Title: A new lineage of eukaryotes illuminates early mitochondrial genome reduction

Authors: Jan Janouškovec[1,2,3,6,*,#], Denis V. Tikhonenkov[3,4,*,#], Fabien Burki[3,5] , Alexis T. Howe[3], Forest L. Rohwer[2], Alexander P. Mylnikov[4], Patrick J. Keeling[3,*]

Author affiliations:
[1] University College London, Department of Genetics, Evolution and Environment, London, UK
[2] San Diego State University, Biology Department, San Diego, CA, USA
[3] University of British Columbia, Botany Department, Vancouver, BC, Canada
[4] Institute for Biology of Inland Waters, Russian Academy of Sciences, Borok, Russia
[5] Science for Life Laboratory, Program in Systematic Biology, Uppsala University, Uppsala, Sweden
[6] Lead Contact
[*] Correspondence: janjan.cz@gmail.com (JJ), tikho-denis@yandex.ru (DVT), pkeeling@mail.ubc.ca (PJK)
[#] authors contributed equally

Keywords: origin of eukaryotes; mitochondrial genome evolution; cytochrome *c* maturation; gene transfer; phylogenomics; microbial diversity; cell ultrastructure; ancoracyst; *Ancoracysta twista*

SUMMARY

The origin of eukaryotic cells represents a key transition in cellular evolution and is closely tied to outstanding questions about mitochondrial endosymbiosis [1,2]. For example, gene-rich mitochondrial genomes are thought to be indicative of an ancient divergence, but this relies on unexamined assumptions about endosymbiont-to-host gene transfer [3–5]. Here, we characterize *Ancoracysta twista*, a new predatory flagellate that is not closely related to any known lineage in 201-protein phylogenomic trees and has a unique morphology, including a novel type of extrusome (ancoracyst). The *Ancoracysta* mitochondrion has a gene-rich genome with a coding capacity exceeding all other eukaryotes except the distantly related jakobids and *Diphylleia*, and uniquely possesses heterologous, nucleus- and mitochondrion-encoded cytochrome *c* maturase systems. To comprehensively examine mitochondrial genome reduction, we also assembled mitochondrial genomes from picozoans and colponemids, and re-annotated existing mitochondrial genomes using HMM gene profiles. This revealed over a dozen previously overlooked mitochondrial genes at the level of eukaryotic supergroups. Analyzing trends over evolutionary time demonstrates that gene transfer to the nucleus was non-linear, occurred in waves of exponential decrease, and that much of it took place comparatively early, massively independently, and with lineage-specific rates. This process has led to differential gene retention, suggesting that gene-rich mitochondrial genomes are not a product of their early divergence. Parallel transfer of mitochondrial genes and their functional replacement by new nuclear factors are important in models for the origin of eukaryotes, especially as major gaps in our knowledge of eukaryotic diversity at the deepest level remain unfilled.

RESULTS AND DISCUSSION

A new lineage of predatory eukaryote with unique ultrastructural traits

We surveyed predatory protists in a sample collected from the surface of tropical aquarium brain coral. One subculture, maintained by us for one year, was an unusual flagellate formally described here as *Ancoracysta twista* (STAR Methods). *Ancoracysta* are small, ovoid cells that actively feed on the marine bodonid *Procryptobia sorokini*, and could not be sustained on bacteria alone (Figures 1A,B and S1A-L). Transmission electron microscopy reveals a complex, four-layer envelope (theca) covered by a thin, dense glycocalyx (Figure 1C-F). The cells move quickly with a twirling motion and rapid directional changes (Movie S1) propelled by two heterodynamic flagella: the anterior flagellum has fine hairs (mastigonemes) on the proximal end (Figure S1O) and often curves to the dorsal cell surface, whereas the posterior flagellum has a short vane (Figure 1G) and runs through a longitudinal ventral groove with the vane against the groove (Figures 1A-D and S1T,U). Intact prey is ingested via a cytostome and transferred through a cytopharynx to a conspicuous posterior food vacuole (Figures 1A-C and S1A,F-K). The cell has several conventional mitochondria with lamellar cristae (Figures 1H and S1X). The cytoplasm also contains a new type of vesicle-enclosed extrusome, the ancoracyst, which are likely discharged to immobilize prey. The extrusome has an amphora-shaped base containing seven cylinders around a central shaft, a striated neck and an anchor-shaped cap with seven structured sectors (Figures 1C,D,I-L and S1Q,U). Taken together, *A. twista* is structurally unique but also shares similarities in the organization of the flagella, longitudinal groove, pellicle and extrusomes with other poorly-studied eukaryotes: colponemids, jakobids, cryomonads, *Telonema*, *Metopion*, and *Metromonas* (see STAR Methods for discussion and Figures 1 and S1) [6–14].

The phylogenetic relationship of *A. twista* to other eukaryotes was assessed using small and large subunit ribosomal RNA genes, but these failed to resolve a close relationship with any other lineage with statistical support (Figure S2). To examine its phylogenetic position more deeply, we sequenced and assembled the *Ancoracysta* transcriptome by Illumina RNA-seq and added *Ancoracysta* data to a combined matrix of 201 conserved proteins previously used to address the relationships between major groups of eukaryotes (STAR Methods) [15]. Maximum Likelihood and Bayesian inferences from this 201-gene dataset again failed to place *Ancoracysta* within any known group of eukaryotes and showed conflicting positions even when using site-heterogeneous models (Figure 2). The Maximum Likelihood LG+Γ4+F+C60 model supported a sister relationship to a grouping of haptophytes and centrohelids with high support (99% ultrafast bootstrap and 97% non-parametric PMSF bootstraps; Figure 2A; STAR Methods). The Bayesian GTR+CAT+Γ4 placed *Ancoracysta* at even deeper position in the same general part of the tree (Figure 2B) although incomplete convergence of the MCMC chains was obtained and this topology was rejected by four statistical tests under the LG+Γ4+F+C60 model at the significance level of 0.05 (Figure 3C; STAR Methods). A deep, unresolved position of *Ancoracysta* was also found by using a set of 36 concatenated mitochondrial proteins (Figure S3). Relationships between taxa in this part of the eukaryotic tree have been difficult to resolve [15–18], and the exact phylogenetic position of *Ancoracysta* remains uncertain based on available genomic data, but the phylogenies do show that it does not branch within any currently recognized lineage.

Gene-rich mitochondrial genome and redundant cytochrome *c* maturation systems

The most gene-rich, bacterial-like mitochondrial genomes are found in the jakobids [3,19], and this ancestral state has led to suggestions that they represent an early-diverging eukaryotic lineage, branching somewhere near the root of the eukaryotic tree [2–5]. But reconciling this conclusion with other lines of evidence has been difficult [20], primarily because we do not really understand the process of gene transfer and genome reduction during early mitochondrial evolution, or even the patterns of gene presence and absence across eukaryotic diversity. Since *Ancoracysta* has no close affinities to other groups, we sequenced its complete mitochondrial genome, resulting in a circular-mapping, 52.7 kb genome with a 2 kb inverted repeat (Figure 3A; STAR Methods). The genome is gene-rich, containing 47 conserved protein-coding genes, nine unidentified open reading frames, three ribosomal RNA genes (including the 5S rDNA), and a large set of 25 tRNA genes, which can decode all codons in the genome (UGA encodes tryptophan). Mitochondrial genomes of other organisms that branch in the same overall region of the tree as *Ancoracysta* tend to encode many fewer genes (e.g., *Ancoracysta* encodes three ribosomal proteins, *rpl11*, *rpl19* and *rpl27* found in no mitochondrial genomes of these lineages: Figures 3A and 4A). Indeed, the gene-richness of *Ancoracysta* mitochondrion is exceeded only by those of the distantly related jakobids and *Diphylleia* (Figure 3B) [21].

Interestingly, the mitochondrial genome of *Ancoracysta* encodes four subunits of the bacterial cytochrome *c* maturation System I (*ccmA*, *ccmB*, *ccmC* and *ccmF;* Figure 3A), which is rare and patchily distributed in mitochondria. Most mitochondria instead use a nucleus-encoded holocytochrome *c* synthase (HCCS, System III), whose distribution is mutually exclusive with System I [22]. System I was apparently ancestral to all eukaryotes and was functionally replaced by the HCCS (which is absent in prokaryotes), but whether this happened once or multiple times is an unresolved debate, the implications of which have been hypothesized to apply to rooting the tree of eukaryotes [2]. Distinguishing between these scenarios has been difficult because the HCCS phylogeny is largely unresolved and no eukaryote has been found to contain both systems [23]. However, a unique HCCS is present in the *Ancoracysta* transcriptome (which does not appear to be a culture contaminant, see STAR Methods; Figure 3C), so it represents the first organism to contain both System I and III. Moreover, the *Ancoracysta* HCCS is closely related to homologues from stramenopiles, which are not themselves monophyletic (Figures 3C and S4A), suggesting that HCCSs can spread between distantly related eukaryotes horizontally. *Ancoracysta* could therefore represent a rare state of redundancy between the HCCS and System I (which is also, at least partly, transcribed in *Ancoracysta*; STAR Methods), possibly as part of an ongoing functional replacement. Topological inconsistencies in the HCCS phylogeny, the rarity of Systems I and III co-existence and the presence of HCCS in at least one organism that was predicted to never contain it (*Percolomonas cosmopolitus*; Figure S4A) all argue for horizontal spread of the HCCS and against its single ancestral acquisition pointing to a position of the eukaryotic root [2].

The early mitochondrial genome reduction

The presence of gene-rich mitochondrial genomes in apparently distantly-related lineages prompted us to investigate mitochondrial genome reduction more comprehensively by mapping early changes in

gene content. We first assembled new mitochondrial genomes from two additional deep-branching groups of eukaryotes, picozoans and colponemids, by using genome sequence surveys data generated previously (STAR Methods; Figure S4B) [24,25]. The complete mitochondrial genome of the uncultivated picozoan species MS584-11 is a circular-mapping molecule of 49kb with moderate gene density. The large subunit ribosomal RNA gene (*rrl*) is split into two widely-separated fragments (one contains two introns), suggesting that it is either trans-spliced or assembles by base-pairing (Figure S4B). The partial mitochondrial genome of *Colponema vietnamica* is 25kb in size, highly compacted and encodes a split *nad5*, a feature previously observed in other alveolates [25]. Comparing the mitochondrial genomes across eukaryotes revealed that more genes generally not found in related lineages were present in colponemids and picozoans: *sdh3* and *rpl32* in *Colponema,* and *rps1* in the picozoan (Figure S4B).

To see how common "rare" mitochondrial genes might really be, we re-annotated existing mitochondrial genomes by using custom Hidden Markov Model (HMM) profiles generated from alignments of all mitochondrial genes (STAR Methods). New candidate genes were verified by searches against public HMM and protein fold databases, and by inspecting potential synteny, which revealed ten strong cases and four weaker cases of previously overlooked genes (Table S1). Combining these with the three new mitochondrial genomes described here shows that ancestral mitochondrial genomes of at least nine supergroups were more gene-rich than previously thought (Figure 4A). This has significant implications for how these genes are interpreted. For example, *rps16* was previously known only in amoebozoan mitochondrial genomes, but is now identified in two other deep lineages, malawimonads and apusomonads. It forms a syntenic cluster with *rpl19* in malawimonads, amoebozoans, and proteobacteria (Figure 4B) [5], suggesting they are retained from an ancestral operon. Overall, the broad but patchy distribution of rare mitochondrial genes (which does not appear to be due to horizontal transfer according to the phylogenies of these genes; STAR Methods) is indicative of widespread instances of parallel gene transfers to the nucleus in distantly related lineages. This is most apparent in our re-analysis of jakobid mitochondrial genomes (Figure 4C): while they are gene-rich, in fact only one mitochondrion-encoded gene in jakobids (*rpl1*) has been transferred to the nucleus of all other eukaryotes, and they lack three genes that are found in mitochondrial genomes of other supergroups (*rps16*, *rpl36* and *rpl23*). The *rpo* genes and *secY* are also unique to jakobid mitochondrial genomes, but have never been transferred to the nucleus (the former has been functionally substituted by a nucleus-encoded T3/T7 phage-like RNA polymerase, which is also present in the *A. twista* transcriptome).

Projecting changes in mitochondrially encoded proteins onto a time-calibrated phylogeny of eukaryotes [26,27] reveals that the rate of gene transfer to the nucleus changed profoundly in time (Table S2). Much of the gene transfer occurred early, partly before and partly after the last eukaryotic common ancestor, but in either case prior to the origin of eukaryotic supergroups (primary reduction; Figure 4D,E). We note that transfers at this early stage of mitochondrial reduction were mostly stochastic with respect to gene function. Fitting of two competing, biologically plausible functions through mitochondrial coding capacity projected in time strongly favoured exponential over linear decay (F-test p-value=2.83e-06 for all data points). In four exemplary lineages the exponential

reduction was even more pronounced (F-test p-values=2.04e-3 to 7.05e-15), with notably similar curvature and horizontal shift coefficients but with a lineage-specific vertical shift (Figure 4E; STAR Methods). Since multiple independent transfers are probably underestimated, the projected gene transfer numbers are minimal estimates, but given the saturation of sampling within some groups and the relatively low rate of new gene discovery in mitochondrial genomes in general, the difference is unlikely to change the observed patterns. In some groups, such as animals, euglenozoans and myzozoans, steep secondary reductions in mitochondrial genome content also took place, which once again likely represent exponential decrease. In most mitochondria, however, the later stage of reduction is characterized by slow, ongoing transfer of individual genes (Figure 4D,E).

Altogether (Figure 4), these patterns show that much of the gene transfer from the mitochondrion to the nucleus occurred comparatively early and was massively parallel in nature. Parallel transfers over shorter time frames have similarly been observed and are perhaps best studied in angiosperm mitochondria [28–30]. Here, most lineages have static gene content, but certain lineages have undergone many gene transfers to the nucleus, resulting in highly uneven rates of transfer from stasis to transfer frequencies higher than that of synonymous substitutions. Moreover, these linages are patchily distributed on the phylogeny of plants, meaning many genes have been relocated in parallel [28–30]. Reconciling this with the long-term patterns described here, it seems that certain lineages of angiosperm mitochondria have recently entered a state of secondary reduction.

There remains a relatively well-conserved core of mitochondrial genes that are rarely transferred [4,31,32], and as this core is approached transfer rates would typically slow down since there may be a variety of barriers to their transfer or targeting [32,33]. What triggers accelerated reduction is not known; it may be precipitated by the transfer of key mitochondrial genes that play a prominent role in the mitochondrial interactome and regulatory networks, by changes to the targeting machinery [19], changes to intrinsic characteristics of the genome such as substitution frequencies [34], or a mix of factors. The retention of gene-rich mitochondrial genomes may equally be due to a mixture of causes. On one hand the stochastic nature of gene transfer means some genomes will simply be larger by chance, with GC content and protein hydrophobicity possibly affecting gene transfer success [32], but on the other hand larger genomes could result from a lack of triggers for exponential decay. For example, jakobid genomes may be large due to their continued reliance (by chance) on the bacterial polymerase or more likely on the bacterial SecY protein import system, whereas many secondary reductions are associated with the loss or functional replacement of whole respiratory chain complexes (e.g., in myzozoans, euglenozoans; Figure 4D), which are hypothesized to be intertwined with the redox regulation of mitochondrial gene expression [33]. Exponential reduction in other cases (e.g., in the ancestor of animals or the apparently recently initiated ongoing reduction in angiosperms) have less clear causes but could operate by similar mechanisms, which act to accelerate the much more gradual, parallel movement of mitochondrial genes to the nucleus as observed in most eukaryotes today (Figure 4D,E).

Implications

Combined morphological and molecular phylogenetic evidence demonstrates that *Ancoracysta twista* represents a newly discovered, deep-branching lineage of eukaryotes (Figures 1 and 2). Similarities in

cell structure between it and certain distantly-related taxa suggest *Ancoracysta* will be valuable in illuminating deep transitions in eukaryotic morphology, particularly concerning longitudinal grooves, complex pellicles, and extrusomes (Figure 1). Genomic data from *Ancoracysta* holds similar promise: its gene-rich mitochondrial genome and redundancy in cytochrome *c* maturation systems (Figures 3) already rejuvenate long-standing debates on where the root of eukaryotes lies and how mitochondrial genomes evolved. Our conclusion that neither gene-rich mitochondrial genomes nor the HCCS can be used to to pinpoint the root of eukaryotes highlights the importance of phylogenomic approaches to this question [20]. Our results also emphasize the importance of parallel gene transfer and exponential, lineage-specific decay of mitochondrial genomes over time (Figure 4). Further insight into this process will rely on understanding triggers for differential reduction of mitochondrial genomes in different groups and their still elusive diversity. The novelty of *A. twista*, whose existence remained undetected by two centuries of morphological inquiries and largely also by environmental clone libraries (a single related sequence was found; STAR Methods), emphasizes that our knowledge of eukaryotic diversity is far from complete and that more such organisms likely exist.

AUTHOR CONTRIBUTIONS

Conceptualization, JJ, DVT, APM, and PJK; Investigation, JJ, DVT, ATH, and APM; Formal analysis, JJ, DVT, and FB; Visualization, JJ and DVT; Supervision, FLR and PJK; Funding acquisition, DVT and PJK; Writing - original draft, JJ, DVT, and PJK; Writing - review and editing, all authors.

ACKNOWLEDGEMENTS

This work was supported by grants from the Canadian Institutes for Health Research (MOP-42517) to PJK, from the Russian Foundation for Basic Research (17-04-00899, 15-29-02518) to DVT. JJ was supported by a UCL Excellence Fellowship, CIFAR Global Scholar Fellowship, and UBC Four Year PhD Fellowship. FB was supported by a grant from the Tula Foundation to the Centre for Microbial Biodiversity and Evolution at UBC and is now a SciLifeLab fellow at Uppsala University. Electron microscopy investigation was supported by the Russian Science Foundation (14-14-00515).

REFERENCES

1. Pittis, A.A., and Gabaldón, T. (2016). Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. Nature *531*, 101–104.

2. Cavalier-Smith, T. (2010). Kingdoms Protozoa and Chromista and the eozoan root of the eukaryotic tree. Biol. Lett. *6*, 342–5.

3. Lang, B.F., Burger, G., O'Kelly, C.J., Cedergren, R., Golding, G.B., Lemieux, C., Sankoff, D., Turmel, M., and Gray, M.W. (1997). An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. Nature *387*, 493–497.

4. Gray, M.W., Lang, B.F., and Burger, G. (2004). Mitochondria of protists. Annu. Rev. Genet. *38*, 477–524.

5. Kannan, S., Rogozin, I.B., and Koonin, E.V. (2014). MitoCOGs: clusters of orthologous genes from mitochondria and implications for the evolution of eukaryotes. BMC Evol. Biol. *14*, 237.

6. Patterson, D.J. (1990). Jakoba libera (Ruinen, 1938), a heterotrophic flagellate from deep oceanic sediments. J. Mar. Biol. Assoc. U. K. *70*, 381–393.

7. Thomsen, H.A., Buck, K.R., Bolt, P.A., and Garrison, D.L. (1991). Fine structure and biology of Cryothecomonas gen. nov. (Protista incertae sedis) from the ice biota. Can. J. Zool. *69*, 1048–1070.

8. Mylnikov, A.P., Mylnikova, Z.M., and Tsvetkov, A.I (1999). The ultrastructure of the marine carnivorous flagellate Metopion fluens. Tsitologiya *41*, 581–585.

9. Mylnikov, A.P., and Tikhonenkov, D.V (2009). The new alveolate carnivorous flagellate (Colponema marisrubri sp. n., Colponemida, Alveolata) from the Red Sea. Zool. Zhurnal *88*, 1163–1169.

10. Myl'nikova, Z.M., and Myl'nikov, A.P. (2010). Biology and morphology of freshwater rapacious flagellate Colponema aff. loxodes Stein (Colponema, Alveolata). Inland Water Biol. *3*, 21–26.

11. Myl'nikova, A.A., and Myl'nikov, A.P. (2011). Ultrastructure of the marine predatory flagellate Metromonas simplex Larsen et Patterson, 1990 (Cercozoa). Inland Water Biol. *4*, 105–110.

12. Yabuki, A., Eikrem, W., Takishita, K., and Patterson, D.J. (2013). Fine Structure of Telonema subtilis Griessmann, 1913: A Flagellate with a Unique Cytoskeletal Structure Among Eukaryotes. Protist *164*, 556–569.

13. Mylnikov, A.P., and Mylnikov, A.A. (2014). Structure of the flagellar apparatus of the bacterivorous flagellate Histiona aroides Pascher, 1943 (Jakobida, Excavata). Inland Water Biol. *7*, 331–337.

14. Tikhonenkov, D.V., Janouškovec, J., Mylnikov, A.P., Mikhailov, K.V., Simdyanov, T.G., Aleoshin, V.V., and Keeling, P.J. (2014). Description of Colponema vietnamica sp.n. and Acavomonas peruviana n. gen. n. sp., two new alveolate phyla (Colponemidia nom. nov. and Acavomonidia nom. nov.) and their contributions to reconstructing the ancestral state of alveolates and eukaryotes. PLOS ONE *9*, e95467.

15. Burki, F., Kaplan, M., Tikhonenkov, D.V., Zlatogursky, V., Minh, B.Q., Radaykina, L.V., Smirnov, A., Mylnikov, A.P., and Keeling, P.J. (2016). Untangling the early diversification of eukaryotes: a phylogenomic study of the evolutionary origins of Centrohelida, Haptophyta and Cryptista. Proc. R. Soc. B Biol. Sci. *283*, 20152802.

16. Burki, F., Okamoto, N., Pombert, J.-F., and Keeling, P.J. (2012). The evolutionary history of haptophytes and cryptophytes: phylogenomic evidence for separate origins. Proc. R. Soc. B Biol. Sci. *279*, 2246–2254.

17. Yabuki, A., Kamikawa, R., Ishikawa, S.A., Kolisko, M., Kim, E., Tanabe, A.S., Kume, K., Ishida, K., and Inagki, Y. (2014). Palpitomonas bilix represents a basal cryptist lineage: insight into the character evolution in Cryptista. Sci. Rep. *4*. doi:10.1038/srep04641.

18. Cavalier-Smith, T., Chao, E.E., and Lewis, R. (2015). Multiple origins of Heliozoa from flagellate ancestors: New cryptist subphylum Corbihelia, superclass Corbistoma, and monophyly of Haptista, Cryptista, Hacrobia and Chromista. Mol. Phylogenet. Evol. *93*, 331–362.

19. Tong, J., Dolezal, P., Selkrig, J., Crawford, S., Simpson, A.G.B., Noinaj, N., Buchanan, S.K., Gabriel, K., and Lithgow, T. (2011). Ancestral and Derived Protein Import Pathways in the Mitochondrion of Reclinomonas americana. Mol. Biol. Evol. *28*, 1581–1591.

20. Derelle, R., Torruella, G., Klimeš, V., Brinkmann, H., Kim, E., Vlček, Č., Lang, B.F., and Eliáš, M. (2015). Bacterial proteins pinpoint a single eukaryotic root. Proc. Natl. Acad. Sci. *112*, E693–E699.

21. Kamikawa, R., Shiratori, T., Ishida, K.-I., Miyashita, H., and Roger, A.J. (2016). Group II Intron-Mediated Trans -Splicing in the Gene-Rich Mitochondrial Genome of an Enigmatic Eukaryote, Diphylleia rotans. Genome Biol. Evol. *8*, 458–466.

22. Allen, J.W.A., Jackson, A.P., Rigden, D.J., Willis, A.C., Ferguson, S.J., and Ginger, M.L. (2008). Order within a mosaic distribution of mitochondrial c-type cytochrome biogenesis systems? FEBS J. *275*, 2385–402.

23. Nishimura, Y., Tanifuji, G., Kamikawa, R., Yabuki, A., Hashimoto, T., and Inagaki, Y. (2016). Mitochondrial Genome of *Palpitomonas bilix* : Derived Genome Structure and Ancestral System for Cytochrome *c* Maturation. Genome Biol. Evol. *8*, 3090–3098.

24. Yoon, H.S., Price, D.C., Stepanauskas, R., Rajah, V.D., Sieracki, M.E., Wilson, W.H., Yang, E.C., Duffy, S., and Bhattacharya, D. (2011). Single-cell genomics reveals organismal interactions in uncultivated marine protists. Science *332*, 714–7.

25. Janouškovec, J., Tikhonenkov, D.V., Mikhailov, K.V., Simdyanov, T.G., Aleoshin, V.V., Mylnikov, A.P., and Keeling, P.J. (2013). Colponemids represent multiple ancient alveolate lineages. Curr. Biol. *23*, 2546–52.

26. Eme, L., Sharpe, S.C., Brown, M.W., and Roger, A.J. (2014). On the Age of Eukaryotes: Evaluating Evidence from Fossils and Molecular Clocks. Cold Spring Harb. Perspect. Biol. *6*, a016139–a016139.

27. Sharpe, S.C., Eme, L., Brown, M.W.W., and Roger, A.J. (2015). Timing the Origins of Multicellular Eukaryotes Through Phylogenomics and Relaxed Molecular Clock Analyses. In Evolutionary Transitions to Multicellular Life, I. Ruiz-Trillo and A. M. Nedelcu, eds. (Dordrecht: Springer Netherlands), pp. 3–29. Available at: http://link.springer.com/10.1007/978-94-017-9642-2_1.

28. Adams, K.L., Daley, D.O., Qiu, Y.-L., Whelan, J., and Palmer, J.D. (2000). Repeated, recent and diverse transfers of a mitochondrial gene to the nucleus in flowering plants. Nature *408*, 354–357.

29. Adams, K.L., Qiu, Y.-L., Stoutemyer, M., and Palmer, J.D. (2002). Punctuated evolution of mitochondrial gene content: High and variable rates of mitochondrial gene loss and transfer to the nucleus during angiosperm evolution. Proc. Natl. Acad. Sci. *99*, 9905–9912.

30. Adams, K.L., and Palmer, J.D. (2003). Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. Mol. Phylogenet. Evol. *29*, 380–395.

31. Maier, U.-G., Zauner, S., Woehle, C., Bolte, K., Hempel, F., Allen, J.F., and Martin, W.F. (2013). Massively Convergent Evolution for Ribosomal Protein Gene Content in Plastid and Mitochondrial Genomes. Genome Biol. Evol. *5*, 2318–2329.

32. Johnston, I.G., and Williams, B.P. (2016). Evolutionary Inference across Eukaryotes Identifies Specific Pressures Favoring Mitochondrial Gene Retention. Cell Syst. *2*, 101–111.

33. Allen, J.F. (2003). Why Chloroplasts and Mitochondria Contain Genomes. Comp. Funct. Genomics *4*, 31–36.

34. Lynch, M., Koskella, B., and Schaack, S. (2006). Mutation pressure and the evolution of organelle genomic architecture. Science *311*, 1727–30.

35. Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., *et al.* (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. *8*, 1494–1512.

36. Boisvert, S., Laviolette, F., and Corbeil, J. (2010). Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. J. Comput. Biol. *17*, 1519–33.

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

38. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313.

39. Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol. Biol. Evol. *32*, 268–274.

40. Lartillot, N., Rodrigue, N., Stubbs, D., and Richer, J. (2013). PhyloBayes MPI: Phylogenetic Reconstruction with Infinite Mixtures of Profiles in a Parallel Environment. Syst. Biol. *62*, 611–615.

41. Finn, R.D., Clements, J., and Eddy, S.R. (2011). HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. *39*, W29-37.

42. Soding, J., Biegert, A., and Lupas, A.N. (2005). The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. *33*, W244–W248.

43. Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N., and Sternberg, M.J.E. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. Nat. Protoc. *10*, 845–858.

44. Mignot, J.., and Brugerolle, G. (1975). Étude ultrastructurale du flagelle phagotrophe Colponema loxodes Stein. Protistologica *XI*, 429–444.

45. Hausmann, K. (1978). Extrusive Organelles in Protists. Int. Rev. Cytol. *52*, 197–276.

46. Mylnikov, A.P., and Mylnikov, A.A (2008). The structure of extrusive organelles in alveolate and other heterotrophic flagellates. Tsitologiya *50*, 406–412.

47. Luft, J.H. (1961). Improvements in Epoxy Resin Embedding Methods. J. Cell Biol. *9*, 409–414.

48. Wang, H.-C., Minh, B.Q., Susko, E., and Roger, A.J. (2017). Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. Syst. Biol. Available at: https://doi.org/10.1093/sysbio/syx068.

49. Cury, J.C., Araujo, F.V., Coelho-Souza, S.A., Peixoto, R.S., Oliveira, J.A.L., Santos, H.F., Dávila, A.M.R., and Rosado, A.S. (2011). Microbial Diversity of a Brazilian Coastal Region Influenced by an Upwelling System and Anthropogenic Activity. PLoS ONE *6*, e16553.

50. Zhao, S., Shalchian-Tabrizi, K., and Klaveness, D. (2013). Sulcozoa revealed as a paraphyletic group in mitochondrial phylogenomics. Mol. Phylogenet. Evol. *69*, 462–468.

FIGURE LEGENDS

Figure 1. Morphology and ultrastructure of *Ancoracysta twista*. (A, B) – light micrographs of typical rigid, not flattened cells; (C, D) – longitudinal sections of the cell; (E, F) – cell coverings include the plasmalemma and envelope (theca) consist of three external layers and one lamellar layer divided by wide space, which is filled with vesicles in some places; (G) – conventional 9+2 microtubule structure in cross section and vane of the posterior flagellum containing amorphous material; (H) – mitochondrion with lamellar cristae; (I-L) – extrusomes (ancoracysts) including longitudinal sections of intact (I) and discharged (L) organelles and transverse sections across the amphoroid base (J) and anchor-shaped cap (K). Abbreviations: ab – amphoroid base of extrusome, ac – anchor-shaped cap of extrusome, af – anterior flagellum, cp – cytopharynx, cs – cytostome, en – cell envelope, ev – enveloping vesicle, ex – extrusome, fd – fold, fv – food vacuole, gr – groove, m – mitochondrion, n – nucleus, pf – posterior flagellum, pl – plasmalemma, rs – reserve substance, sn - striated neck of extrusome, vs – vesicles. Scale bars: A, B – 10 μm; C, D – 1 μm; E – 0,5 μm; F, G, K – 0,1 μm; H–J, L – 0,2 μm. See Figure S1 and Movie S1 for additional data on *A. twista* morphology.

Figure 2: *Ancoracysta* represents a new, deep-branching lineage of eukaryotes. (A) Maximum likelihood tree inferred in IQ-TREE by using the LG+Γ4+F+C60 model. Values at branches correspond to ultrafast bootstraps (1000 replicates, LG+Γ4+F+C60 model), non-parametric bootstraps (100 replicates, LG+Γ4+F+C60+PMSF model) and Bayesian posterior probabilities (dashes = different topology, black dots = 100/100/1 support). (B) Phylobayes tree (CAT+GTR+Γ4 model; supergroups were abbreviated to the first three letters). Clades were shortened for clarity as indicated (in %). (C) Probabilities of tree topologies in parts A and B as evaluated by Kishino-Hasegawa (KH), Shimodaira-Hasegawa (SH), Expected Likelihood Weight (ELW) and approximately unbiased (AU) tests under the LG+Γ4+F+C60 model in IQ-TREE (STAR Methods). Unresolved deep-branching position of *A. twista* was also revealed by phylogenies of the ribosomal DNA operon and 36 concatenated mitochondrial genes (Figures S2 and S3).

Figure 3: A gene-rich mitochondrial genome in *Ancoracysta*. (A) The map of the *Ancoracysta* mitochondrial genome with genes colour-coded as to their function and the inverted repeat shaded in grey. Genes on the outside of circle are transcribed in the clockwise direction. System I cytochrome *c* maturation genes are shown with black arrows. (B) Protein coding capacity of mitochondrial genomes across eukaryotes. Orthologous sequences that are demonstrably ancestral to all mitochondria are

shown, and all species with >40 mitochondrial genome-encoded proteins are specifically listed (only representatives for groups with <40 genes are shown, for clarity). Species that are relatively closely related to *Ancoracysta* in the tree are shown in blue. (C) Maximum likelihood phylogeny of holocytochrome *c* synthase (LG+Γ4+F model, 1000 ultrafast bootstrap replicates, >50 are shown). Outgroup was shortened for clarity as indicated (in %). STR=Stramenopiles. Full tree in Figure S4A.

Figure 4: Mitochondrial genome reduction in eukaryotes. (A) New mitochondrial genes in eukaryotic supergroups as identified by Hidden Markov Model searches (see Table S1 for details) are shown by coloured dots: blue = high-confidence, purple = low-confidence, and grey = known genes. A reconstruction of ancestral mitochondrial protein-coding capacities in eukaryotes is shown below, where the impact of newly-identified genes is shown by "+" (including high-confidence genes only). (B) The newly identified mitochondrial *rps16* in *Malawimonas* is in a conserved operon arrangement with *rpl19* (green), also found in amoebozoans and bacteria (grey genes are never encoded in mitochondria). (C) Rare mitochondrial protein-coding genes (present in 3 or less supergroups) are highly differentially shared. Jakobid genes that are never nucleus-encoded are inside dashed lines. (D) Time-calibrated evolution of the mitochondrial protein complement reveals an early, exponential and lineage-specific decrease in coding capacity. Ancestral mitochondrial protein-coding gene numbers were projected by on the dated phylogeny in reference [26] by the mean date (filled circles) and connected by lines (the position of malawimonads is inconsistent with our phylogeny, dashed line). Supergroups are colour-coded and abbreviated to the first three letters. (E) Loss of protein-coding genes from the mitochondria on examples of four lineages can be explained by a simple exponential decrease function corresponding to primary and secondary reductions. See Table S2 for primary data.

STAR METHODS

CONTACT FOR RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Jan Janouškovec (janjan.cz@gmail.com).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Formal taxonomic diagnosis

Ancoracysta n. gen. Janouškovec, Tikhonenkov, Burki, Howe, Rohwer, Mylnikov et Keeling 2017
Assignment. Eukaryota
Diagnosis. Unicellular protist with two heterodynamic flagella inserted into separate pockets and covered by a thin layer of dense glycocalyx. The cell body is enclosed in a four-layer envelope (theca) with vesicles in the interspace. Mitochondrial cristae are lamellar. The extrusomes consist of an amphora-like structure with seven cylinders inside and an anchor-shaped cap subdivided into seven sectors.
Etymology. Named after unusual extrusomes resembling anchor ("ancora", Latin) in longitudinal section.
Zoobank Registration. urn:lsid:zoobank.org:act:5E86E67F-0194-45AB-ABDB-A4A9A599AA49.
Type species. *Ancoracysta twista*
Ancoracysta twista n. sp. Janouškovec, Tikhonenkov, Burki, Howe, Rohwer, Mylnikov et Keeling 2017
Diagnosis. Cell oval, 7-14 μm length, 3-5 μm width, rigid, not flattened. Large cytostome is situated at the anterior end of the cell. The anterior flagellum, 8-10 μm in length, curves to dorsal cell surface, and bears fine mastigonemes on the proximal end. The posterior flagellum, 12-14 μm, has a short vane filled with amorphous material and directed backward, which lies along the ventral side of the cell close to the longitudinal ventral groove. The cells are characterized by fast twirling movement and frequently change direction. The organism is eukaryovorous and takes up prey using the cytostome, forming a large food vacuole at the posterior end. The posterior end of starving specimens is pointed and does not contain the food vacuole. Cysts have not been found.
Type material. A block of chemically fixed resin-embedded cells of the type strain, TD-1, is deposited in Marine Invertebrate Collection, Beaty Biodiversity Museum, University of British Columbia as MI-PR211. This constitutes the name-bearing type of the new species (a hapantotype).
Figure 1A illustrates a live cell of strain TD-1.
Type locality. Unknown. Surface of tropical brain coral from an aquarium.
Etymology. The species name indicates twirling movement.
Gene sequence. The accession number for the 18S rRNA gene sequence is GenBank: MG202010.
Zoobank Registration. urn:lsid:zoobank.org:act:1FC88CDD-F805-47E9-A1A9-1523D8457EDA
Comparison. Two separate flagellar pockets and thin mastigonemes on the proximal end of the anterior flagellum are typical of *Colponema loxodes* and *C. marisrubri* [9,10,44]. Colponemid and jakobid species also have a posterior flagellum with a structurally similar vane (lacking paracrystals) positioned in the longitudinal ventral groove (albeit armored) [6,13,14]. The complex envelope of *A. twista* resembles to a great degree the cell coverings of *Colponema marisrubri* [9] and somewhat less those in the cryomonads [7]. *C. marisrubri* also carries a toxicyst that has the closest structural resemblance to the ancoracyst from all extrusomes [45,46]. Vesicle-embedded, sectorized extrusomes in *Metopion fluense*, *Metromonas simplex* and *Telonema subtilis* could be also related to the ancoracyst but are more dissimilar in their symmetry and cap structure [8,11,12].

METHOD DETAILS

Cell isolation, culturing and microscopy

A sample containing seawater and coral mucus was drawn from the surface of an unspecified species of tropical brain coral on 26 April 2010 at the Birch Aquarium at Scripps Institution of Oceanography, University of California San Diego, USA. The sample was incubated overnight and then serially diluted in 0.22um filtered aquarium water. Cells of *Ancoracysta twista* were found in one well of the culture plate after several days and the contents of the well were transferred into a parafilm-sealed petri dish of PES medium prepared from filtered autoclaved seawater and supplemented with 250-500 ul of 802 Sonneborn's *Paramecium* medium per liter of PES. The culture was stored at room temperature and further subcultured periodically until 8 September 2010 after which time two parafilm-sealed subsamples remained at room temperature for eight months. In April 2011, *Ancoracysta* cells were found in the plates and isolated in a clonal culture (strain TD-1) by using a glass micropipette. Cells of *Ancoracysta* TD-1 were subsequently propagated on the marine bodonid *Procryptobia sorokini* strain B-69 grown in Schmalz-Pratt's medium by using the bacterium *Pseudomonas fluorescens* as prey [14]. *Ancoracysta* actively preyed upon *Procryptobia* but could not be propagated on *Pseudomonas* alone; how it persisted in subsamples for eight months remains unclear. The *Ancoracysta twista* TD-1 culture was lost after one year of continuous passaging.

Light microscopy observations were made by using the Zeiss AxioScope A.1 equipped with a DIC contrast water immersion objective (63x) and the analog video camera AVT HORN MC-1009/S. For transmission electron microscopy (TEM), cells were centrifuged and fixed in a cocktail of 0.6% glutaraldehyde and 2% OsO_4 on Schmaltz-Pratt medium for 15–30 min at 1°C then dehydrated in alcohol and acetone series (30, 50, 70, 96, and 100%, 20 minutes in each step). Afterward, the cells were embedded in a mixture of Araldite and Epon [47]. Ultrathin sections were prepared with a LKB ultramicrotome (Sweden) and observed by using the JEM 1011 transmission electron microscope (JEOL, Japan).

DNA and RNA sequencing and assembly

Cells were harvested by slow centrifugation (1000 x *g*) of the culture through a 0.7 μm filter (Vivaclear; attention was paid not to stir up or include liquid from the bottom of the culture plate containing primarily the prey cells). Total genomic DNA and RNA were extracted from the filter by using the MasterPure Complete DNA and RNA Purification Kit (Epicentre) and RNAqueous-Micro Kit (Ambion), respectively. RNA was reverse-transcribed and amplified as double-strand cDNA by using the SMARTer Pico PCR cDNA Synthesis Kit (Clontech). Total genomic DNA of *Colponema vietnamica* Colp-7a was isolated as described previously [25]. Illumina paired-end 100 bp reads were generated from all three samples. Genomic reads of *Ancoracysta* and *Colponema vietnamica* Colp-7a were assembled *de novo* in Ray v2.0 [36] by using the default settings. Single cell-derived genomic reads and genomic read assembly of the picozoan sp. MS584-11 were acquired from the authors of a published study [24]. Transcriptomic reads of *Ancoracysta* were assembled *de novo* in Inchworm (Trinity v2.0.3) under default settings [35]. The SMARTer adapters were trimmed and 97111 contigs of the minimum size of 200 nucleotides were deposited in GenBank: PRJNA413804.

QUANTIFICATION AND STATISTICAL ANALYSIS

Nuclear gene mining, alignments and phylogenies

Ancoracysta sequences were included into the alignments of 263 proteins previously used in eukaryote-wide phylogenies [15,16]. Briefly, the *Ancoracysta* predicted proteins were searched by BLASTP at the evalue cutoff of 1e-20 by using all sequences in the individual alignments as queries and top four unique hits were retrieved. Amino-acid sequences were aligned in MAFFT-linsi v7.245 (default parameters) [37] and stripped of hypervariable sites in TRIMAL v1.4 (maximum number of gaps allowed per site: 20%). Single protein-trees were built by using RAxML v8 [38] with 100 rapid bootstraps and the LG+Γ4+F model, and non-orthologous proteins were removed from alignments upon manual inspection of the individual trees. The final matrix was limited to protein alignments in

which both *Ancoracysta* and at least 50 of the total 100 operational taxonomic units were present. This resulted in 201 protein alignments, which were concatenated into a supermatrix of 44349 amino acid sites. Maximum Likelihood analyses were performed with IQ-TREE v1.3.8 and v1.5.4 [39] under the LG+Γ4+F+C60 model with 1000 ultrafast bootstrap replicates. A second branch support metric was generated by computing 100 non-parametric bootstrap replicates by using the efficient posterior mean site frequency method (PMSF in IQ-TREE) [48] and the LG+Γ4+F+C60 tree as a guide tree to estimate PMSF profiles (Figure 2A). We also used the CAT+GTR+Γ4 topology (see below) as a guide tree in the Maximum Likelihood reconstruction with the LG+Γ4+F+C60+PMSF model: the *Ancoracysta* relationship to haptophytes and centrohelids was again recovered although with lower bootstrap support (70%, not shown). The Bayesian phylogeny was inferred in Phylobayes-MPI v1.5 [40] under the CAT+GTR+Γ4 running two independent Markov chain Monte Carlo (MCMC) chains with the –dc option (to remove constant sites). 7800 generations were allowed before stopping the chains, and the burnin period was determined after plotting the evolution of the log-likelihood (Lnl) across the iterations. Convergence between the chains was assessed by examining the difference in frequency between all bipartitions, but was not achieved (maxdiff = 1) although the position of *Ancoracysta* was congruent in between the two chains. Likelihood ratio tree topology tests (Kishino-Hasegawa, KH; Shimodaira-Hasegawa, SH; Expected Likelihood Weight, ELW; approximately unbiased, AU) were used to compare the CAT+GTR+Γ4 (logL=-2855811.158) and LG+Γ4+F+C60 (logL=-2855811.158) topologies under the LG+Γ4+F+C60 model in IQ-TREE v1.5.4 [39] (Figure 3C).

Contiguous sequences of the complete ribosomal DNA (rDNA) operon were identified in the *de novo* DNA read assembly of *Ancoracysta twista* (few highly polymorphic sites were found by reverse read mapping and all of them were in the outer intergenic spacer).The picozoan sp. MS584-11 was found on two closely adjacent contigs in the original DNA read assembly. The rDNA operon phylogeny in Figure S2 was inferred from a matrix containing 4061 sites and 108 representative eukaryotic rDNAs, which was generated by -localpair alignment in MAFFT v. 7.215 [37] and hypervariable site removal in Gblocks v. 0.91b (b1=70%, b2=75%, b3=12, b4=4, b5=h parameters). Maximum likelihood analysis was inferred in RAxML 8.2.4 [38] by using 20 random starts, the GTRGAMMA model and 300 nonparametric bootstrap replicates for branch supports.

To evaluate environmental distribution of *Ancoracysta twista*, its 18S rDNA was used as a query to search the environmental sequence clone library in GenBank. A single short sequence, HM227078, with a 99% identity was identified, which was derived from a coastal tropical upwelling region of Brazil [49].

Mitochondrial gene mining, alignments and phylogenies

The concatenated phylogeny of 36 mitochondrial genes (Figure S3) was prepared by including *Ancoracysta* mitochondrial and nuclear sequences in a published set of 42 processed alignments [50] and by removing those alignments (*atp2*, *cytc2*, *mtif2*, *mtrf1*, *sco1* and *suclg1*) where *Ancoracysta* was not absent. Alignments were trimmed by BMGE v1.1 (-b 4 and -g 0.4 setting) and merged in a single matrix, and Maximum Likelihood phylogeny was computed in IQ-TREE v1.5.4 (LG+Γ4+I+F model with 300 non-parametric bootstrap replicates.

The HCCS of *A. twista* (Figure 3C) was found on two overlapping transcriptomic contigs, whose continuity was verified by reverse mapping of genomic and transcriptomic reads. The N-terminus of the mature protein could not be assembled. It is highly unlikely that the HCCS sequence could be derived from bacteria in the culture or the bodonid prey since all bacteria and all euglenozoans lack it. There is also no evidence of contamination from a stramenopile in the sequence data sets - the only contaminants found in the sequence data are fungi, but fungal HCCSs are well represented in public databases and the *Ancoracysta* sequence is not closely related to them. The HCCS phylogeny (Figures 3C and S4A) was inferred from a final matrix of 165 amino acid sites and 101 sequences of

representative eukaryotes in IQ-TREE v1.5.4 (LG+Γ4+I +F model and 1000 ultrafast bootstrap replicates). The matrix was prepared by modifying a published dataset [23] by using GenBank and iMicrobe MMETSP as resources, aligning sequences MAFFT v. 7.215 (--localpair and --maxiterate 1000 settings) [37] and removing hypervariable sites in BMGE v1.1 (-b 3 and -g 0.4 settings).

The expression of the mitochondrion-encoded *ccm* genes was evaluated by searching their presence in the whole-cell transcriptome. Although the protocol for generating this transcriptome involved reverse transcription with the oligodT primer, which typically results in under-representation of mitochondrial transcripts in the data, the *ccmA* and *ccmF* transcripts were present (those for *ccmB* and *ccmC* were absent).

To test whether any of the 16 least abundant mitochondrion-encoded genes (Figure 4C) could originate by horizontal transfer, their individual phylogenies with a representative sampling of bacteria were inferred (LG+Γ4+I+F model in IQ-TREE v1.5.4 with 100 ultrafast bootstrap replicates). Few of the tree topologies were well resolved (many of the genes are short and divergent in sequence), but at least nine of them (*atp3, cox11, rpl1, rpl23, rpl27, rpl34, rpl35, rpl36* and *tufA*) were consistent with a vertical origin in the mitochondrial endosymbiont: they had nuclear-encoded homologs in other eukaryotes and their sequences were related to alphaproteobacteria; also see [5,20]. The remaining seven mitochondrion-encoded genes (*rpl18, rps16, rpoA, rpoB, rpoC, rpoD* and *secY*) were highly divergent in sequence to the point that assigning homologous sites in their alignments was difficult. In phylogenies, most of them did not group with alphaproteobacteria and formed long branches of an unresolved position among bacteria. However, none of the trees unambiguously suggests an origin by horizontal transfer into the mitochondrial genome; note also that *rps16* is part of a conserved proteobacterial operon (Figure 4B), and *rpl18* and *rps16* are nucleus-encoded in other eukaryotes [5]. Because horizontal transfer into mitochondrial genomes is extremely rare and an isolated undetected event would not challenge conclusions about parallel gene loss, we conclude that this process could not have measurably impacted analyses in Figure 4.

Mt DNA assembly and annotation

Two overlapping mitochondrial genome contigs were identified by tblastn searches in the *Ancoracysta* genomic DNA assembly and joined into a single circular-mapping scaffold. The structure of the genome was confirmed by reverse mapping of paired-end genomic and transcriptomic reads across the boundaries of the two original contigs and inverted repeats in Consed v23 at the final average coverage of 14.8 genomic and 455 transcriptomic reads per site (Figure 2A). Besides the inverted repeat, no identical repeats longer than 35 base pairs were identified. The picozoan sp. MS584-11 mitochondrial genome (Figure S4B) was identified on eight overlapping contigs, which were assembled into a single circular-mapping sequence. The genome structure was verified by reverse mapping of reads at the final average coverage of 235 reads per site. The longest identical repeat of 123 nucleotides was unambiguously mapped in the genome by reverse read mapping; other repeats were 85 nucleotides or shorter. The mitochondrial genome of *Colponema vietnamica* strain Colp-7a (Figure S4B) was assembled from four overlapping contigs identified in the *de novo* genomic DNA assembly (no identical repeats longer than 20 base pairs were identified). Mitochondrial genes were annotated in Artemis v16 and by using BLASTN, BLASTP, HMMER3 [41] and tRNAScan-SE searches online.

Identifying new mitochondrial protein-coding genes

HMMER v3.1b2 was used to create Hidden Markov Model (HMM) profiles from MAFFT (--linsi algorithm) alignments of all mitochondrial proteins of alphaproteobacterial origin. Six frame protein translations (>30 amino acids) of both newly-sequenced and selected published mitochondrial genomes were then searched for similarity against the HMM profiles. All mitochondrial genome sequences published before Aug 1, 2017 were included in the searches except for densely sampled groups (metazoans, viridiplantae and stramenopiles) and groups with highly-reduced mitochondiral genomes

(trypanosomatids and apicomplexans) where representative mitochondrial genome sequences were selected by following the literature. Candidates for newly identified genes at the e-value cutoff of 1e-4 were curated individually by examining their overlap with other genes in the genome, comparing their sequence lengths and conservation in alignments of corresponding genes and assessing their protein fold and sequence conservation by using the online prediction tools: HMMScan [41], HHPred [42] and PHYRE2 [43]. Gene candidates that gave positive hits in the above databases that were not known previously in their respective eukaryotic supergroups were sorted into high and low confidence categories and were listed in Table S1. An updated matrix of mitochondrial protein-coding genes of demonstrable alphaproteobacterial origin was then generated and used to draw their count in individual species (Figure 3B) and along the early evolution of eukaryotes (Figure 4A), and to infer the distribution of rare mitochondrial genes (Figure 4C). Syntenic *rps16* and *rpl19* (Figure 4B) were identified by comparing malawimonad and amoebozoan mitochondrial genomes and their operons in alpha-proteobacterial genomes were identified by using SyntTax, http://archaea.u-psud.fr/synttax/Default.aspx.

Time-calibrated analysis of mitochondrial genome reduction

To study reduction of mitochondrial genome in time, we used molecular mean date estimates for eukaryotic group splits from a published multiprotein phylogeny calibrated by the fossil record (time estimates were acquired from the authors of the original studies and are listed in Table S2) [26,27]. The ancestral number of mitochondrial protein-coding genes was assigned to 75 of the 84 dated tree nodes by using data from all descendant mitochondrial genomes available. Nine dated splits were omitted because mitochondrial genomes in one or both of the daughter lineages are not available or are believed to be absent in those lineages altogether (Table S2). Present day numbers of mitochondrion-encoded proteins (time=0) were used for species in the original phylogeny only, where applicable (75 data points; Table S2); we found that these were representative of their lineages and their selection had a negligible effect on projections of the early mitochondrial genome reduction. In one case of an apparent outlier, the mitochondrial genome of *Thraustochytrium aureum* was used as representative for early-branching stramenopiles in place of the reduced mitochondrial genome of *Blastocystis hominis*. In 17 other cases, mitochondrial genome from a related organism was used as a proxy for missing data in the original species (Table S2) as such that the species position in the original phylogeny would not change. In finding the best functional fit through the data points, we compared linear (f(x)=Gene_count ~ (a * Time_mya) + b and exponential decay functions (f(x)=Gene_count~ exp(a + b * Time_mya) + c) as two competing and biologically plausible scenarios. The two functions were fit in R by using the nonlinear (weighted) least-squares estimate function “nls()”, and compared statistically by the ANOVA F-test function “anova()”. The exponential decay function was strongly preferred over the linear when using data points for all organisms (F=23.751,p-value=2.83e-06) and for four lineages as shown in Figure 4E (i.e, all data points connecting the eukaryotic root with their modern representatives). These lineages were jakobids (excluding *Jakoba libera* as an outlier for the group; F=34.424, p-value=2.04e-03), stramenopiles (F=165.13, p-value=3.51e-10), close relatives of animals (the choanoflagellate *Monosiga* and filasterians *Capsaspora* and *Ministeria*; F=299.48, p-value=3.24e-08), and animals, both when starting at the eukaryotic root (F=179.15, p-value=4.75e-12) and in their common ancestor (F=630.66, p-value=7.05e-15). The three estimated exponential coefficients (i.e., vertical shift=a, curvature=b, and horizontal shift=c) were strongly justified by the data in most cases, as determined by t-test p-values: all organisms (a: 2.3e-01, b: 3.4e-04, c: 4.2e-33); jakobids (a: 3e-02, b: 1.9e-02, c: 7.9e-12); stramenopiles (a: 2.2e-05, b: 7.3e-07, c: 1.2e-21); choanoflagellate + filasterians (a: 1.1e-06, b: 3.4e-08, c: 2e-25); animals from the eukaryotic root (a: 8.9e-05, b: 6e-06, c: 5e-11); animals from the animal ancestor (a: 1.1e-06, b: 3.4e-08, c: 2e-25).

DATA AND SOFTWARE AVAILABILITY

The accession numbers for sequences reported in this paper are GenBank: MG202006-MG202010 and PRJNA413804.

FIGURE LEGEND – SUPPLEMENTAL MOVIE

Movie S1 (related to Figure 1). Characteristic, fast twirling movement of *Ancoracysta twista*.