000000. The *H*. *salminicola* annotated transcriptome shotgun assembly project has been deposited at DDBJ/EMBL/GenBank under the accession GHBP00000000. The version described in this paper is the first version, GHBP01000000.

We also extracted KEGG IDs for the transcripts out of the Trinotate table using custom scripts, and mapped the pathways using the KEGG database (23). Specifically, we compared the enzyme of the oxidative phosphorylation pathways among *H*. *salminicola*, *M*. *squamalis* and *K*. *iwatai*. KEGG IDs of *K*. *iwatai* were obtained following Chang et al. (24). A figure comparing the KEGG IDs *H*. *salminicola*, *M*. *squamalis*, and *K*. *iwatai* was drawn with https://www.kegg.jp/kegg/tool/map_pathway.html (Fig. S8).

5- Gene prediction and annotation of the nuclear genomes

We used the MAKER2 pipeline (25) incorporating the semi-Hidden Markov model (HMM) based Nucleic Acid Parser (SNAP v.20131129) (26) and Augustus v.3.3 (27) *ab initio* gene predictors in order to quantify the overall genome characteristics of our final, filtered, genome assemblies of *M*. *squamalis* and *H*. *salminicola*, and to obtain predicted coding regions of the genomes to further analyze for the presence and absence of certain genes.

MAKER2 (v. 2.31.9) (25) was run on each species in two rounds, the first round included: 1) species-specific transcriptomic evidence from our final Trinity assemblies; 2) protein evidence from the longest ORFs predicted by Transdecoder v.3.0.1 (http://transdecoder.github.io, last accessed October 2018) for each transcript; 3) we incorporated our previously published transcriptome of *M*. *cerebralis* (24) as an EST from a closely related species for the annotation of *M*. *squamalis*; and 4) the precompiled HMM profile of the nematode *Brugia malayi* to train SNAP. The HMM profile of *B*.

malayi was chosen for training because it was shown to produce better protein predictions for *Hydractinia echinata* (28). Indeed the low GC content of the genome of *B*. *malayi* is comparable to those of many cnidarians including Myxozoa (24). This first round resulted in a species-specific HMM profile created by SNAP for each assembly and provided information about the genomic location of the input transcriptomic and protein sequences in our new genomic assemblies for *M*. *squamalis* and *H*. *salminicola*. Using this information, we extracted the genomic regions to which the protein and EST evidence mapped, plus the 1,000bp upstream and downstream of these regions. This sequence data were used as input for BUSCO v.2.0 (29) to develop a set of Augustus gene prediction parameters for each species. Both the Augustus parameters and HMM profiles were used in a second run of MAKER2 for each species to do *ab initio* gene prediction without the specific protein and EST evidence.

After each round of annotation, GFF3 files were exported, and annotation statistics were calculated using the gt-stats function of the GenomeTools package (30) with the "-addintrons" option, to infer the presence of introns between exonic features. We found that while the second round of annotation using solely the *ab initio* predictors increased the number of predicted coding regions for *H*. *salminicola*, and hence appeared to give us further information about gene content in this species, for *M*. *squamalis* the second round actually reduced the number of predicted genes. Therefore, for further analyses we used the predicted protein set from the second round of annotation for *H*. *salminicola* and from the initial round of annotation for *M*. *squamalis*.

6- Functional annotation of the nuclear genomes

The predicted proteins identified by MAKER2 were annotated based on their sequence similarity to *T*. *kitauei* proteins. Blastp (15) searches were performed using the predicted protein sets of *H*. *salminicola* and *M*. *squamalis* against the NCBI non-redundant (nr) protein database (with the parameters -evalue 1e-5 -max_target_seqs 1 -max_hsps 1). Annotations were only retrieved and incorporated to the gff file submitted to NCBI, if the best hits originated from *T*. *kitauei*.

The *M*. *squamalis* annotated genome shotgun assembly project has been deposited at DDBJ/EMBL/GenBank under the accession QWKW00000000. The version described in this paper is the first version, QWKW01000000. The *H*. *salminicola* annotated genome shotgun assembly project has been deposited at DDBJ/EMBL/GenBank under the accession SGJC00000000. The version described in this paper is the first version, SGJC01000000.

7- Building protein databases to search for presence and absence of specific genes

Protein sequences from the fly (*Drosophila melanogaster*) were used as starting point of our databases since the mitochondrial proteome of this species is well characterized (31). Unlike human, whose mitochondrial proteome is also well characterized (32), the fly did not undergo two whole genome duplications (33), which facilitates the identification of orthologues. However, in some cases, our databases were completed with human sequences (see below).

Database of fly and human proteins interacting with mtDNA and mtRNA

First, mt-ribosomal proteins of the fly were downloaded from FlyBase (34) (http://flybase.org/reports/FBgg0000059.html; last accessed 11/2017). Specifically, for each mt-ribosomal gene, the protein sequence of the longest transcript was downloaded. Second, fly proteins which were catalogued in the MitoDrome database (http://mitodrome.ba.itb.cnr.it/; last accessed 01/2018) (31) under functional classification "DNA and RNA", and "Protein synthesis: Others" were retrieved. Finally, to complete this database of proteins interacting with mtDNA and mtRNA, we downloaded from NCBI the sequences of human proteins known to bind to mtRNA and described in Rackham et al. (35).

We searched for fly homologues of these human protein sequences using blastp against the fly proteome (GCF_000001215.4_Release_6_plus_ISO1_MT_protein.faa; downloaded on 29/11/17) (36). We then ran a reciprocal blastp search using only the best hits, with an E-value cutoff of 1e-05, as query against the Human proteome

(GCF_000001405.38_GRCh38.p12; downloaded on 27/01/18). All fly proteins that returned the human protein that had been used as query in the first search were retained (Data table S2).

Database of fly mt metabolic proteins

We downloaded from the MitoDrome database (http://mitodrome.ba.itb.cnr.it/drome.php; last accessed 01/17/2018) the collection of fly proteins corresponding to the different mitochondrial metabolic pathways. For each gene, the protein sequence of the longest transcript was downloaded. The proteins interacting with mtDNA and mtRNA were excluded, as they were already examined using the methods above. The accession number for each fly protein was assigned according to the fly proteome from NCBI GCF_000001215.4_Release_6_plus_ISO1_MT_protein.faa. (Data table S1).

Database of enzymes involved in anaerobic metabolism in protists

In anaerobic protists, the modifications of energy metabolism are often coupled with the horizontal gene transfer of prokaryotic genes (reviewed in (37, 38)). For example, a key enzyme that has been transferred independently in several anaerobic protists is the pyruvate:ferredoxin oxidoreductase. A list of key metabolic proteins that originated from horizontal gene transfer in anaerobic protists was established from (37, 38) (Data table S3).

8- Identification of *Hydra*, *Polypodium* and myxozoan nuclear encoded mt proteins based on reciprocal BLAST search

The fly and human protein sequences from the databases described in the above paragraphs were the starting points of reciprocal BLAST (15) searches conducted to identify homologous copies in myxozoans (i.e., *H. saminicola; M. squamalis; Kudoa iwatai*, *Myxobolus cerebralis, Thellohanelus kitauei)* and in the Myxozoa outgroup *Polypodium hydriforme*. Our personal observations of myxozoan transcriptomes suggest that intron retention (39) is a common phenomenon in Myxozoa. This observation is not surprising since intron retention has been found to be an important regulatory mechanism in Cnidaria (40, 41). The corollary, however, is that in myxozoan transcriptomes the

longest transcript might retain introns. Consequently, complete CDS are not easily inferred from transcriptome sequence. We thus found that blastx searches against transcriptome assemblies led to a higher number of protein coding gene identification than blastp searches against predicted peptide sequences or translated ORF. Similarly, HMM approaches (42) which are based on peptide sequences did not lead to a higher number of protein identification for *H. salminicola* and *M. squamalis* in our preliminary analyses. As a case in point, our blastx searches allowed us to recognize all the *Kudoa* proteins identified by Muthye and Lavrov (43) who used HMM profiles, confirming that the use of HMM profiles did not improve protein identifications in our case (Data tables S1-S2).

Identification of nuclear encoded mt proteins in the *Hydra* proteome

Because Myxozoa are fast evolving we first identified homologous copies of these proteins in *Hydra vulgaris*. Specifically, blastp (15) searches were conducted against the *Hydra* proteome (GCF_000004095.1_Hydra_RP_1.0_protein.faa; downloaded on 29/11/17) (44). We then ran a reciprocal blastp search using only the best hits, with an E-value cutoff of 1e-05, as query either (a) against the proteome of *D. melanogaster* (GCF_000001215.4_Release_6_plus_ISO1_MT_protein.faa; downloaded on 29/11/17) (36), when a *D. melanogaster* sequence was used as query, or (b) against the human proteome (GCF_000001405.38_GRCh38.p12; downloaded on 27/01/18), when a human sequence was used as query.

All *Hydra* proteins that returned the fly (or human) protein that had been used as queries in the first search were retained. These *Hydra* sequences were then used as queries to identify homologous copies in Myxozoa (Data tables S1-S2). In the absence of *Hydra* homologs, the fly or human sequences were used.

Identification of nuclear encoded mt proteins in myxozoan and *Polypodium* transcriptomes

The *Hydra,* fly and human proteins were used as queries to conduct tblastn searches against the transcriptome of *H. saminicola; M. squamalis; K. iwatai* (NCBI entry

GBGI01000000.1) (24), *M. cerebralis* (NCBI entry GBKL01000000) (24), and *P. hydriforme* (NCBI entry GBGH01000000) with an E-value cutoff of 1e-05. For each species, we then ran reciprocal blastx searches using only the first hits as queries against the proteome of *H. vulgaris* (with *Hydra* queries), the fly proteome (with fly queries) or the human proteome (with human queries) (see proteome accession above). All myxozoan proteins that returned the original query protein (from *Hydra* or fly or human) were retained (Data tables S1-S2).

When a hit was found for a myxozoan species but not for another, tblastx searches were performed using the myxozoan sequence as query against the transcriptome of the other species. As above, reciprocal tblastx searches were then performed, this time against the transcriptome of the first species to support the homology of the hit, with an E-value cutoff of 1e-05. Similarly, when a hit was found in *Polypodium* but not in myxozoans tblastx searches were performed using the *Polypodium* sequence as query, followed by the corresponding reciprocal search.

Identification of nuclear encoded mt proteins in the *Thelohanellus kitauei* proteome

Because transcriptome assemblies are not available in public databases for the myxozoan *Thelohanellus kitauei;* the *Hydra,* fly and human proteins were used as queries to conduct blastp searches against the proteome of *T. kitauei* (GCA_000827895.1_ASM82789v1_protein.faa; downloaded on 29/11/17) (45). Reciprocal blastp searches using only the best hits, with an E-value cutoff of 1e-05 were then performed against the *Hydra*, fly and human proteomes respectively (see proteome accession above). All *T. kitauei* proteins that returned the original query protein (from *Hydra* or fly or human) were retained (Data tables S1-S2).

For proteins that had no hit in the *T. kitauei* proteome using the *Hydra,* fly and human proteins, but had hits in other myxozoan transcriptomes, we performed blastx searches with the myxozoan transcript sequences. If a hit was found in the *T. kitauei* proteome with an E-value cutoff of 1e-05, reciprocal tbastn searches were performed with the *T. kitauei* hit against the corresponding myxozoan transcriptome.

Identification of nuclear encoded mt proteins, in myxozoan genomes

Genes found in the transcriptome assemblies of *K*. *iwatai*, *M*. *squamalis* and *H*. *salminicola* were identified in the corresponding DNA assemblies using blastn searches. (i.e., sequences with 100% identity were considered as hits). The DNA assembly of *K*. *iwatai* was downloaded from NCBI (entry JRUX01000000.1).

We also searched, in these three genomes, for protein-coding genes that had not been identified in the transcriptomes. Specifically, for each species, tblastn searches were performed using proteins of *H*. *vulgaris* and *T*. *kitauei* as query against the genome sequences. Reciprocal blastx searches were performed with the genomic hits that received an E-value < 1e-05 hit against the corresponding proteome.

Identification of nuclear encoded mt proteins in the predicted proteome of *H. salminicola* and *M. squamalis*

Genes found in *H*. *vulgaris* and *T*. *kitauei* proteomes were searched for, using blastp, in the proteomes of *H*. *salminicola* and *M*. *squamalis* predicted by MAKER2 (see above details). Reciprocal blastp searches were performed with the predicted protein hits that received an E-value < 1e-05 hit against the corresponding proteome.

When a hit was found in the predicted proteome for either *H*. *salminicola* or *M*. *squamalis* but not for the other, blastp searches were performed using the corresponding predicted protein sequence as query against the predicted proteome of the other species. As above, reciprocal blastp searches were then performed, this time against the predicted proteome of the first species to support the homology of the hit with an E-value cutoff of 1e-05.

Identification of pseudogenes in *H. salminicola* genome

All genome hits that did not have either a transcriptome or a proteome hit in *H*. *salminicola* were examined manually. First, genomic reads were mapped to the corresponding DNA contig with the program Geneious 9.0.5 (Biomatters Ltd.) using the following parameters: Maximum gap per read=10%; Maximum gap size=15; Minimum overlap=25; Minimum overlap identity=80%; Word length=18; index word length=13;

Maximum mistmatches per read=20% Maximum ambiguity=4; “Map multiple best matches = to all”; “only map reads which = map nearby”. This step allowed us to verify that the coverage was homogeneous and sufficiently deep (about 300) along the genomic contig and to verify the absence of assembly errors.

The DNA sequence of *H*. *salminicola* was then aligned with the homologous DNA region identified in *T*. *kitauei* and *M*. *squamalis*. The RNA transcript of *T*. *kitauei* and *M*. *squamalis* were also aligned to identify intron exon boundaries. The alignments were performed with MAFFT v7.308 (46) under the L-INS-i algorithm as implemented in Geneious 9.0.5, and then corrected manually. The open reading frame of *H*. *salminicola* was examined manually for unambiguous cases of frameshift and stop codon. Such cases are presented in Fig. S7.

Search for anaerobic metabolism enzymes of bacterial origins

Eukaryote proteins involved in anaerobic metabolisms (Data table S3) were used as query to perform tblastn and blastp against the transcriptome assembly and MAKER2 protein predictions of *H*. *salminicola* respectively. However, no hit received an E-value < 1e-05.

9- Reconstruction of the mt genome of M. squamalis

Blastn and tblastx searches were conducted against the filtered genome (assembled with IDBA) using sequence JWZT01002463 of the closely related myxozoan *Thelohanellus kitauei* as query. This sequence has been found to encode mitochondrial (mt) genes (47).

The BLAST search identified “scaffold_890” as a putative mitochondrial sequence. The Perl script Novoplasty v2.6.3 (48) was then used to reconstruct a first draft of the mitochondrial sequence of *Myxobolus squamalis* under the “mito” mode. The program was run using scaffold_890 as seed and using the *M*. *squamalis* reads filtered from putative contaminations. The insert size was set to “367” and read length to “151” based on the characteristics of the Illumina sequencing run. Because myxozoans can have unusually large mt genomes (47, 49) the parameter “Genome Range” was increased to “12,000-50,000”. All other options were kept to default settings.

Three versions of the mitochondrial genome assembly were reconstructed using Novoplasty. These three circular sequence versions differed by about 100 nucleotide substitutions and had slight differences in their start and end regions. A majority consensus sequence was built based on the alignment of these three sequences. The start and end of the obtained consensus sequence were found to be identical, allowing the consensus sequence to be circularized. This circular sequence was called Myxobolus_mt_Assembly_v1. We then mapped reads to the Myxobolus_mt_Assembly_v1 sequence to verify that read mapping supported the circular conformation chosen and that no major drop in coverage existed along the circular sequence. Read mapping was performed with the program Geneious 9.0.5 (Biomatters Ltd.). The mapping options were Medium-Low Sensitivity, Fine tuning none (fast / read mapping), Map multiple best matches to all, "only map reads which map nearby" and "trim paired read overhangs". Visual examination of the existing SNP using Geneious 9.0.5 indicates that our sequence data was composed of two major haplotypes, one present in 60% and the other in 40% of the reads. This observation is likely since the DNA was extracted from several plasmodia, and we thus have a multi-isolate sample. From these data we extracted the majority consensus sequence from the reads. The new sequence was called: Myxobolus_mt_Assembly_v2. We then remapped reads to Myxobolus_mt_Assembly_v2 using relatively strict parameters for the custom sensitivity options. Specifically, we used Maximum gap per read=2%; Maximum gap size=10; Minimum overlap=75; Minimum overlap identity=98%; Word length=10; index word length=10; Maximum mismatches per read=2% Maximum ambiguity=16. We also used "Map multiple best matches to all", "only map reads which map nearby" and "trim paired read overhangs". The resulting mapping on Myxobolus_mt_Assembly_v2 contained drops in coverage suggesting that the sequence obtained was a chimera of the two haplotypes.

The eleven selected regions from Myxobolus_mt_Assembly_v2 which presented the highest coverage were extracted for further analysis. Several rounds of read mapping with high sensitivity were performed on these regions. Each round of read mapping

allowed us to elongate the regions with the highest coverage. For example, requiring a minimum overlap of 25 bp between a read and the mt contig allowed us to extend the sequence by 75 bp, since the reads are 100 bp long. The majority rule consensus was kept at each mapping round. Reads were mapped until all sequence fragments overlapped and a complete and circular mitochondrial sequence could be assembled. The new sequence was called: Myxobolus_mt_Assembly_v3_final and has been submitted to the National Center for Biotechnology Information database under accession number MK087050. A final read mapping was performed with Bowtie2 (12) (2.2.3) for both genomic and transcriptomic reads. The coverage of the final consensus sequence is presented in Fig. S9. The mean base coverage with DNA reads was 185 (Std Dev: 82; Minimum: 29; Maximum: 398).

The annotation of the mitochondrial genome of *M. squamalis* was performed manually with the help of the automated gene annotation tools, MFannot (http://megasun.bch.umontreal.ca/cgi-bin/mfannot/mfannotInterface.pl, last accessed June 20, 2018), RNAweasel (http://megasun.bch.umontreal.ca/RNAweasel/, last accessed June 20, 2018), and MITOS (50) (http://mitos.bioinf.uni-leipzig.de/index.py, last accessed June 20, 2018) with parameter "Genetic Code 04". All other parameters were set to default.

Only two canonical protein coding genes, *nad1* and *nad5,* could be recognized based on sequence alignment with other myxozoan mitochondrial sequences. We also identified 13 ORFs longer that 100 nucleotides. Blastp and tblastn searches were conducted using the translated ORFs as query against the nr databases (nucleotide and protein) of NCBI. No significant similarity (with an E-value cutoff of $< 10^{-5}$) with any NCBI sequence could be observed for these ORFs. Each ORF (i.e., either canonical or unknown) was submitted to the TMHMM Server v. 2.0 (http://www.cbs.dtu.dk/services/TMHMM/) (51) to detect the presence of putative transmembrane domains which are present in most mt protein coding genes.

Annotation of repeats was performed with the help of the Geneious plugin repeat finder version 1.0. Specifically, repeats longer than 150 bp were annotated. We allow

repeats to include insertions and contain up to 5% sequence divergence and we considered nested repeats to belong to the same repeat group. The start and end of each ORF, either canonical (*nad1* and *nad5*) or unknown, were chosen in order to both reduce the overlap between consecutive ORFs and to minimize the length of the non-coding region between ORFs.

The location of the rRNA gene identified by MITOS and MFannot was also supported by the presence of numerous RNA reads mapping to these regions (Fig. S9B). The boundaries of the rRNA have been inferred from the flanking genes in their 3'-end. Because the 5'-ends of both rRNA genes were flanked by repeats, the 5'-end boundaries were inferred based on the presence of rRNA reads.

10- Phylogenetic reconstructions based on mt protein sequence

Phylogenetic analyses of cnidarian relationships were performed based on the *nad1* and *nad5* genes. The phylogenetic data matrix was built following the methodology described in Yahalomi et al. (47) using the same taxon sampling with the addition of *Myxobolus squamalis*. Bayesian phylogenetic reconstructions were performed with phylobayes version 4.1b (52) both under the GTR + Γ model and the CAT + Γ mixture model.

For the GTR + Γ model, four chains of Markov Chains Monte Carlo (MCMC) were run for 400,000 cycles and sampled every 10 cycles. After a manual examination of the trace files produced by Phylobayes, the first 50,000 trees were discarded as burn-in. The consensus tree was thus based on 35,000 sampled trees. The maximum and average difference between chains observed at the end of the run, were 0.021 and 0.002, respectively. The maximum discrepancy and the minimum effective size of the different chain parameters were all < 0.03 and > 10,000 respectively.

For the CAT + Γ model, four chains of Markov Chains Monte Carlo (MCMC) were run for 3,000,000 cycles and sampled every 50 cycles. After a manual examination of the trace files produced by Phylobayes, the first 500,000 trees were discarded as burn-in. The consensus tree was thus based on 50,000 sampled trees. The maximum and

average difference between chains observed at the end of the run, were 0.025 and 0.001, respectively. The maximum discrepancy and the minimum effective size of the different chain parameters were all < 0.03 and > 10,000 respectively. These different results indicate a good convergence of the GTR + Γ and CAT + Γ chains. The phylogenetic dataset is available in the TreeBASE repository (http://purl.org/phylo/treebase/phylows/study/TB2:S23827?x-access-code=e9bba2fefab72915b2e4ee97923768d7&format=html).

SUPPLEMENTARY RESULTS AND DISCUSSION

1- Characteristics of the mitochondrial sequence of *M. squamalis*

The mt genome of *M. squamalis*, presented in Fig. S5, has similar characteristics to those of *Kudoa* species (47, 49). Specifically, the presence of repeats in non-coding regions, the absence of tRNAs, and extremely fast evolutionary rates. Because of this extreme rate, only four canonical mt genes (*rnl, rns, nad1* and *nad5*) could be identified. This mt genome differs, however, by the presence of several non-coding regions rather than a single large non-coding region. Although only two protein coding genes could be identified, 13 additional open reading frames (ORF) were detected. Among these 13 ORFs, 10 were predicted to include transmembrane domains. Because canonical mitochondrial genes involved in the electron transport chain include transmembrane domains, this finding suggests that these 10 ORFs could represent canonical mitochondrial protein genes whose fast evolutionary rate hindered their identification. The presence of three genes without predicted transmembrane domains, is surprising as it suggests the presence of novel protein coding genes encoded within the mitochondrial genomes of *M. squamalis*. Such proteins were not identified in the mitochondrial genomes of *Enteromyxum* and *Kudoa* (47, 49) and could thus represent false negative results. To confirm if these ORFs are artifacts or genuine myxozoan genes, additional work should be done to identify whether homologous ORFs are present in the mitochondrial genome of other *Myxobolus* species. All genes, both canonical and predicted, were found to be oriented on the same strand.

2- Phylogenetic reconstructions based on mt protein sequences

The phylogenetic tree reconstructed based on mitochondrial protein genes is presented in Fig. S6. All myxozoan sequences, including the newly obtained sequence of *M. squamalis*, show an extreme rate of evolution in agreement with Yahalomi et al. (47). The consequence of this high evolutionary rate is a lack of similarity between myxozoan and cnidarian sequences. Therefore, the phylogenetic reconstructions based on mt sequence do not resolve the relationships among cnidarians, nor the phylogenetic position of Myxozoa within cnidarians. Relationships among myxozoan mt sequences, however, agree with molecular trees based on 18S rRNAs (53): *Myxobolus* and *Thelohanellus*,

which both belong to the fresh-water/oligochaete host lineage, form the sister clade of *Kudoa* and *Enteromyxum*, which belong to the marine/polychaete host lineage. The monophyly of these two groups is recovered with high support (PP=0.97-1.0). This confirms our hypothesis that the *M. squamalis* sequence is a genuine mitochondrial sequence and not contamination from another eukaryote.

3- The oxidative phosphorylation pathways of *H. salminicola, M. squamalis* and *K. iwatai*

The comparison of the KEGG IDs found among the RNA transcript and mitochondrial sequences of *H. salminicola*, *M. squamalis*, and *K. iwatai* is presented in Fig. S8. The results support the inferences observed based on Blast searches. Specifically, no KEGG ID corresponding to Complex I proteins was found for *H. salminicola*, and KEGG IDs corresponding to complex IV were only found for *K. iwatai* transcripts. Since complex IV is the complex interacting with O_2 molecules, this suggests that *H. salminicola*, and possibly *M. squamalis*, might not be undergoing standard cellular respiration. By contrast, all three myxozoans possessed KEGG IDs belonging to complex II, which is part of the Krebs cycle, as well as complex V (the ATP synthase complex). This suggests that a proton gradient is still present across the inner organelle membrane in *H. salminicola*. However, the complete anaerobic pathway still remains to be elucidated and the electron acceptor is still unknown for *H. salminicola*. Intriguingly, *H. salminicola* could possess a novel anaerobic pathway, as homologues of the anaerobic metabolism enzymes utilized by anaerobic protists could not be identified in this species (Data table S3).

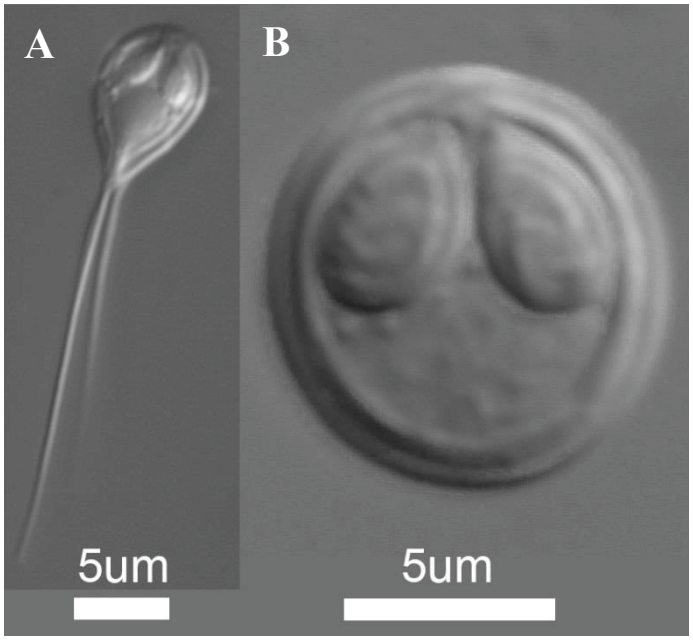


Fig. S1. Myxospores of the two myxozoan parasites studied. A. *Henneguya salminicola*, B. *Myxobolus squamalis*

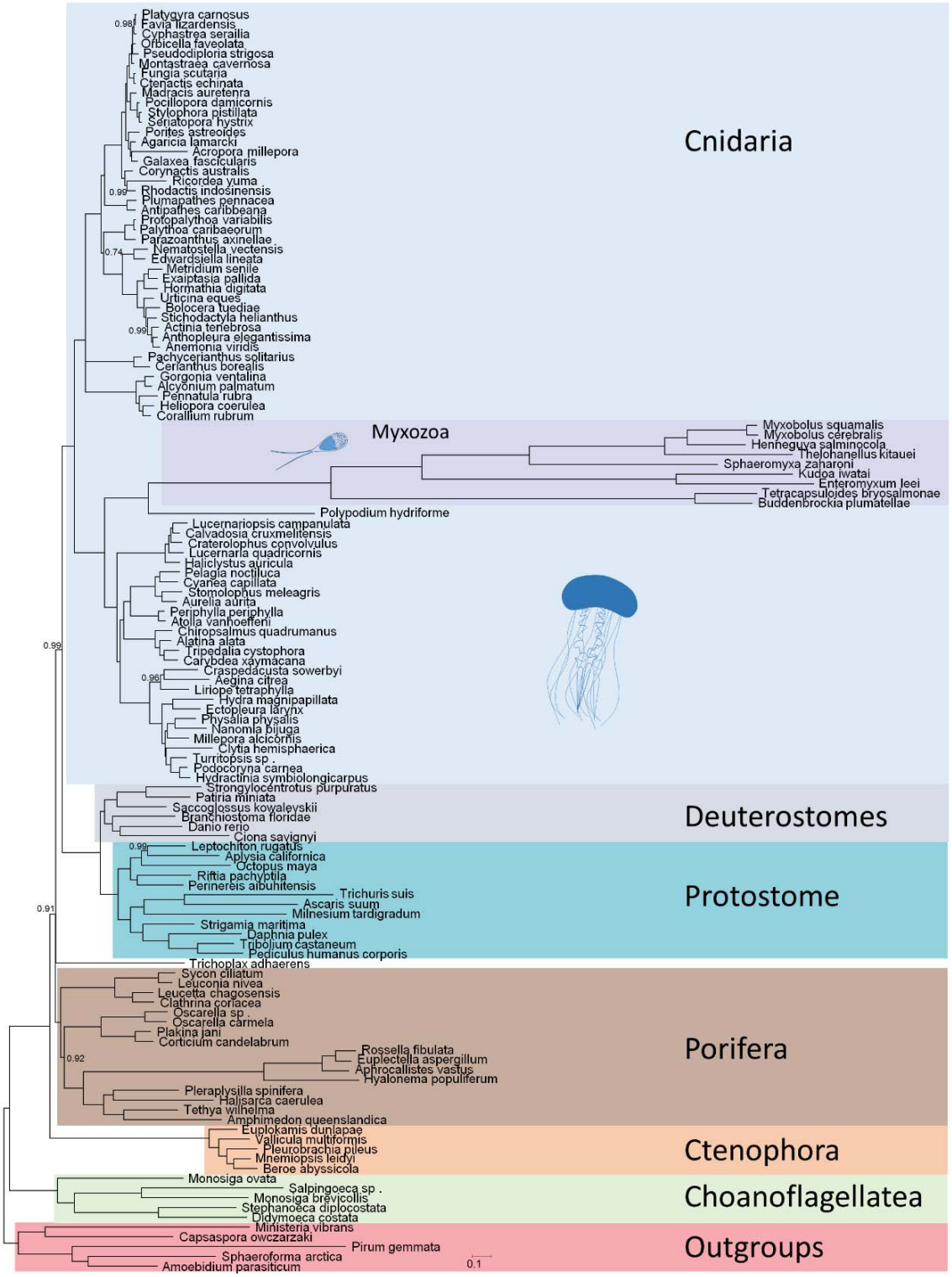


Fig. S2. Metazoan phylogenetic relationships inferred from a supermatrix of 11,352 amino acid positions for 129 species. Bayesian majority-rule consensus tree reconstructed under the CAT + Γ model from two independent Markov-chain Monte Carlo chains. Node supports (posterior probabilities) are only indicated for those which did not received maximum support (1.0). Branches with a support under 0.5 are collapsed.

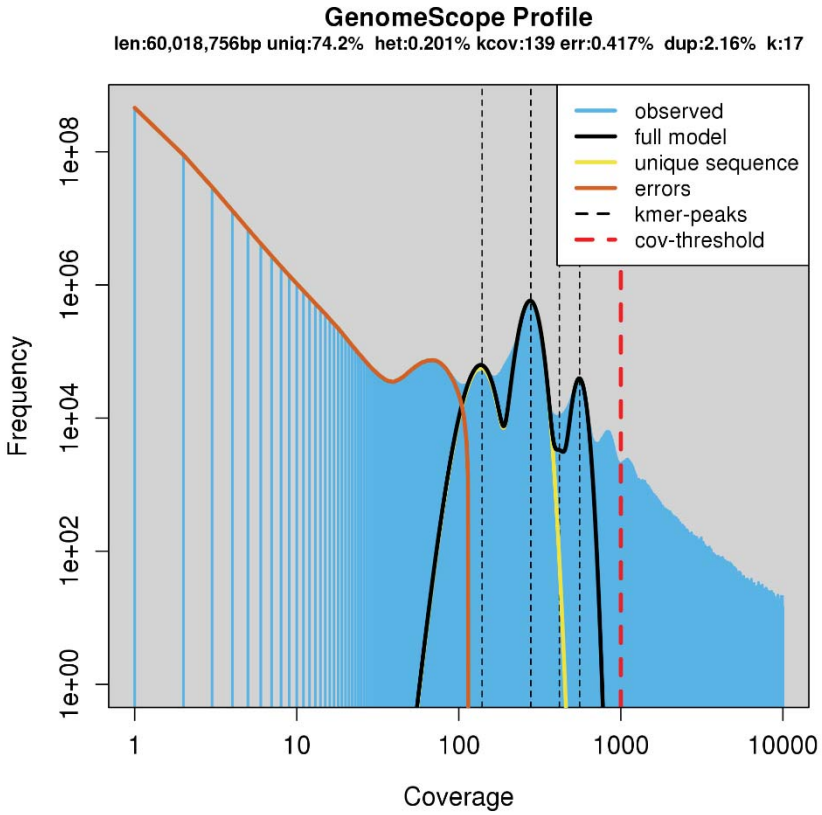


Fig. S3. GenomeScope (54) profile of *H. salminicola*. Heterozygosity was estimated to be between 0.18-0.22%. The Haploid length of the genome and the repetitive component of the genome were estimated to be between 59,8-60.0 Mb and 15.46-15.51 Mb respectively. Finally the model fit was between 85.0-87.1%.

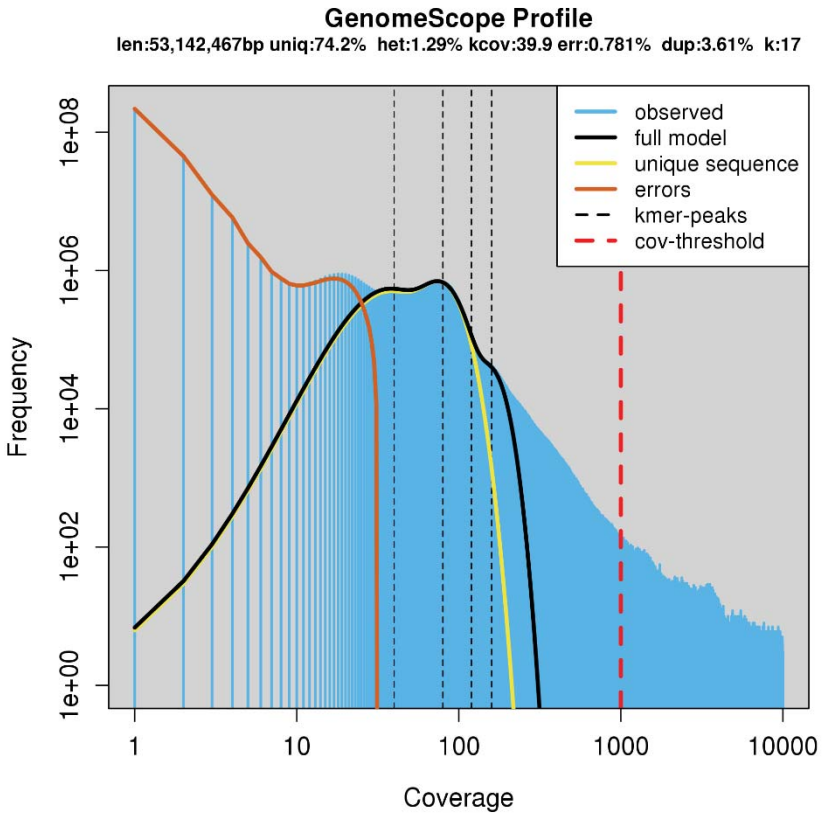


Fig. S4. GenomeScope (54) profile, *M. squamalis*. Heterozygosity was estimated to be between 1,26-1.32%. The Haploid length of the genome and the repetitive component of the genome were estimated to be between 52,6-53.1 Mb and 13.55-13.69 Mb respectively. Finally the model fit was between 91.4-94.6%.

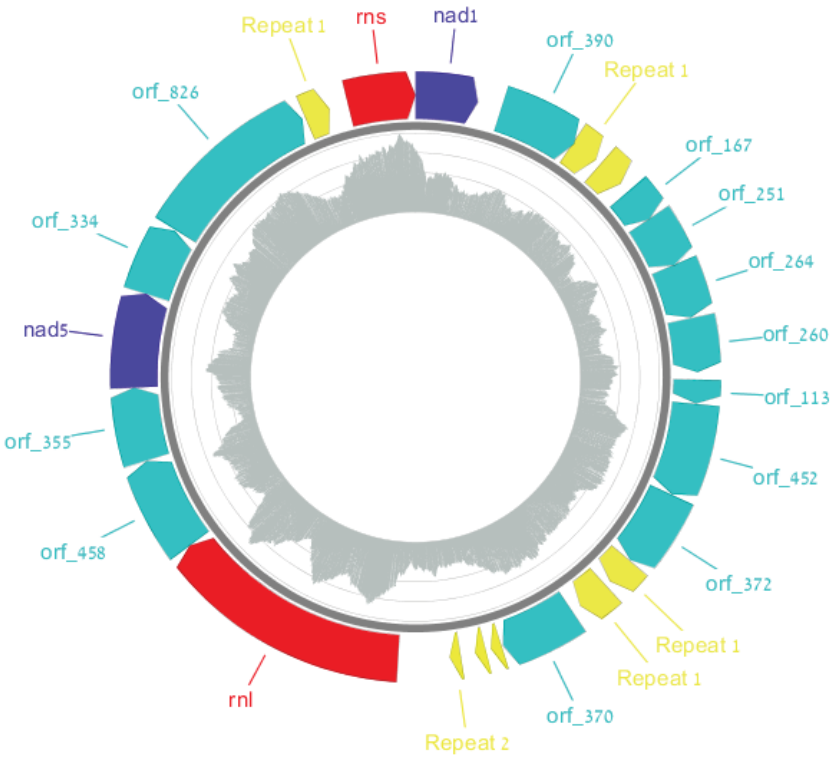


Fig. S5. Illustration of the mitochondrial genome of *Myxobolus squamalis*. Genes are shown in dark and light blue, rns and rnl genes are indicated in red, and repeated elements are in yellow. The GC content is indicated in grey inside the circle.

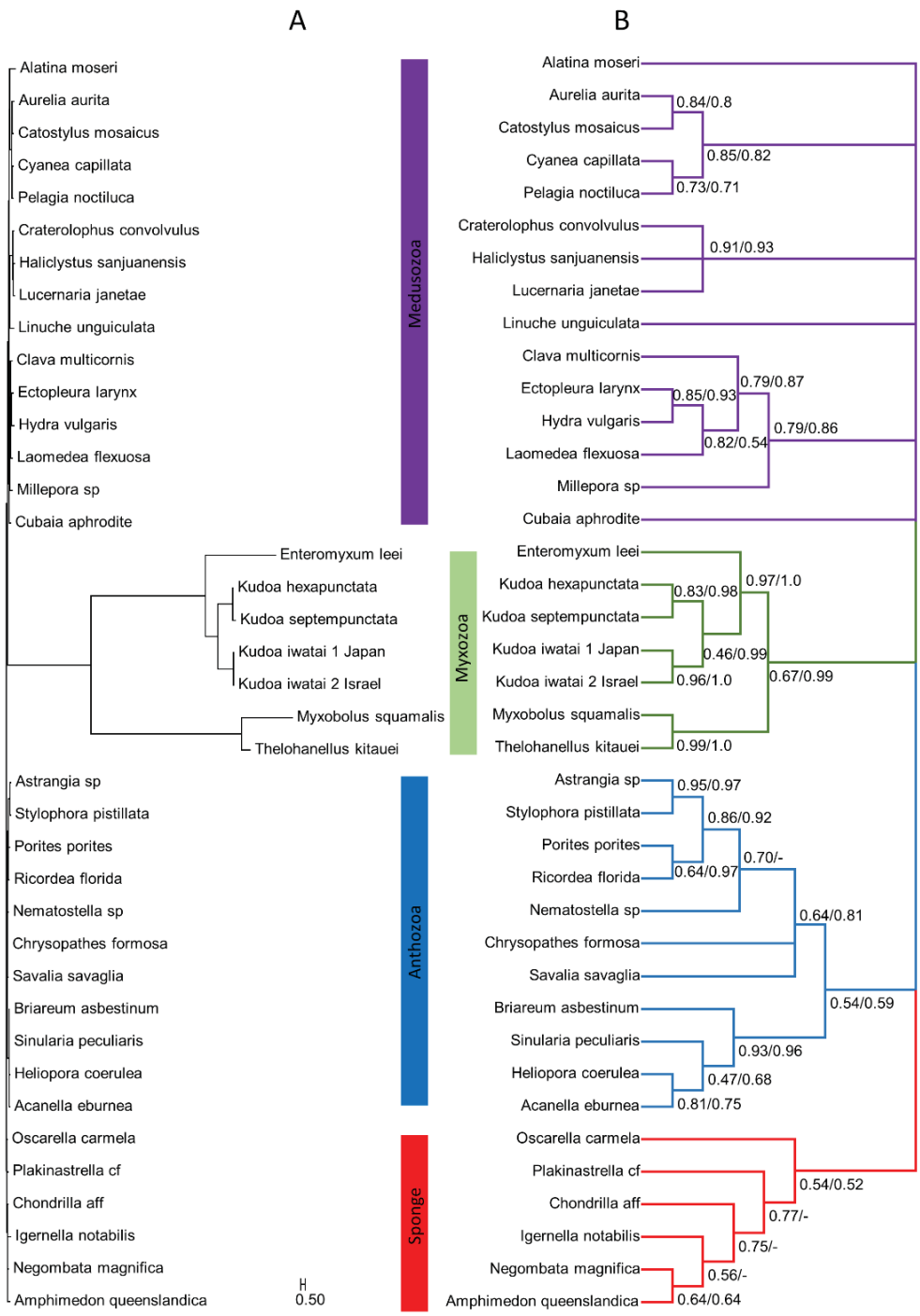


Fig. S6. Phylogenetic tree of cnidarians based on a concatenated alignment of two mitochondrial proteins, *nad1* and *nad5*. (A) Bayesian consensus tree from four independent MCMC runs obtained using the CAT model (41 taxa and 318 amino-acid sites). Note the long branches leading to Myxozoa, attributable to the very fast sequence evolution rates. (B) Topology of the same Bayesian tree showing only the nodes where posterior probability is >0.5. Remaining nodes are labeled with the posterior probability values of both CAT and GTR+Γ models.

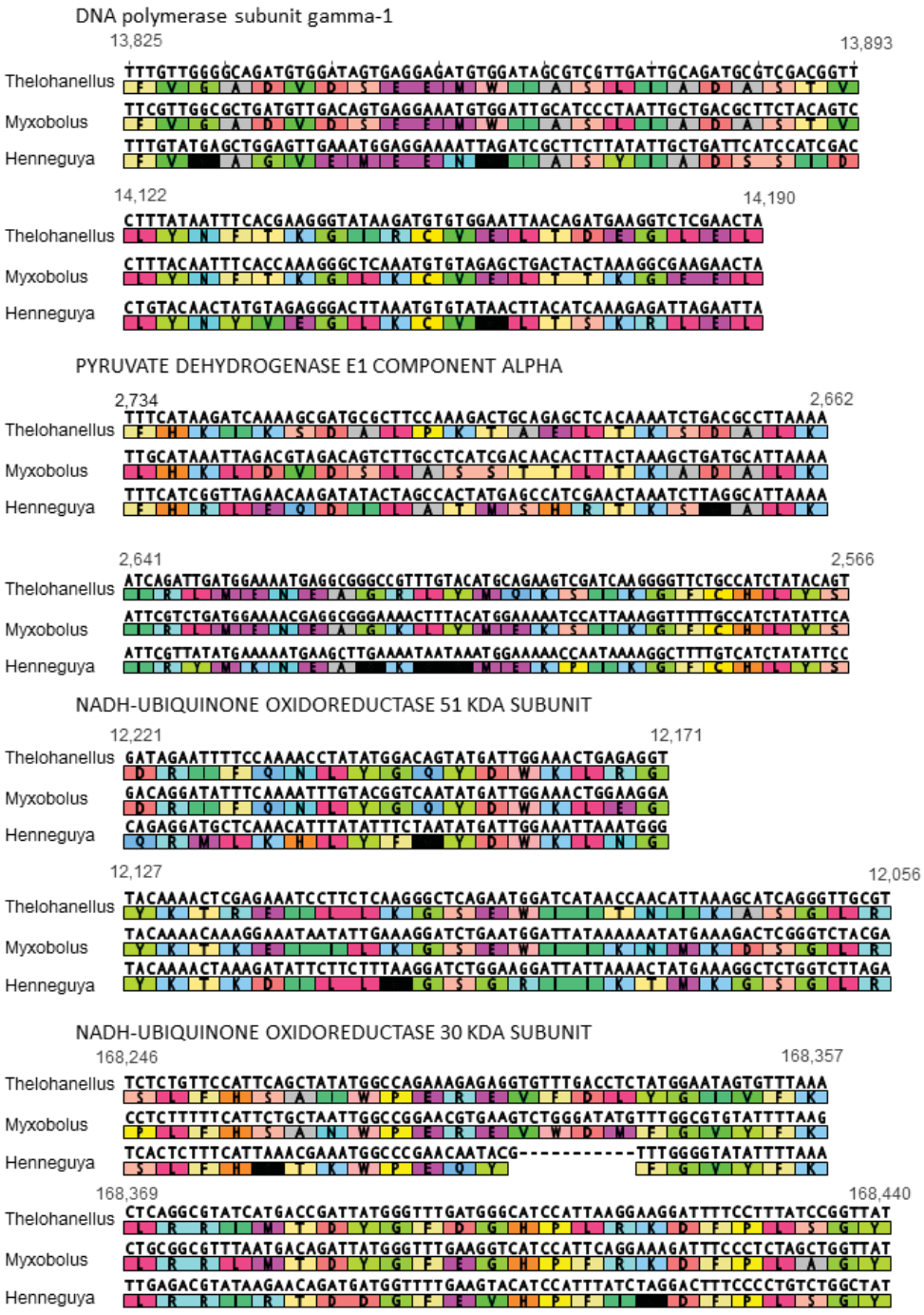


Fig. S7. Alignments of four nuclear encoded mt proteins genes revealing the presence of pseudogenes in *H. salminicola*. The species considered are the myxobolids *Thelohanellus kitauei*, *Myxobolus squamalis*, and *Henneguya salminicola*. The numbering of the DNA position, indicated above the DNA sequence, is based on the *Thelohanellus kitauei* sequence. Amino acid sequences are indicated below the DNA sequences where Stop codons are shown in black and deletion are marked by a dash.

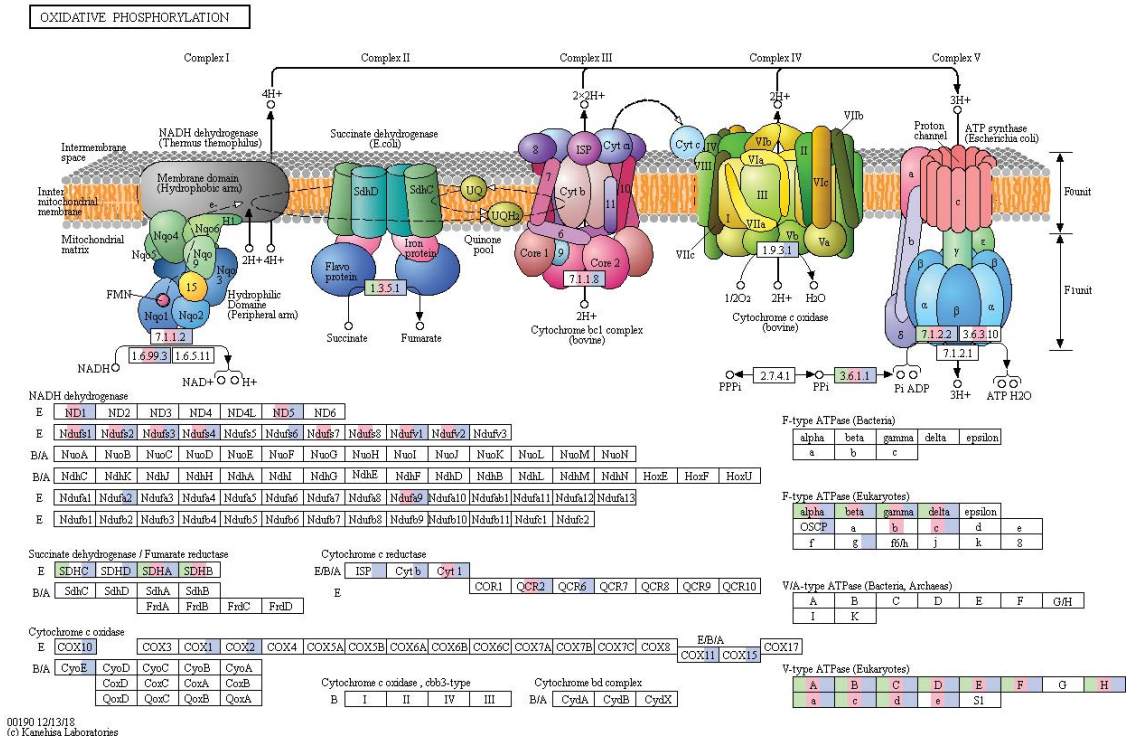


Fig. S8. Oxidative phosphorylation pathways of *H*. *salminicola*, *M*. *squamalis* and *K*. *iwatai*. The proteins identified in the mitochondrial genome and the transcriptome of *H*. *salminicola*, *M*. *squamalis* and *K*. *iwatai* are indicated in green, red and blue, respectively. The absence of a protein is indicated in white. Results corresponding to pseudogenes in *H*. *salminicola* (i.e., NuoI/comp_48827 and NduFs1/comp_72978) were removed and therefore appear white in this figure.

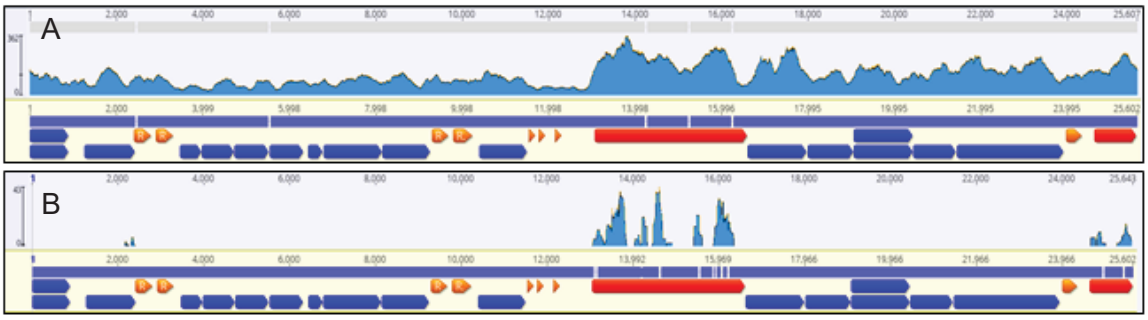


Fig. S9. Per-base coverage analysis of the *M*. *squamalis* mt chromosome. Illumina reads mapped with Bowtie2 (12) to the *Myxobolus squamalis* mt genome (Accession: MK087050). The mapping of the genomic reads (A) is presented above the mapping of the transcriptomic genes (B). The rnl and rns genes are indicated by red arrows while protein and repeats are indicated by blue and orange arrows, respectively.

Additional data table S1 (separate file). Genes involved in mitochondrial metabolic pathways (other than those involved in replication and translation of the mt genome).

NCBI accession numbers of the *Drosophila* protein query and corresponding blast hits in the *Hydra* and myxozoan genomes, proteomes and transcriptomes (see *SI Methods*). The annotation for class, function, and gene originated from the MitoDrome database (http://mitodrome.ba.itb.cnr.it/). Genes involved in more than one class/pathway have further information present in the "other class" columns. Sequences identified as pseudogenes are indicated in orange.

Additional data table S2 (separate file). Nuclear genes involved in the replication and translation of the mitochondrial genome.

NCBI accession numbers of *Drosophila* and human protein queries and corresponding blast hits in the genomes, transcriptomes and Maker-generated proteomes of *Hydra*, myxozoans, and *Polypodium* (see *SI Methods*). Protein annotations originated from flybase (https://flybase.org/) Sequences identified as pseudogenes are indicated in orange.

Additional data table S3 (separate file). Enzymes involved in anaerobic metabolism in protists.

Enzyme name, Accession number and species of sequences used as query against the proteome and transcriptome of *H. salminicola*.

References

1. D. M. Hillis, M. T. Dixon, Ribosomal DNA: molecular evolution and phylogenetic inference. *Q*. *Rev*. *Biol*. **66**, 411-453 (1991).
2. S. L. Hallett, A. Diamant, Ultrastructure and small-subunit ribosomal DNA sequence of *Henneguya lesteri* n. sp. (Myxosporea), a parasite of sand whiting *Sillago analis* (Sillaginidae) from the coast of Queensland, Australia. *Dis*. *Aquat*. *Org*. **46**, 197-212 (2001).
3. A. Diamant, C. M. Whipps, M. L. Kent, A new species of *Sphaeromyxa* (Myxosporea: Sphaeromyxina: Sphaeromyxidae) in devil firefish, *Pterois miles* (Scorpaenidae), from the northern Red Sea: morphology, ultrastructure, and phylogeny. *J*. *Parasitol*. **90**, 1434-1442 (2004).
4. C. M. Whipps *et al*., First report of three *Kudoa* species from eastern Australia: *Kudoa thyrsites* from mahi mahi (*Coryphaena hippurus*), *Kudoa amamiensis* and *Kudoa minithyrsites* n. sp. from sweeper (*Pempheris ypsilychnus*). *J*. *Eukaryot*. *Microbiol*. **50**, 215-219 (2003).
5. T. A. Hall, BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp*. *Ser* **41**, 95-98 (1999).
6. G. L. Hoffman, *Parasites of North American Freshwater Fishes, 2nd edition* (Cornell University Press, 1999), pp. 539.
7. D. M. L. Hervio *et al*., Taxonomy of *Kudoa* species (Myxosporea), using a small-subunit ribosomal DNA sequence. *Can*. *J*. *Zool*. **75**, 2112-2119 (1997).
8. M. L. Kent *et al*., Recent advances in our knowledge of the Myxozoa. *J*. *Eukaryot*. *Microbiol*. **48**, 395-413 (2001).
9. T. M. Polley, S. D. Atkinson, G. R. Jones, J. L. Bartholomew, Supplemental description of *Myxobolus squamalis* (Myxozoa). *J*. *Parasitol*. **99**, 725-728 (2013).
10. S. Andrews, *FastQC: a quality control tool for high throughput sequence data*. *Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc* (2010).
11. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17** (2011).
12. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat*. *Methods* **9**, 357-359 (2012).
13. K. A. Christensen *et al*., Chinook salmon (*Oncorhynchus tshawytscha*) genome and transcriptome. *PLoS ONE* **13**, e0195461 (2018).
14. Y. Peng, H. C. Leung, S. M. Yiu, F. Y. Chin, IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420-1428 (2012).
15. C. Camacho *et al*., BLAST+: architecture and applications. *BMC Bioinform*. **10**, 421 (2009).
16. S. Lien *et al*., The Atlantic salmon genome provides insights into rediploidization. *Nature* **533**, 200-205 (2016).
17. C. Berthelot *et al*., The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat*. *Commun*. **5**, 3657 (2014).

18. W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659 (2006).
19. M. G. Grabherr *et al*., Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat*. *Biotechnol*. **29**, 644-652 (2011).
20. B. Li, C. Dewey, RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform*. **12**, 323 (2011).
21. G. Parra, K. Bradnam, I. Korf, CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061-1067 (2007).
22. D. M. Bryant *et al*., A tissue-mapped axolotl *de novo* transcriptome enables identification of limb regeneration factors. *Cell Rep*. **18**, 762-776 (2017).
23. M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, K. Morishima, KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res*. **45**, D353-D361 (2017).
24. S. E. Chang *et al*., Genomic insights into the evolutionary origin of Myxozoa within Cnidaria. *Proc*. *Natl*. *Acad*. *Sci*. *U.S.A*. **112**, 14912–14917 (2015).
25. C. Holt, M. Yandell, MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinform*. **12**, 491 (2011).
26. I. Korf, Gene finding in novel genomes. *BMC Bioinform*. **5**, 59 (2004).
27. M. Stanke, A. Tzvetkova, B. Morgenstern, AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biol*. **7 Suppl 1**, S11.11–S11.18 (2006).
28. A. Török *et al*., The cnidarian *Hydractinia echinata* employs canonical and highly adapted histones to pack its DNA. *Epigenetics Chromatin* **9**, 36 (2016).
29. R. M. Waterhouse *et al*., BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol*. *Biol*. *Evol*. **35**, 543–548 (2018).
30. G. Gremme, S. Steinbiss, S. Kurtz, *GenomeTools*: A comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Trans*. *Comput*. *Biol*. *Bioinform*. **10**, 645-656 (2013).
31. D. D'Elia *et al*., The MitoDrome database annotates and compares the OXPHOS nuclear genes of *Drosophila melanogaster*, *Drosophila pseudoobscura* and *Anopheles gambiae*. *Mitochondrion* **6**, 252-257 (2006).
32. M. Morgenstern *et al*., Definition of a high-confidence mitochondrial proteome at quantitative scale. *Cell Rep* **19**, 2836-2852 (2017).
33. P. Dehal, J. L. Boore, Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol*. **3**, e314 (2005).
34. A. J. Rey, H. Attrill, S. J. Marygold, C. FlyBase, Using FlyBase to find functionally related *Drosophila* genes. *Methods Mol*. *Biol*. **1757**, 493-512 (2018).
35. O. Rackham, T. R. Mercer, A. Filipovska, The human mitochondrial transcriptome and the RNA-binding proteins that regulate its expression. *Wiley Interdiscip*. *Rev*. *RNA* **3**, 675-695 (2012).
36. M. D. Adams *et al*., The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185-2195 (2000).
37. A. J. Roger, S. A. Muñoz-Gómez, R. Kamikawa, The origin and diversification of mitochondria. *Curr*. *Biol*. **27**, R1177-R1192 (2017).

38. M. Müller *et al*., Biochemistry and evolution of anaerobic energy metabolism in eukaryotes. *Microbiol*. *Mol*. *Biol*. *Rev*. **76**, 444-495 (2012).
39. X.-D. Fu, Exploiting the hidden treasure of detained introns. *Cancer Cell* **32**, 393-395 (2017).
40. C. Huang, D. Leng, S. Sun, X. D. Zhang, Re-analysis of the coral *Acropora digitifera* transcriptome reveals a complex lncRNAs-mRNAs interaction network implicated in *Symbiodinium* infection. *BMC Genomics* **20**, 48 (2019).
41. Y. Moran *et al*., Intron retention as a posttranscriptional regulatory mechanism of neurotoxin expression at early life stages of the starlet anemone *Nematostella vectensis*. *J*. *Mol*. *Biol*. **380**, 437-443 (2008).
42. J. Mistry, R. D. Finn, S. R. Eddy, A. Bateman, M. Punta, Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res*. **41**, e121 (2013).
43. V. Muthye, D. V. Lavrov, Characterization of mitochondrial proteomes of nonbilaterian animals. *IUBMB Life* **70**, 1289-1301 (2018).
44. J. A. Chapman *et al*., The dynamic genome of *Hydra*. *Nature* **464**, 592-596 (2010).
45. Y. Yang *et al*., The genome of the myxosporean *Thelohanellus kitauei* shows adaptations to nutrient acquisition within its fish host. *Genome Biol*. *Evol*. **6**, 3182-3198 (2014).
46. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol*. *Biol*. *Evol*. **30**, 772-780 (2013).
47. D. Yahalomi *et al*., The multipartite mitochondrial genome of *Enteromyxum leei* (Myxozoa): eight fast-evolving megacircles. *Mol*. *Biol*. *Evol*. **34**, 1551-1556 (2017).
48. N. Dierckxsens, P. Mardulyn, G. Smits, NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res*. **45**, e18 (2017).
49. F. Takeuchi *et al*., The mitochondrial genomes of a myxozoan genus *Kudoa* are extremely divergent in Metazoa. *PLoS ONE* **10**, e0132030 (2015).
50. M. Bernt *et al*., MITOS: Improved de novo metazoan mitochondrial genome annotation. *Mol*. *Phylogenet*. *Evol*. **69**, 313-319 (2013).
51. A. Krogh, B. Larsson, G. von Heijne, E. L. L. Sonnhammer, Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J*. *Mol*. *Biol*. **305**, 567-580 (2001).
52. N. Lartillot, T. Lepage, S. Blanquart, PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25**, 2286-2288 (2009).
53. I. Fiala, P. Bartošová-Sojková, C. M. Whipps, "Classification and phylogenetics of Myxozoa" in *Myxozoan evolution, ecology and development,* B. Okamura, A. Gruhl, J. L. Bartholomew, Eds. (Springer International Publishing, 2015), pp. 85-110.
54. G. W. Vurture *et al*., GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202-2204 (2017).