Microbial predators form a new supergroup of eukaryotes

https://doi.org/10.1038/s41586-022-05511-5

Received: 20 June 2022

Accepted: 2 November 2022

Published online: 07 December 2022

Check for updates

Denis V. Tikhonenkov[1,2,13]✉, Kirill V. Mikhailov[3,4,13], Ryan M. R. Gawryluk[5,13], Artem O. Belyaev[1,6], Varsha Mathur[7,8], Sergey A. Karpov[9,10], Dmitry G. Zagumyonnyi[1,2], Anastasia S. Borodina[1,11], Kristina I. Prokina[1,12], Alexander P. Mylnikov[1,14], Vladimir V. Aleoshin[3,4] & Patrick J. Keeling[7]

Molecular phylogenetics of microbial eukaryotes has reshaped the tree of life by establishing broad taxonomic divisions, termed supergroups, that supersede the traditional kingdoms of animals, fungi and plants, and encompass a much greater breadth of eukaryotic diversity[1]. The vast majority of newly discovered species fall into a small number of known supergroups. Recently, however, a handful of species with no clear relationship to other supergroups have been described[2–4], raising questions about the nature and degree of undiscovered diversity, and exposing the limitations of strictly molecular-based exploration. Here we report ten previously undescribed strains of microbial predators isolated through culture that collectively form a diverse new supergroup of eukaryotes, termed Provora. The Provora supergroup is genetically, morphologically and behaviourally distinct from other eukaryotes, and comprises two divergent clades of predators—Nebulidia and Nibbleridia—that are superficially similar to each other, but differ fundamentally in ultrastructure, behaviour and gene content. These predators are globally distributed in marine and freshwater environments, but are numerically rare and have consequently been overlooked by molecular-diversity surveys. In the age of high-throughput analyses, investigation of eukaryotic diversity through culture remains indispensable for the discovery of rare but ecologically and evolutionarily important eukaryotes.

Before the advent of high-throughput sequencing methods, cultivation and microscopy were the main approaches for exploring the diversity of microbial organisms. Molecular surveys of microbial communities have bypassed the restrictive lack of cultivation methods for most of microbial life, and led to an explosive increase in the known diversity of bacteria and archaea[5,6]. The same molecular strategies also revealed new eukaryotic groups[7–9], but notably fewer than for prokaryotes. This is due in part to the fact that much of the eukaryotic diversity was already recognized through morphological studies, but also because even deep molecular survey data predominantly uncover relatively abundant taxa. Rare taxa are more easily overlooked, and eukaryotes include an entire ecological class of organisms that tend to be numerically rare—predators[10]. Recent years have witnessed a resurgence of cultivation as a method to discover new microbial predators. These rare but important organisms often appear as 'orphan' lineages in the tree of life, and have already substantially impacted our understanding of early eukaryotic evolution[2–4,11–13]. Beyond highlighting the blind spots of molecular survey data, the orphan lineages also raise an important biological question as to whether these organisms are phylogenetically isolated relicts, or the tip of an iceberg of more elusive diversity.

Like their animal counterparts, microbial predators are expected to be comparatively rare in nature. But rarity does not preclude either a high level of diversity or ecological importance any more than it does for animals that fill similar ecological niches. Continued discovery of new lineages will be important for resolving many issues in the eukaryotic tree of life, but it is also important that each newly discovered lineage is examined in some detail to better understand the structure of their diversity, how they have evolved, and the roles they might have played in evolution and still have in ecology.

Morphology of new microbial predators

Ten new microbial predators were isolated from geographically distinct marine habitats, including coral reefs of Curaçao, nearshore sediments of the Black and Red seas, and the water columns of the North-East Pacific and Arctic oceans. These strains are all small, fast-swimming

[1]Papanin Institute for Biology of Inland Waters, Russian Academy of Sciences, Borok, Russian Federation. [2]AquaBioSafe Laboratory, University of Tyumen, Tyumen, Russian Federation. [3]Belozersky Institute for Physico-Chemical Biology, Lomonosov Moscow State University, Moscow, Russian Federation. [4]Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russian Federation. [5]Department of Biology, University of Victoria, Victoria, British Columbia, Canada. [6]Department of Zoology and Ecology, Penza State University, Penza, Russian Federation. [7]Department of Botany, University of British Columbia, Vancouver, British Columbia, Canada. [8]Department of Zoology, University of Oxford, Oxford, UK. [9]Zoological Institute, Russian Academy of Sciences, Saint Petersburg, Russian Federation. [10]Department of Invertebrate Zoology, Faculty of Biology, Saint Petersburg State University, Saint Petersburg, Russian Federation. [11]Department of Zoology and Parasitology, Voronezh State University, Voronezh, Russian Federation. [12]Ecologie Systématique Evolution, CNRS, Université Paris-Saclay, AgroParisTech, Gif-sur-Yvette, France. [13]These authors contributed equally: Denis V. Tikhonenkov, Kirill V. Mikhailov, Ryan M. R. Gawryluk. [14]Deceased: Alexander P. Mylnikov. ✉e-mail: tikho-denis@yandex.ru

and superficially unremarkable flagellates that prey on other microbial eukaryotes. To obtain the isolates, water samples were enriched with *Pseudomonas fluorescens* bacteria to stimulate the growth of bacterivorous nanoflagellates, which in turn stimulated the growth of eukaryovorous protists. New strains were isolated by micropipette and propagated in a predator–prey culture on the bodonid *Procryptobia sorokini* as a steady food source.

The general morphological features of the new strains include a ventral feeding groove, a complex cell envelope, extrusive organelles and two heterodynamic flagella inserted into separate pockets. This same overall body plan describes the previously discovered orphan species *Ancoracysta twista*[3] and a strain formerly known as *Colponema marisrubri*[14], which here we rename *Nebulomonas marisrubri*. However, these similarities are only cursory and are shared with other distantly related protist groups; individually, these organisms are fundamentally different structurally and behaviourally, and probably occupy different niches in microbial communities. Notably, different strains of these predators exhibit different modes of feeding—one group feeds by nibbling on their prey, and the other group engulfs whole prey. We refer to these two groups as nibblerids and nebulids, respectively (see the Supplementary Discussion for taxonomic diagnoses).

The nebulids comprise the species *A. twista* and *N. marisrubri*. They are approximately 10-µm-long ovoid flagellates that phagocytose entire prey cells. Nibblerids, which include *Ubysseya fretuma* gen. nov., sp. nov. and four new species united under *Nibbleromonas* gen. nov., are much smaller (about 3 µm) (Fig. 1a–o) and have sickle-shaped starved cells with a distinct thorn under the ventral groove that contains five or six large complex extrusive organelles (Fig. 1p,s,t) that are used for attacking prey. Nibblerids can also engulf whole prey (Supplementary Video 1), but more characteristically feed by a unique behaviour whereby they bite off and ingest a part of a large prey cell by closing their ventral groove (Fig. 1s,u and Supplementary Video 2) and using tooth-like protrusions that nibble pieces of the larger prey (Fig. 1s). This feeding mode is unique, and demonstrates how pico-sized flagellates can feed on larger cells, which is often not considered in the modelling of microbial food webs.

Nibblerids are also ultrastructurally unique (see the Supplementary Discussion for a morphological description) and different from *Ancoracysta*[3]. Characteristic morphological features include 1–2 dorsal layers of alveolar vesicles beneath the plasma membrane (Fig. 1p,q), the internal membranes used as a depot for the formation of a food vacuole around the prey (note the absence of the internal membrane in Fig. 1r), micropores between the alveoli (Fig. 1p (inset)), a row of equidistant cytoplasmic microtubules supporting the cell coverings (Fig. 1q), a flagellar transition zone with an axosome, a curved transverse plate at the level of the cell surface and a transition cylinder distal to the transverse plate (Fig. 1v), wide bands of microtubules armouring the walls of the ventral groove (Fig. 1s,u), a posterior flagellum with two opposite longitudinal folds (Fig. 1r (inset)), a large mitochondrion with sac-like cristae and a filamentous inclusion (Fig. 1p,t,w) and a microbody next to the mitochondrion (Fig. 1w).

The two longitudinal folds seen in nibblerid flagella is a rare trait among eukaryotes that is otherwise found only in malawimonadids, some metamonads and discobids. The peculiar filamentous inclusion in the mitochondrion is characteristic of tubular cristae in some ochrophytes (Chrysophyta, Xanthophyta). The characteristics shared with distant relatives suggest that these aspects of their body plan may be very ancient, potentially reminiscent of the ancestral state of several large eukaryotic supergroups.

The new strains form an ancient lineage

We obtained transcriptomes for the new strains and investigated their position in the phylogeny of eukaryotes using a 320-gene dataset encompassing a broad spectrum of eukaryotic diversity[15]. Bayesian inference and maximum-likelihood tree reconstructions, performed using site-heterogeneous models (Methods), revealed a new supra-kingdom-level group of eukaryotes, here named Provora (devouring voracious protists) (Fig. 2). The nibblerids and nebulids form two deeply diverging lineages of Provora. The phylogenetic position of Provora relative to other established eukaryotic groups varies slightly depending on the phylogenetic method, with conflicting placements in the Bayesian inference and maximum-likelihood reconstructions. The Bayesian inference tree places Provora sister to a supergroup comprising TSAR (the SAR supergroup plus Telonemia) and Haptista with 0.95 posterior probability (Fig. 2). By contrast, the maximum-likelihood analysis strongly favours (98% bootstrap support) a union of Provora with another group of uncertain phylogenetic affinity, the Hemimastigophora, and places both as sister to TSAR and Haptista (Extended Data Fig. 1a).

To examine the possible impacts of mutational saturation and compositional bias on the phylogeny, we conducted analyses using site-elimination and alignment recoding approaches[16] (Methods). Elimination of the fastest-evolving sites or the most heterogeneous partitions produces phylogenies that are broadly congruent with the original maximum-likelihood and Bayesian inference trees. Removal of compositionally heterogeneous partitions preserves the original maximum-likelihood tree topology when up to 70% of the alignment is eliminated (Supplementary Table 1). With up to 40% of the fastest-evolving sites eliminated, the original maximum-likelihood tree topology remains unchanged, and support for the grouping of Provora with Hemimastigophora decreases only slightly (from 98% to 84% bootstrap support) (Supplementary Table 1). When 50% of the fastest-evolving sites are eliminated, the analysis switches to weakly supporting the sister position of Provora to TSAR + Haptista (65% bootstrap support), recovering the relationship obtained in the Bayesian inference tree (Fig. 2). Further removal of variable sites quickly destabilizes the entire tree, including the TSAR clade, which was shown to require a substantial alignment length to maintain stability[17], and the Provora itself, splitting the group into the individual Nebulidia and Nibbleridia clades.

Bayesian inference with the six-state recoded alignment yields a monophyletic Provora in position sister to Haptista with a low posterior probability (0.58 pp) (Extended Data Fig. 1c). The alternative, which receives 0.42 pp, places Provora sister to TSAR + Haptista—similar to the non-recoded dataset (Fig. 2). Applying four-state recoding to further decrease the effects of saturation and compositional biases appears to also dissolve much of the phylogenetic signal for deep tree nodes. With the four-state recoding, we obtained paraphyletic Provora and unresolved relationships between major lineages in Diaphoretickes (Extended Data Fig. 1d).

An approximately unbiased test with a range of possible phylogenetic relationships for Provora and Hemimastigophora did not reject 10 out of the 63 tested topologies at the 5% significance level when analysing the full dataset (Supplementary Table 2). The approximately unbiased test is most restrictive when 20% to 30% of the sites are eliminated. Specifically, after eliminating 20% of sites by the evolutionary rate, the approximately unbiased test rejects all but two of the topologies: those recovered by the Bayesian inference and maximum-likelihood analyses. The test highlights the sister relationship to the Haptista + TSAR assemblage as a unique non-conflicting solution for the placement of Provora—this tree topology is observed in the Bayesian inference analyses with the native and six-state recoded data (Fig. 2 and Extended Data Fig. 1b,c), and it is the only other alternative in the maximum-likelihood analyses that avoids rejection by the test in the site-elimination series. Overall, the phylogenomic analyses cannot currently distinguish between the alternatives, but do strongly support the monophyly of Provora and show that they are distinct and distantly related to other eukaryotes.

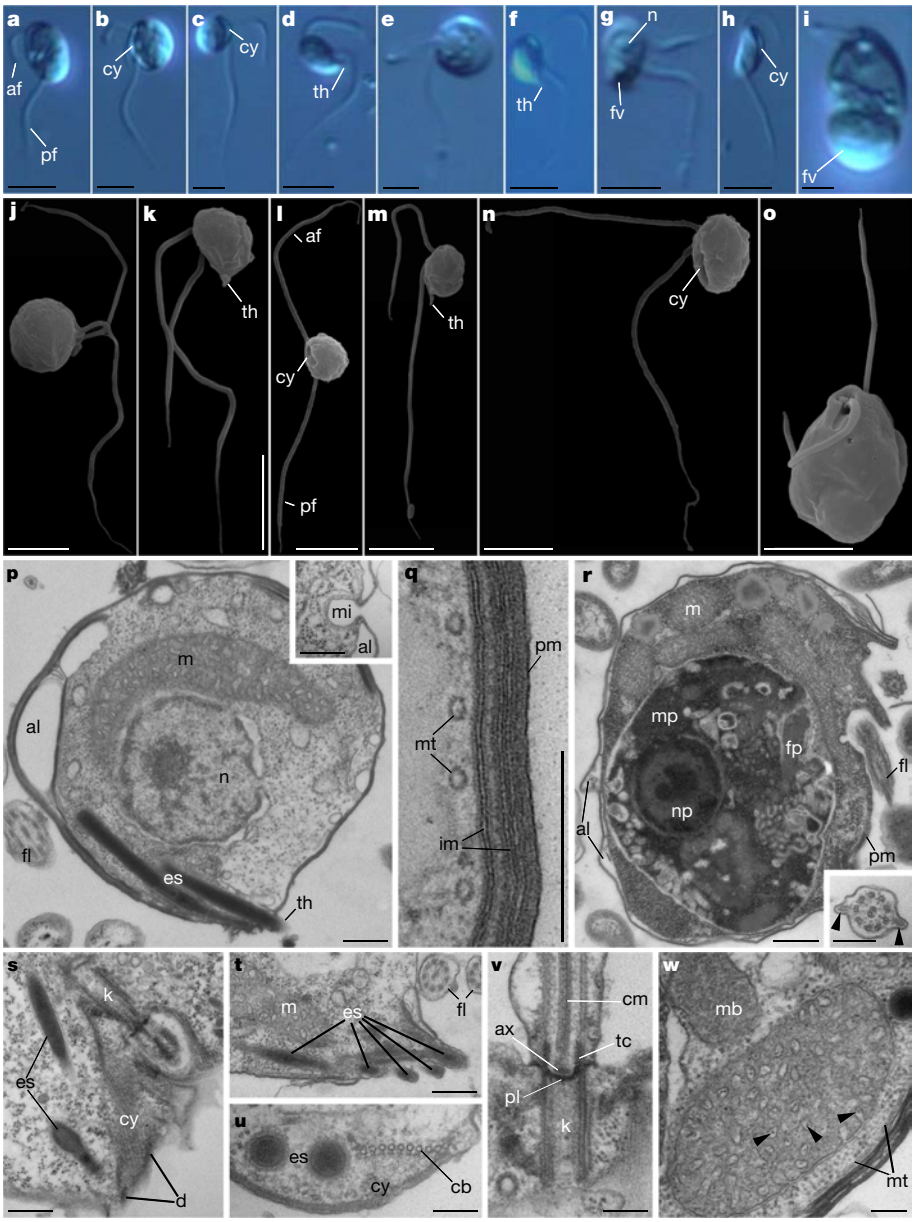


Fig. 1 | Cell morphology. a–i, Living cells, visualized by light microscopy, showing *U. fretuma* (**a,b**), *Nibbleromonas arcticus* (**c,d**), *Nibbleromonas kosolapovi* (**e**), *Nibbleromonas curacaus* (**f**), *Nibbleromonas quarantinus* (**g,h**), *N. marisrubri* (**i**). **j–o**, Cells, visualized by scanning electron microscopy, showing *U. fretuma* (**j**), *N. arcticus* (**k**), *N. kosolapovi* (**l,m**), *N. quarantinus* (**n**), *N. marisrubri* (**o**). **p–w**, Cell sections, visualized by transmission electron microscopy (exemplified by *N. quarantinus* (**p**,**q**,**s–w**) and *N. arcticus* (**r**)). **p**, Section through the middle part of the starving cell, showing the non-uniformity of the cell coverings and thorn; the inset shows a longitudinal section of a micropore with typical alveoli. **q**, Complex multimembrane coverings with underlying microtubules. **r**, Cell with engulfed prey; the inset shows a transverse section of the posterior flagellum with two longitudinal folds (arrowheads). **s**, Section through the base of the flagellum and cytostomal ventral groove with protruding 'denticles'. **t**, Longitudinal section of a thorn with five extrusomes. **u**, Cross-section of the cytostomal band of microtubules with the plate facing into the cytostomal ventral groove. **v**, Longitudinal section of kinetosome and transition zone of flagellum. **w**, Mitochondrion with sac-like cristae containing filamentous inclusions (arrowheads) and a microbody. af, anterior flagellum; al, alveoli; ax, axosome of flagellum; cb, cytostomal band of microtubules; cm, central microtubules of flagellum; cy, cytostomal ventral groove; d, denticles; es, extrusomes; fl, flagellum; fp, flagellum of prey; fv, food vacuole; im, inner membranes; k, kinetosome of flagellum; m, mitochondrion; mb, microbody; mi, micropore; mp, mitochondrion of prey; mt, microtubules; n, nucleus; np, nucleus of prey; pf, posterior flagellum; pl, transversal plate; pm, plasma membrane; tc, transitional cylinder; th, thorn. Scale bars, 3 µm (**a–o**), 400 nm (**p**, main image), 100 nm (**p**, inset), 400 nm (**r**, main image) 150 nm (**r**, inset) and 200 nm (**q** and **s–w**). These experiments were repeated 50 (**a–i**) and 3 (**j–w**) times with similar results.

Provora is distributed globally

To characterize the distribution of Nibbleridia and Nebulidia species in nature, we comprehensively searched 18S rRNA gene (SSU) surveys from diverse environments (Supplementary Data 1). We retrieved amplicons belonging to Provora globally and predominantly in marine environments with wide ecological variety, including coral reefs, open ocean surfaces, the deep chlorophyll maximum, mesopelagic waters and marine sediments (5,000 m), and also found evidence for their presence in brackish and fresh waters, but not in soil. Provora appear in relatively low abundance in all surveys (Extended Data Fig. 1e).

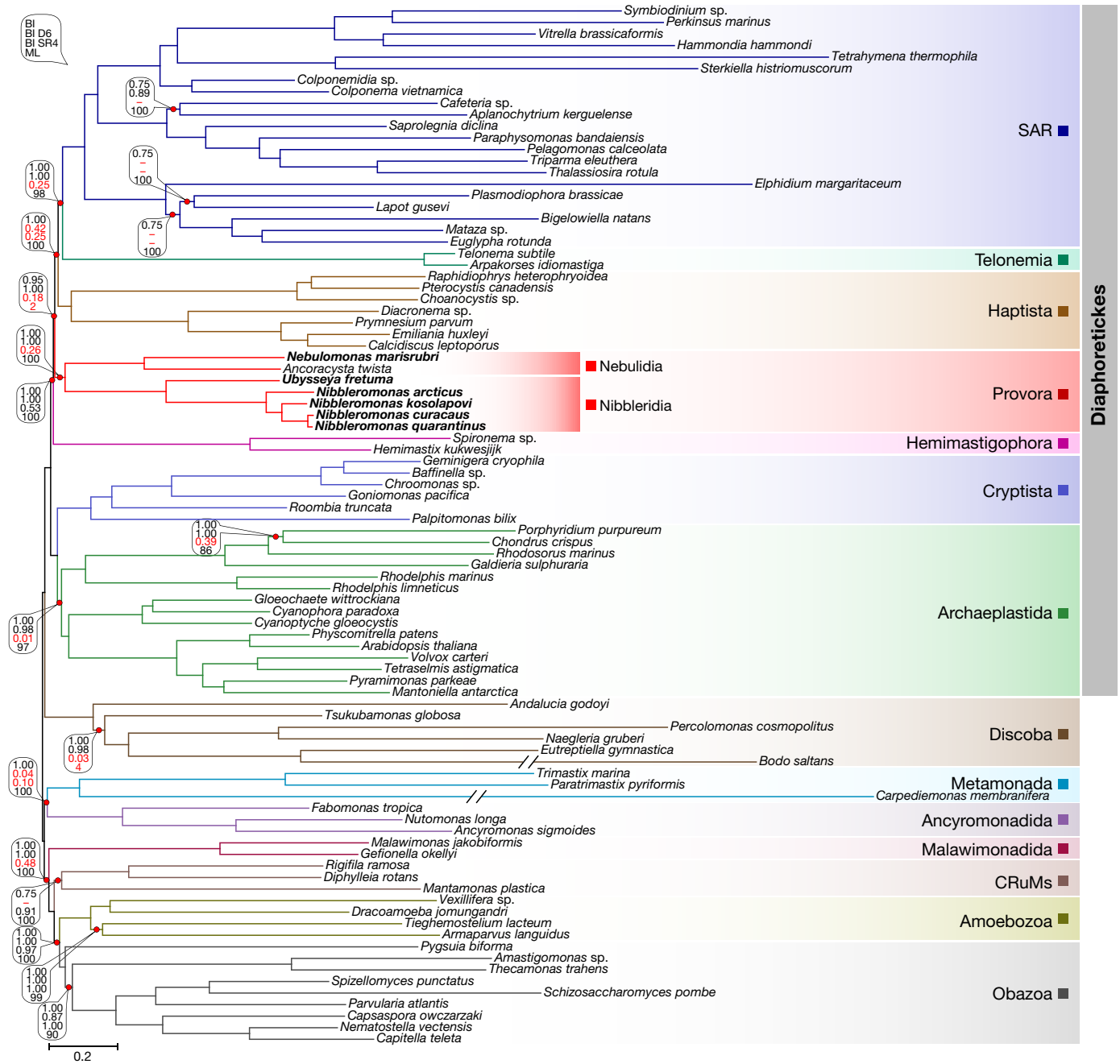


Fig. 2 | Phylogeny of eukaryotes reconstructed with a concatenated 320-gene dataset. A Bayesian inference consensus tree obtained using PhyloBayes with four independent analysis chains (CAT + GTR + G4 model), featuring support values obtained in the analyses with the recoded alignments, and the maximum-likelihood analysis (posterior mean site frequency (PMSF) model, bootstrap with 100 replicates). Tree nodes with incongruence between analyses or simply lacking maximal support values in at least one type of analysis are marked with red circles, and the corresponding support values are shown. Support values from top to bottom, the PhyloBayes posterior probability with the native dataset, the PhyloBayes posterior probability with the Dayhoff 6-recoded dataset, the PhyloBayes posterior probability with the SR4-recoded dataset and the maximum-likelihood bootstrap support percentage. Support values for bipartitions that were not recovered in the consensus tree for the corresponding analysis are given in red. The newly described species of Provora are given in bold. The branches of *Bodo saltans* and *Carpediemonas membranifera* were shortened by 30% for the illustration.

Note that, although high-throughput environmental sequencing did sample these organisms, the deep evolutionary divergence of Provora means that phylogenetic trees based on the SSU hypervariable regions used in such surveys cannot recover their phylogenetic relationship without support from a broader phylogenomic framework. As a result, such sequences are consistently misidentified, annotated as unclassified orphans, or even more often simply excluded from analyses or ignored owing to their low numbers. Comparing the SSU survey data with the ten strains now characterized by culturing and microscopy analysis suggests that the diversity of Provora at the genus level is even higher than represented among cultured representatives (Extended Data Fig. 2).

Characteristics of gene family content

Finding that Provora are distantly related to all other eukaryotes, we surveyed their gene content to establish some of their basic features, and to compare the two main subgroups to one another. At the highest

level, Provora appear to have gene-rich genomes and, despite their apparent low abundance, there is no evidence of accelerated evolution often associated with small population sizes[18]—no excessive gene loss was observed (Extended Data Fig. 3a), and phylogenomic data show that their genes are among the least divergent in eukaryotes, as reflected in their short branch lengths (Fig. 2).

Functional annotation and trophic mode analysis of the transcriptomic data in Provora is consistent with a predatory lifestyle. No characteristic proteins of plastid-bearing lineages, such as plastid import proteins, are detected in the transcriptomes of provorans. Microtubule-associated proteins, which are crucial for flagellar motility, are conserved in Provora (Extended Data Fig. 3), and they possess a rich suite of proteases and lysosomal nutrient-sensing complexes, including Ragulator–Rag, GATOR1, GATOR2 and KICSTOR, that are involved in the regulation of cell growth (Supplementary Data 2). A comparison of protein domains with other eukaryotes shows an abundance of proteins involved in calcium signalling in Provora (Supplementary Data 3), including an enriched repertoire of calcium-activated ion channels of the intermediate/small conductance potassium channel family, anoctamin family chloride channels and proteins with an interaction module for cellular calcium sensors (IQ calmodulin-binding motif)[19]. Phylogenetic analysis with eukaryotic members of the inositol trisphosphate receptors, which orchestrate the release of calcium ions from the endoplasmic reticulum stores[20], infers multiple deep lineages and independent expansions in Nibbleridia and Nebulidia (Extended Data Fig. 4), suggesting that these receptors and the calcium signalling system have an important role in the coordination of cellular behaviours in Provora.

Among the protein domains that are most prominently enriched in Provora relative to other eukaryotes, we found a family of membrane-attack complex and perforin domains (MACPF). Members of the MACPF family are known predominantly as pore-forming cytolytic proteins that function in the immune systems of animals and plants[21,22], or in host cell invasion by parasitic protists[23], and were also reported to constitute lethal toxins of the sea anemone extrusive organelles[24], which are analogous to the extrusomes of Provora. Protein domain searches identified 7 to 30 proteins with MACPF domains in the transcriptomic data of the Provora species. The family is equally abundant in Nibbleridia and Nebulidia and shows multiple lineage-specific expansions (Extended Data Fig. 5). MACPF domains in Provora are found in association with EGF-like domains, and many sequences are predicted with a secretory signal peptide, supporting probable extracellular targeting of these proteins.

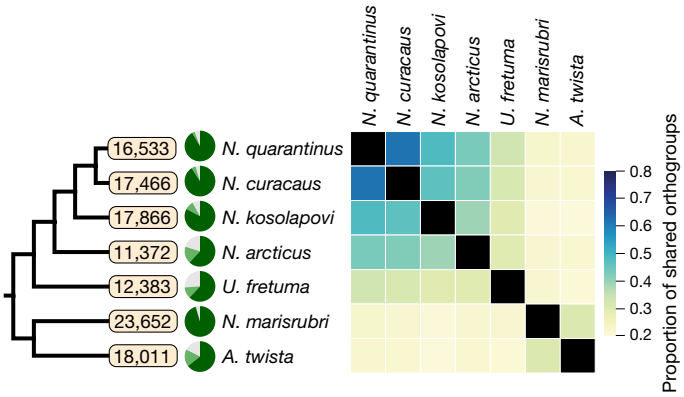


Fig. 3 | Estimated gene family diversity in Provora. Gene family counts were estimated using OrthoFinder orthogroups. The total counts of inferred orthogroups (including species-specific orthogroups) are provided for each species. The proportions of shared orthogroups to the total orthogroup counts in pairwise comparisons of species (arithmetic mean) are shown by a heat map. BUSCO completeness estimates (eukaryota_odb9) for the transcriptomes of Provora are represented by pie charts: dark green, complete; light green, fragmented; grey, missing.

The antiquity of the split between the two deep lineages comprising Provora is also reflected in their gene family contents. Nibblerid and nebulid species share only 20–25% of inferred orthologous groups, similar to the proportions shared with distantly related eukaryotic species (Fig. 3 and Extended Data Fig. 6). The orthologous groups also indicate that their genomes are relatively gene rich, providing an estimate of 16–24 thousand families in total for the three representatives of Provora with the highest completeness estimates.

Mitochondrial genomes of Provora

The mitochondrial genome of *A. twista* was previously shown to be unusually gene rich, and this feature was found to be conserved across the whole Provora lineage (Extended Data Figs. 7 and 8 and Supplementary Table 3).

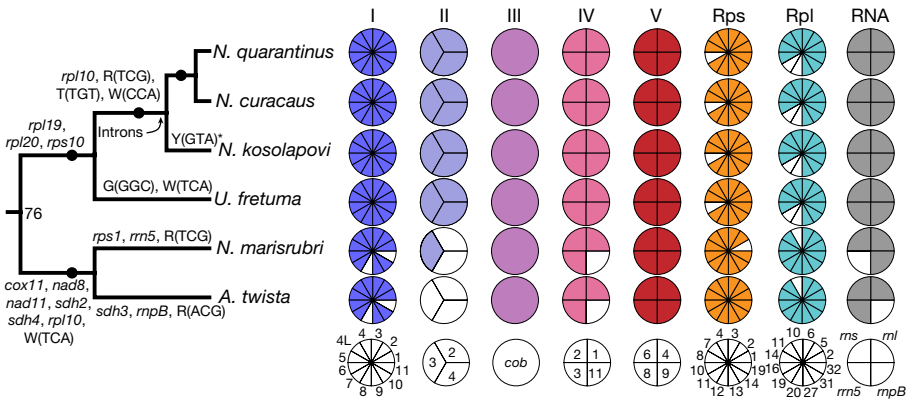


Fig. 4 | Mitochondrial genomes support the distinctness and diversity of Provora. A subsection of a global mitochondrial multiprotein phylogeny focused on Provora is presented, with a Coulson plot showing variation in nibblerid and nebulid mitochondrial genome repertoires. Each functional complex is shown as a pie chart with individual mitochondrial genes as wedges. Empty wedges indicate the absence of a gene; the genes are identified in the legend below. The evolutionary dynamics of mitochondrial genome content is summarized with a tree, listing the gene losses next to the corresponding branches; I–V represent the respiratory chain complexes NADH dehydrogenase (I), succinate dehydrogenase (II), cytochrome *c* reductase (III), cytochrome *c* oxidase (IV) and ATP synthase (V). *rps*, small-subunit ribosomal proteins; *rpl*, large-subunit ribosomal proteins. 'RNA' indicates RNA-encoding genes: *rns*, small-subunit ribosomal RNA; *rnl*, large-subunit ribosomal RNA; *rrn5*, 5S ribosomal RNA; *rnpB*, RNA component of RNase P. *cob* corresponds to apocytochrome *b*. Mitochondrial tRNA genes are specified according to the single-letter amino acid code, with anticodon sequences in parentheses. Ultrafast bootstrap scores are included as a measure of statistical support, and broadly support the conclusions of Fig. 1. The solid black dots indicate full support.

Their mitochondrial genomes share a conserved set of 51 proteins, with only minor variations, such as patchy presence of a few ribosomal proteins, tRNAs and bacteria-like *rnpB* (Fig. 4 and Extended Data Fig. 9a). In many cases, the missing genes are found in the transcriptomes as putatively nucleus-encoded homologues, suggesting that the variability is the result of functional endosymbiotic gene transfers. Most of the differences in the genome size are due to species-specific variations in the number and size of mitochondrial group I introns and the associated homing endonuclease genes, which apparently arose within the genus *Nibbleromonas*, potentially aided by lateral transfer from fungal mitochondria (Extended Data Fig. 9b).

Two noteworthy functional variations that distinguish Nibbleridia and Nebulidia affect electron-transport-chain complexes and their assembly factors (Fig. 4). All mitochondrial genomes in Provora encode a type I cytochrome *c* maturation system (*ccmA*, *ccmB*, *ccmC* and *ccmF*), inherited from the ancestor of mitochondria, and Nebulidia also possess a nucleus-encoded type III cytochrome *c* maturation system (holocytochrome *c* synthase; HCCS), as reported previously in *A. twista* (Extended Data Fig. 10), which has replaced the type I system in most eukaryotes. The presence of dual cytochrome *c* maturation systems in *N. marisrubri* and *A. twista* suggests that both systems have co-existed over extended evolutionary time, arguing against the proposed ongoing replacement of type I system[3], and suggests that comparisons of nibblerid and nebulid mitochondria may provide unique insights into the evolution of cytochrome *c* biogenesis in eukaryotes. Together, both transcriptomic data and mitochondrial genomes of Provora emphasize the deep evolutionary distance between its lineages with, for example, mitochondrial diversity exceeding all known diversity of metazoan mitochondria.

Conclusions

Provora is an ancient supergroup of eukaryotes that rivals traditional Kingdoms of animals, fungi or plants in terms of antiquity and the level of divergence between its few described members. It incorporates the orphan species *A. twista*, revealing it to be the first clue of a diverse major lineage that has gone undetected through thousands of environmental molecular surveys, rather than a remote relict. Despite their diversity and global distribution, Provora are numerically rare, but as eukaryovorous predators, their rarity relative to other microbes is not surprising and does not indicate a lack of ecological impact any more than a lion's rarity compared to wildebeest does. These findings underscore how high-throughput sequencing methods are valuable, but alone are insufficient for understanding the diversity and phylogeny of eukaryotes: all methods have different biases, and culturing continues to be a crucial tool for discovering rare and genetically divergent lineages of ecological importance, and deducing their biology and relationship to other established groups.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-022-05511-5.

1. Keeling, P. J. & Burki, F. Progress towards the tree of eukaryotes. *Curr. Biol.* **29**, R808–R817 (2019).
2. Gawryluk, R. M. R. et al. Non-photosynthetic predators are sister to red algae. *Nature* **572**, 240–243 (2019).
3. Janouškovec, J. et al. A new lineage of eukaryotes illuminates early mitochondrial genome reduction. *Curr. Biol.* **27**, 3717–3724 (2017).
4. Lax, G. et al. Hemimastigophora is a novel supra-kingdom-level lineage of eukaryotes. *Nature* **564**, 410–414 (2018).
5. Oren, A. Prokaryote diversity and taxonomy: current status and future challenges. *Philos. Trans. R. Soc. Lond. B* **359**, 623–638 (2004).
6. Shu, W. S. & Huang, L. N. Microbial diversity in extreme environments. *Nat. Rev. Microbiol.* **20**, 219–235 (2022).
7. Massana, R., del Campo, J., Sieracki, M. E., Audic, S. & Logares, R. Exploring the uncultured microeukaryote majority in the oceans: reevaluation of ribogroups within stramenopiles. *ISME J.* **8**, 854–866 (2014).
8. de Vargas, C. et al. Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**, 1261605 (2015).
9. Flegontova, O. et al. Extreme diversity of diplonemid eukaryotes in the ocean. *Curr. Biol.* **26**, 3060–3065 (2016).
10. Ahlering, M. A. & Carrel, J. E. Predators are rare even when they are small. *Oikos* **95**, 471–475 (2001).
11. Hehenberger, E. et al. Novel predators reshape holozoan phylogeny and reveal the presence of a two-component signaling system in the ancestor of animals. *Curr. Biol.* **27**, 2043–2050 (2017).
12. Tikhonenkov, D. V. et al. 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 (2014).
13. Tikhonenkov, D. V. et al. New lineage of microbial predators adds complexity to reconstructing the evolutionary origin of animals. *Curr. Biol.* **30**, 4500–4509 (2020).
14. Mylnikov, A. P. & Tikhonenkov, D. V. The new alveolate carnivorous flagellate *Colponema marisrubri* sp. n. (Colponemida, Alveolata) from the Red Sea. *Zool. Zh.* **88**, 1163–1169 (2009).
15. Strassert, J. F. H., Irisarri, I., Williams, T. A. & Burki, F. A molecular timescale for eukaryote evolution with implications for the origin of red algal-derived plastids. *Nat. Commun.* **12**, 1879 (2021).
16. Rodriguez-Ezpeleta, N. et al. Detecting and overcoming systematic errors in genome-scale phylogenies. *Syst. Biol.* **56**, 389–399 (2007).
17. Strassert, J. F. H., Jamy, M., Mylnikov, A. P., Tikhonenkov, D. V. & Burki, F. New phylogenomic analysis of the enigmatic phylum Telonemia further resolves the eukaryote tree of life. *Mol. Biol. Evol.* **36**, 757–765 (2019).
18. Lanfear, R., Kokko, H. & Eyre-Walker, A. Population size and the rate of evolution. *Trends Ecol. Evol.* **29**, 33–41 (2014).
19. Bahler, M. & Rhoads, A. Calmodulin signaling via the IQ motif. *FEBS Lett.* **513**, 107–113 (2002).
20. Schaffer, D. E., Iyer, L. M., Burroughs, A. M. & Aravind, L. Functional innovation in the evolution of the calcium-dependent system of the eukaryotic endoplasmic reticulum. *Front. Genet.* **11**, 34 (2020).
21. Morita-Yamamuro, C. et al. The *Arabidopsis* gene *CAD1* controls programmed cell death in the plant immune system and encodes a protein containing a MACPF domain. *Plant Cell Physiol.* **46**, 902–912 (2005).
22. Rosado, C. J. et al. The MACPF/CDC family of pore-forming toxins. *Cell. Microbiol.* **10**, 1765–1774 (2008).
23. Ishino, T., Chinzei, Y. & Yuda, M. A *Plasmodium* sporozoite protein with a membrane attack complex domain is required for breaching the liver sinusoidal cell layer prior to hepatocyte infection. *Cell. Microbiol.* **7**, 199–208 (2005).
24. Satoh, H., Oshiro, N., Iwanaga, S., Namikoshi, M. & Nagai, H. Characterization of PsTX-60B, a new membrane-attack complex/perforin (MACPF) family toxin, from the venomous sea anemone *Phyllodiscus semoni*. *Toxicon* **49**, 1208–1210 (2007).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature Limited 2022

Methods

Cell isolation and culture establishment

U. fretuma (clone TD-3) was obtained from a sea water sample taken in the Strait of Georgia, British Columbia, Canada (49° 10′ 366′′ N, 123° 28′ 50′′ W) at 220 m depth, salinity 35‰, using a Niskin bottle on 13 June 2017. *N. kosolapovi* clone Colp-32 was isolated from Arctic waters of the Kara Sea (75° 53′ 16.8′′ N, 89° 30′ 28.8′′ E), at 20 m depth (total depth 52 m), water temperature 0.66 °C, salinity 32.8‰ on 19 September 2015. *N. arcticus* clone Colp-45 was obtained from Arctic waters of the East Siberian Sea (71° 27′ 59.8′′ N, 152° 53′ 59.3′′ E), at 11 m depth, water temperature 2.76 °C, salinity 25.1‰ on 5 September 2017. *N. quarantinus* clones Colp-41 and Colp-44 were isolated from the sample of silty sand (salinity 18‰) taken in the shoreland of Quarantine Bay (44° 36′ 41.4′′ N, 33° 30′ 6.2′′ E) in Sevastopol city, Crimea, Black Sea on 13 May 2017. *N. curacaus* clones Cur-5 and Cur-12 were obtained from the sea waters (salinity 35‰) of the eastern point of the Curaçao island (12° 12′ 32.3′′ N, 68° 48′ 58.8′′ W) on 24 April 2018, scraping from the sponges *Agelas conifera* Schmidt 1870 and *Callyspongia vaginalis* Lamarck 1814, respectively, at 24.7 m depth. *N. marisrubri* clones Colp-4b, Colp-4c and Cur-8 were isolated from the Red Sea, Sharm El Sheikh, Egypt (27° 50′ 50.5′′ N, 34° 18′ 59.4′′ E), scraping from coral at 75 m depth, April 2015 (Colp-4b); from the scraping from stone (salinity 18‰) in Kazachya Bay (44° 34′ 18.8′′ N 33° 24′ 40.2′′ E) in Sevastopol city, Crimea, Black Sea, on 1 September 2018 (Colp-4c); and from the coral sand at 24.7 m depth at the eastern point of the Curaçao island (12° 12′ 32.3′′ N, 68° 48′ 58.8′′ W), on 24 April 2018 (Cur-8).

The water samples were enriched for *P. fluorescens* bacterium Migula, 1895 at the rate of 0.15 ml of suspension (around 25 million bacteria cells) per 5 ml of sample. The samples were examined on the third, sixth and ninth day of incubation in accordance with methods described previously[25]. After isolation using a glass micropipette, clones were propagated on the bodonid *P. sorokini* strain B-69, which were grown in marine Schmalz–Pratt medium or artificial marine medium (RS-R11040, Red Sea) using the bacterium *P. fluorescens* as food[12]. No microbial eukaryotes other than *P. sorokini* were used in enrichment. Feeding of the provorans on heterotrophic *Spumella*-like heterotrophic chrysophytes and *Pteridomonas* spp. (Pedinellales) was also observed in natural samples. Isolated clones TD3, Colp-32, Colp-41, Colp-44, Colp-45 and Colp-4c are currently being stored in a collection of live protozoan cultures at the Papanin Institute for Biology of Inland Waters, Russian Academy of Sciences and the University of British Columbia; however, clones Cur-5, Cur-12, Cur-8 and Colp-4b perished after several months of cultivation.

Light and electron microscopy

Light microscopy observations were performed using the Zeiss AxioScope A.1 equipped with a DIC water-immersion objective (×63) and an AVT HORN MC-1009/S analogue video camera. For scanning electron microscopy, cells were collected by centrifugation (5,500*g*). Then, 0.5 ml of 2.5% glutaraldehyde (in 0.1 M cacodylate buffer) was added to the 0.5 ml of resuspended cells and kept at 4 °C for 30 min and then processed as described previously[26]. Transmission electron microscopy preparations were performed in accordance with a previously published protocol[26].

Preparation of libraries and sequencing

Cells grown in clonal laboratory cultures were collected when the cultures had reached peak abundance and after the prey had been eaten (light microscopy observations). Cells were collected by centrifugation (1,000*g* at room temperature) onto an 0.8 µm membrane of a Vivaclear mini column (Sartorius Stedim Biotech, VK01P042); this was done separately for RNA and DNA extractions. Total RNA was then extracted using the RNAqueous-Micro Kit (Invitrogen, AM1931) and converted into cDNA using the Smart-seq2 protocol[27]. Moreover, cDNA of clones TD-3, Colp-32, Colp-41, Cur-5, Cur-12 and Colp-4c was obtained from 20 single cells using the Smart-seq2 protocol (cells were manually picked from the culture using a glass micropipette and transferred to a 0.2 ml thin-walled PCR tube containing 2 µl of cell lysis buffer (0.2% Triton X-100 and RNase inhibitor (Invitrogen))). Paired-end libraries were prepared using the NexteraXT protocol (Illumina, FC-131-1024), and sequencing was performed on the Illumina MiSeq platform with read lengths of 2 × 300 bp.

Total DNA was extracted from the filters using the MasterPure Complete DNA and RNA Purification Kit (Epicentre, MC85200). Genomic DNA libraries of clones TD-3, Colp-41, Cur-12 and Colp-4c were generated at The Centre for Applied Genomics, and 150 bp paired-end reads were sequenced on the Illumina HiSeq X machine. Genomic DNA sequencing of clone Colp-32 was performed on the Illumina MiSeq platform with read lengths of 300 bp using the Nextera DNA Sample Prep Kit (Illumina, FC-121-1030) to construct paired-end libraries.

The SSU rRNA genes were amplified by PCR using the general eukaryotic primers GGF (5′-CTTCGGTCATAGATTAAGCCATGC-3′) and GGR (5′-CCTTGTTACGACTTCTCCTTCCTC-3′) for clone TD-3; PF1 and FAD4 (ref. [28]) for clone Colp-4b; EukA and EukB[29] for clones Colp-32, Cur-8, Cur-12 and Colp-4c; and 18SFU and 18SRU[30] for clones Colp-41, Colp-44, Colp-45 and Cur-5. The PCR products were subsequently cloned (Colp-4b, Colp-32, Cur-5, Cur-8, Cur-12 and Colp-4c) or sequenced directly (TD-3, Colp-41, Colp-44 and Colp-45) using Sanger dideoxy sequencing with two additional internal primers 18SintF (5′-GGTAATTCCAGCTCCAATAGCGTA-3′) and 18SintR (5′-GTTTCAGCCTTGCGACCATACT-3′).

Transcriptomic dataset assembly and decontamination

Raw Illumina sequencing reads were merged using PEAR v.0.9.6 and the quality of the paired reads was confirmed in FastQC[31,32]. Adapter and primer sequences were subsequently trimmed using Trimmomatic v.0.36 and transcriptomes were assembled using Trinity (v.2.4.0)[33,34]. The resulting contigs were then filtered for bacterial and kinetoplastid prey contaminants using BlobTools as well as BLASTn and BLASTx searches against the NCBI nt database and the Swiss-Prot database, respectively[35,36]. ORF predictions were carried out using TransDecoder (v.5.5.0)[37]. Predicted peptides in the transcriptomic assemblies of Provora isolates were clustered by CD-HIT[38] with a 90% identity threshold to reduce the redundancy of sequence sets. Before annotating the peptides, we also screened the data for contamination using similarity searches, and discarded sequences of probable bacterial or prey origin. The searches were performed using DIAMOND[39] against the NCBI's non-redundant database using the 'more-sensitive' search mode. The taxonomic data were extracted from the search results using TaxonKit[40]. Transcripts with the best hit to bacterial or euglenozoan (prey) sequences were removed from the assemblies. An additional screening was performed for the *Paraphysomonas*-contaminated transcriptome of *N. curacaus* Cur-5, by querying the transcriptome against *Paraphysomonas imperforata* and *Paraphysomonas bandaiensis*, available in the EukProt database[41]. The clustered and filtered peptide sets for each isolate were evaluated with BUSCO[42] using the eukaryota_odb9 dataset.

Annotation of transcriptomic data

The transcriptomes of Provora isolates were investigated using the KEGG database pathway maps and functional classification system[43]. The KEGG orthology assignments for the cleaned peptide sets were generated by the KEGG Automatic Annotation Server[44] using the bidirectional best-hit method. For comparative analyses of KEGG annotations, we selected 65 eukaryotic species with available genomic data, and similarly conducted assignments of KEGG orthology for each genome using the server. The results of orthology assignments for each organism were collected into a table, incorporating the KEGG BRITE classification system for orthologues (Supplementary Data 2). The KEGG orthology entries were evaluated using the counts of identified

orthologs in each species to highlight entries systematically over- or underrepresented in Provora against a sample of other eukaryotes. We used a simple normalized measure for each KEGG orthology entry, counting the number of species that had less orthologues than the isolates of Provora and subtracting the number of species that had more. The values were calculated for each isolate and an average value was reported for each KEGG orthology entry.

Conservation in the major functional categories defined by the KEGG BRITE classification system was summarized by means of a heat map featuring KEGG orthology entry counts in Provora and other eukaryotic species. The KEGG orthology entries in each species were reduced to the presence/absence data, and entries that appeared only in Diaphoretickes, Discoba or Amorphea were excluded to reconstruct the ancestral eukaryotic complement in accordance with the Dollo parsimony principle and the probable positions for the eukaryotic root[45]. The KEGG orthology counts in the functional categories for each species were normalized to the inferred ancestral eukaryotic entry count. The heat map was created using the Python data visualization library Seaborn[46].

Protein domain families in the cleaned peptide sets were identified using HMMER searches[47] with the PfamScan tool and the Pfam v.32.0 database[48]. The searches were carried out using the default family-specific gathering thresholds. Pfam domain searches were also performed for the collection of proteomes in the EukProt database[41]. The counts of proteins containing each domain family were extracted from the individual search results and assembled in a comparative table (Supplementary Data 3). To highlight the domain families that are enriched in Provora relative to the rest of eukaryotes in the EukProt database, we applied the same measure that was used for evaluating over- or under-representation of the KEGG orthologies. Protein domain architectures for selected groups of proteins were analysed using the SMART domain annotation resource[49], and signal peptides were predicted using SignalP (v.5.0)[50]. Profile searches for selected proteins, such as LAMTOR subunits of the Ragulator complex, were performed with HMMER using the alignments of known family members, constructed with MAFFT[51]. Trophic mode prediction and principal component analysis were performed with the Trophic Mode Prediction Tool[52] using the default settings with the built-in datasets.

Orthogroup analysis

For the identification of orthologous groups of proteins, we combined the transcriptomic data of isolates that originated from the same species: Cur-5 and Cur-12 for *N. curacaus*; Colp-41 and Colp-44 for *N. quarantinus*; Colp-4b, Colp-4c and Cur-8 for *N. marisrubri*. The combined transcriptomic datasets were clustered using CD-HIT[38] with a 90% identity threshold. The duplication values in the clustered datasets were estimated by BUSCO[42] to be between 2.3% and 5.6% with the eukaryota_odb9 dataset. Orthogroup inference was performed using OrthoFinder[53] for the transcriptomic datasets of Provora species and the proteomes of 65 eukaryotic species, selected to broadly sample the eukaryotic diversity and accounting for genome availability. The searches in the OrthoFinder workflow were performed using the BLAST algorithm[54]. The data on the shared orthogroups were extracted from the OrthoFinder output, and the proportions of shared orthogroups in pairwise comparisons were calculated using arithmetic mean. The heat map with the proportions of shared orthogroups was created using the Python data visualization library Seaborn[46].

Phylogenomic dataset construction

For the construction of the phylogenomic dataset we relied on a publicly available collection of 320 orthologous gene groups that cover a broad range of eukaryotes[15]. We limited the existing taxonomic sampling to 69 species for computational tractability, largely following the selection strategy outlined in that study and consulting the provided phylogeny with 733 taxa. The sampling was then extended using the transcriptomic data from the newly described species and also including several important lineages that were available in the EukProt database[41] but were missing in the original collection, such as hemimastigophores, CRuMs, ancyromonadids, colponemids and several other deep-branching members of eukaryotic groups (Supplementary Data 4). Orthologous sequences were identified in the transcriptomes and filtered to remove contaminants using a previously developed dataset-expansion pipeline[13]. We used sequences from the following organisms for eukaryotic contamination filtering: kinetoplastids *B. saltans* and *Trypanosoma cruzi* for *N. marisrubri* Colp-4b, colponemids and *Rhodelphis limneticus*; *P. imperforata* and *P. bandaiensis* for *N. curacaus* Cur-5; parasitic fungus *Malassezia globosa* for colponemids and hemimastigophores; additional fungal species (*Saccharomyces cerevisiae*, *Yarrowia lipolytica* and *Ustilago maydis*) for *Hemimastix kukwesjijk*; *Spodoptera litura* and *Amastigomonas* sp. for *Colponema vietnamica*; and *Trimastix marina* for *Ancyromonas sigmoides* and *Gefionella okellyi*. Orthologous sequences surviving the contamination filter were added to the 320-gene dataset and aligned with MAFFT[51] using the localpair (L-INS-i) algorithm. Single-gene alignments were inspected manually using BioEdit[55], and single-gene phylogenies were reconstructed using IQ-TREE[56] to resolve cases of questionable orthology or contamination where necessary. Specifically, sequences from new isolates and the EukProt database were screened for cross-contamination or residual contaminants surviving the filtering procedure. Cleaned sequence sets from the inspected alignments were then submitted to an automated quality-filtering procedure of PREQUAL[57] with a 0.95 posterior probability filtering threshold, realigned with MAFFT using the localpair (L-INS-i) algorithm, and trimmed with trimAl (ref. [58]) using an automated trimming heuristic followed by a gap threshold filter of 0.7. The resulting 320 trimmed alignments were concatenated by SCaFoS[59] into a data matrix with 104,691 sites (92,911 variable sites) and 94 operational taxonomic units. Each new isolate was present in at least 80% of all genes in the dataset. The recoded versions of the dataset were created with the recode option of PhyloBayes[60] by applying the Dayhoff scheme with six amino acid groups[61] or the SR4 recoding scheme[62] with four groups.

Phylogenomic analyses

Phylogeny reconstructions with the concatenated alignment were performed with the Bayesian inference approach implemented in PhyloBayes[60] and the maximum-likelihood approach of IQ-TREE[56]. PhyloBayes analyses were conducted under the site-heterogeneous CAT-GTR model[63] with four discrete Gamma rate categories; the -dc flag was applied for the input alignment to eliminate constant sites. Four independent chains were run with PhyloBayes for 10,000 cycles and summarized with a 50% burn-in and 0.02 sampling frequency to generate the consensus tree. The recoded alignments were analysed with PhyloBayes using identical parameters, but the computation was extended to 30,000 cycles. Maximum-likelihood tree reconstruction with IQ-TREE was performed using the LG + C60 + F + G4 profile mixture model[64]. Node support for the maximum-likelihood tree was evaluated with nonparametric bootstrapping with 100 replicates and using the PMSF method for the approximation of the profile mixture model[65].

For the site-elimination analyses, we generated a series of alignments by progressively removing the most variable sites or the most compositionally heterogeneous alignment partitions. Approximately 10% of the original alignment was removed in each iteration of the dataset. Site rates in the full alignment were estimated using IQ-TREE concurrent with the tree reconstruction and under the same evolutionary model. Compositional heterogeneity was evaluated using the relative composition frequency variability measure by BaCoCa[66]. Each alignment in the series was analysed by IQ-TREE similarly to the full alignment: tree reconstruction was performed using the LG + C60 + F + G4 model and node support was evaluated using nonparametric bootstrapping with 100 replicates and the PMSF method. Approximately unbiased

tree topology tests[67] were performed with the full alignment and the alignments in the site-elimination series. The approximately unbiased tests were performed in IQ-TREE using the site-wise likelihood calculated under the LG + C60 + F + G4 model for all datasets. Visualization of phylogenetic trees and construction of topologies was performed using MEGA[68].

Mitochondrial genome assembly and annotation

Paired-end 150 bp Illumina genomic DNA reads were trimmed of adapter and low-quality sequences using BBMap (v.37.36) (https://sourceforge.net/projects/bbmap/). Trimmed reads were assembled into contigs with SPAdes (v.3.14.1)[69] using *k*-mer sizes of 21, 33, 55, 77 and 99. Contigs corresponding to putative mitochondrial DNA were identified by querying assemblies with mitochondrial proteins, using tBLASTn. In the case of *N. marisrubri*, a single mitochondrial DNA contig could not be recovered with SPAdes; here, NOVOPlasty (v.4.3)[70] was used with a *k*-mer value of 55 to recover a single circular contig.

Mitochondrial DNA contigs were annotated automatically with MFannot (https://megasun.bch.umontreal.ca/cgi-bin/mfannot/mfannotInterface.pl), using translation table 4 (mold, protozoan and coelenterate mitochondrial). Mitochondrial large subunit ribosomal RNA (*rnl*) genes could not be annotated by MFannot in *N. quarantinus*, *N. curacaus* and *N. kosolapovi* owing to the presence of multiple group I introns, so exon/intron boundaries were assigned manually on the basis of alignment to the intronless *U. fretuma rnl* gene. Manual editing of exon/intron boundaries was performed using the NCBI Genome Workbench (v.3.6.0)[71]. Mitochondrial genome maps were generated with OGDRAW (v.1.3.1)[72].

Predicted secondary structures of mitochondrial *rnpB* genes from *U. fretuma*, *N. quarantinus* and *N. curacaus* were drawn with RNA2Drawer[73] on the basis of the predicted structures of jakobid *rnpB* homologues[74].

Individual mitochondrial protein phylogenies

Alignment of mitochondrion- and nucleus-encoded mitochondrial proteins was performed using MAFFT L-INS-i (v.7.313)[51]. Non-homologous sequences were trimmed with BMGE (v.1.1.2)[75], and phylogenetic trees were reconstructed with IQ-TREE (v.2.0.7)[56], with evolutionary models chosen according to the Bayesian Information Criterion. Either of 1,000 ultrafast or nonparametric bootstrap analyses—specified in each figure—were used as measures of statistical support.

Mitochondrial multiprotein phylogeny

A concatenated phylogeny of 21 mitochondrial-DNA-encoded proteins broadly conserved across eukaryotes was generated using PhyloSuite (v.1.2.2)[76]. Homologues of *atp6*, *atp8*, *atp9*, *cox1*, *cox2*, *cox3*, *cob*, *nad1*, *nad2*, *nad3*, *nad4*, *nad4L*, *nad5*, *nad6*, *nad7*, *nad9*, *rps12*, *rps19*, *rpl2*, *rpl14* and *rpl16* were aligned with MAFFT L-INS-i (v.7.313)[51] using the default parameters, trimmed with trimAL (v.1.2)[58] under the 'strict' setting and concatenated. A maximum-likelihood phylogenetic tree was calculated using IQ-TREE (v.1.6.8)[77] under the LG + F + R8 model of evolution, as determined automatically according to the Bayesian Information Criterion, and 1,000 ultrafast bootstrap replicates were carried out as a measure of statistical support.

Environmental survey

To search for the presence of Provora in nature, we downloaded environmental sequencing datasets[78–86] targeting the 18S rRNA gene (both v4 and v9 regions) from marine, freshwater and soil environments (the full list of studies is provided in Supplementary Data 1). The operational taxonomic units from each study were used as BLAST databases for BLASTn searches against Provora 18S rRNA sequences ($e = 1 \times 10^{-25}$)[54]. All resulting hits were extracted and incorporated into a eukaryotic-wide 18S rRNA gene alignment, realigned using MAFFT (v.7.222) (--auto)[51] and trimmed using trimAL (v.1.2) (gt = 0.6 for the v4 region and gt = 0.8 for the v9 region)[58]. Phylogenies were constructed using IQ-TREE (v.1.6.8)[77] and manually inspected to remove contaminants and ensure that only hits branching within the Provora were retained. The newly characterized operational taxonomic units were also used as queries to search GenBank for Provora sequences using BLAST[54]. Final phylogenies were generated in IQ-TREE (v.1.6.8) with statistical support from 1,000 ultrafast bootstraps[77]).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Raw transcriptome reads from Provora are deposited in GenBank (PRJNA866092), along with the SSU rRNA gene sequences of species (OP101998–OP102010). Assembled transcriptomes, mitochondrial genomes, materials of orthogroup and phylogenetic analyses, along with individual gene alignments, concatenated and trimmed alignments, and maximum-likelihood and Bayesian tree files for the phylogenomic dataset are available at Figshare (https://doi.org/10.6084/m9.figshare.20497143). The following databases were used in this study: NCBI nt (https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nt.gz), NCBI non-redundant database (https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz), Swiss-Prot (https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.fasta.gz), EukProt (https://figshare.com/articles/dataset/EukProt_a_database_of_genome-scale_predicted_proteins_across_the_diversity_of_eukaryotic_life/12417881/2), KEGG (https://www.genome.jp/kegg/), Pfam (http://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam32.0/). The following environmental sequencing datasets were used for 18S rRNA gene analysis: *Tara Oceans* (https://zenodo.org/record/3768510#.Y1ZtKuzMI1I), protists in European coastal waters and sediments (https://doi.org/10.1111/1462-2920.12955), Autonomous Reef Monitoring Structures (ARMS) in Red Sea (https://doi.org/10.1038/s41598-018-26332-5), Stream biofilm eukaryotic assemblages (https://doi.org/10.1016/j.ecolind.2020.106225), Deep sea basin sediments (https://doi.org/10.1038/s42003-021-02012-5), eukaryotic plankton in reef environments in Panama (https://doi.org/10.1007/s00338-020-01979-7), eukaryote communities in a high-alpine lake (https://doi.org/10.1007/s12275-019-8668-8), mountain lake microbial communities (https://doi.org/10.1111/mec.15469), microbial eukaryotes in Lake Baikal (https://doi.org/10.1093/femsec/fix073). A 320-gene dataset was used for constructing alignments for phylogenomic analyses (https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-021-22044-z/MediaObjects/41467_2021_22044_MOESM5_ESM.zip). The new taxa have been registered with the Zoobank database (http://zoobank.org/) under the following accession codes: urn:lsid:zoobank.org:act:9EE01A01-E294-415B-A36F-0FB4373183D0, urn:lsid:zoobank.org:act:A54BD0FB-7FA3-42CB-9D3D-2211FA657DC0, urn:lsid:zoobank.org:act:F6395E20-7BDF-4CBE-95FB-E4CE1E7B8185, urn:lsid:zoobank.org:act:F1E8545D-BAC1-44FF-9B6B-8FEE4AC028BB, urn:lsid:zoobank.org:act:66A5C066-890F-4F25-AAB6-5CDCE2028034, urn:lsid:zoobank.org:act:830A4372-62D9-4CE1-BFD8-9FE9EED67FED, urn:lsid:zoobank.org:act:DFE7080B-6201-455A-99CE-903103CBB049, urn:lsid:zoobank.org:act:A230EC14-DC4B-4F05-8D69-8FE0BAB3DE09, urn:lsid:zoobank.org:act:B8894608-40D4-4D16-A4D9-6F448614F22C and urn:lsid:zoobank.org:act:97B89F6F-72D6-482A-9EA7-88E5C63E6EB6.

25. Tikhonenkov, D. V., Mazei, Y. A. & Embulaeva, E. A. Degradation succession of heterotrophic flagellate communities in microcosms. *Zh. Obs. Biol.* **69**, 57–64 (2008).
26. Tikhonenkov, D. V. et al. On the origin of TSAR: morphology, diversity and phylogeny of Telonemia. *Open Biol.* **12**, 210325 (2022).
27. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
28. Keeling, P. J., Poulson, N. & McFadden, G. I. Phylogenetic diversity of parabasalian symbionts from termites, including the phylogenetic position of *Pseudotrypanosoma* and *Trichonympha*. *J. Eukaryot. Microbiol.* **45**, 643–650 (1998).

29. Medlin, L., Elwood, H. J., Stickel, S. & Sogin, M. L. The characterization of enzymatically amplified eukaryotic 16S-like rRNA-coding regions. *Gene* **71**, 491–499 (1988).
30. Tikhonenkov, D. V., Janouškovec, J., Keeling, P. J. & Mylnikov, A. P. The morphology, ultrastructure and SSU rRNA gene sequence of a new freshwater flagellate, *Neobodo borokensis* n. sp. (Kinetoplastea, Excavata). *J. Eukaryot. Microbiol.* **63**, 220–232 (2016).
31. Andrews, S. FastQC: a quality control tool for high throughput sequence data (Babraham Bioinformatics, 2010); https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
32. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614–620 (2013).
33. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
34. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
35. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
36. Laetsch, D. R. & Blaxter, M. L. BlobTools: interrogation of genome assemblies. *F1000Research* **6**, 1287 (2017).
37. Haas, B. J. et al. Denovo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
38. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
39. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
40. Shen, W. & Ren, H. TaxonKit: a practical and efficient NCBI taxonomy toolkit. *J. Genet. Genomics* **48**, 844–850 (2021).
41. Richter, D. J. et al. EukProt: a database of genome-scale predicted proteins across the diversity of eukaryotes. *Peer Community Journal* **2**, e56 (2022).
42. Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
43. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.* **49**, D545–D551 (2021).
44. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C. & Kanehisa, M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* **35**, W182–W185 (2007).
45. Burki, F. The eukaryotic tree of life from a global phylogenomic perspective. *Cold Spring Harb. Perspect. Biol.* **6**, a016147 (2014).
46. Waskom, M. et al. mwaskom/Seaborn: v0.8.1 (September 2017). *Zenodo* https://doi.org/10.5281/zenodo.883859 (2017).
47. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
48. Finn, R. D. et al. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44**, D279–D285 (2016).
49. Letunic, I. & Bork, P. 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res.* **46**, D493–D496 (2018).
50. Almagro Armenteros, J. J. et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* **37**, 420–423 (2019).
51. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
52. Burns, J. A., Pittis, A. A. & Kim, E. Gene-based predictive models of trophic modes suggest Asgard archaea are not phagocytotic. *Nat. Ecol. Evol.* **2**, 697–704 (2018).
53. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
54. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
55. Hall, T. A. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* **41**, 95–98 (1999).
56. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
57. Whelan, S., Irisarri, I. & Burki, F. PREQUAL: detecting non-homologous characters in sets of unaligned homologous sequences. *Bioinformatics* **34**, 3929–3930 (2018).
58. Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
59. Roure, B., Rodriguez-Ezpeleta, N. & Philippe, H. SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. *BMC Evol. Biol.* **7**, S2 (2007).
60. Lartillot, N., Rodrigue, N., Stubbs, D. & Richer, J. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst. Biol.* **62**, 611–615 (2013).
61. Dayhoff, M., Schwartz, R. & Orcutt, B. in *Atlas of Protein Sequence and Structure* (ed. Dayhoff, M.) 345–352 (National Biomedical Research Foundation, 1978).
62. Susko, E. & Roger, A. J. On reduced amino acid alphabets for phylogenetic inference. *Mol. Biol. Evol.* **24**, 2139–2150 (2007).
63. Lartillot, N. & Philippe, H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* **21**, 1095–1109 (2004).
64. Quang le, S., Gascuel, O. & Lartillot, N. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* **24**, 2317–2323 (2008).
65. Wang, H. C., Minh, B. Q., Susko, E. & Roger, A. J. Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. *Syst. Biol.* **67**, 216–235 (2018).
66. Kück, P. & Struck, T. H. BaCoCa—a heuristic software tool for the parallel assessment of sequence biases in hundreds of gene and taxon partitions. *Mol. Phylogenet. Evol.* **70**, 94–98 (2014).
67. Shimodaira, H. An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* **51**, 492–508 (2002).
68. Kumar, S., Stecher, G. & Tamura, K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **33**, 1870–1874 (2016).
69. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
70. Dierckxsens, N., Mardulyn, P. & Smits, G. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res.* **45**, e18 (2017).
71. Kuznetsov, A. & Bollin, C. J. in *Multiple Sequence Alignment* (ed. Katoh, K.) 261–295 (Springer, 2021).
72. Lohse, M., Drechsel, O., Kahlau, S. & Bock, R. OrganellarGenomeDRAW—a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Res.* **41**, W575–W581 (2013).
73. Johnson, P. Z., Kasprzak, W. K., Shapiro, B. A. & Simon, A. E. RNA2Drawer: geometrically strict drawing of nucleic acid structures with graphical structure editing and highlighting of complementary subsequences. *RNA Biol.* **16**, 1667–1671 (2019).
74. Burger, G., Gray, M. W., Forget, L. & Lang, B. F. Strikingly bacteria-like and gene-rich mitochondrial genomes throughout jakobid protists. *Genome Biol. Evol.* **5**, 418–438 (2013).
75. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* **10**, 210 (2010).
76. Zhang, D. et al. PhyloSuite: an integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. *Mol. Ecol. Resour.* **20**, 348–355 (2020).
77. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
78. Ibarbalz, F. M. et al. Global trends in marine plankton diversity across kingdoms of life. *Cell* **179**, 1084–1097 (2019).
79. Massana, R. et al. Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing. *Environ. Microbiol.* **17**, 4035–4049 (2015).
80. Gendron, E. M. S., Darcy, J. L., Hell, K. & Schmidt, S. K. Structure of bacterial and eukaryote communities reflect in situ controls on community assembly in a high-alpine lake. *J. Microbiol.* **57**, 852–864 (2019).
81. Minerovic, A. D. et al. 18S-V9 DNA metabarcoding detects the effect of water-quality impairment. *Ecol. Indic.* **113**, 106225 (2020).
82. Pearman, J. K. et al. Cross-shelf investigation of coral reef cryptic benthic organisms reveals diversity patterns of the hidden majority. *Sci. Rep.* **8**, 8090 (2018).
83. Rodas, A. M. et al. Eukaryotic plankton communities across reef environments in Bocas del Toro Archipelago, Panamá. *Coral Reefs* **39**, 1453–1467 (2020).
84. Schoenle, A. et al. High and specific diversity of protists in the deep-sea basins dominated by diplonemids, kinetoplastids, ciliates and foraminiferans. *Commun. Biol.* **4**, 501 (2021).
85. Schulhof, M. A. et al. Sierra Nevada mountain lake microbial communities are structured by temperature, resources and geographic location. *Mol. Ecol.* **29**, 2080–2093 (2020).
86. Yi, Z. et al. High-throughput sequencing of microbial eukaryotes in Lake Baikal reveals ecologically differentiated communities and novel evolutionary radiations. *FEMS Microbiol. Ecol.* **93**, fix073 (2017).

Acknowledgements We thank M. Vermeij and the staff at the CARMABI research station for field sampling support; and N. Kosolapova for help with sample collection in the Arctic. This research was supported by grants from the Russian Foundation for Basic Research (to D.V.T., grant no. 20-34-70049), the Tyumen Oblast Government, as part of the West-Siberian Interregional Science and Education Center's project no. 89-DON (2) (to D.V.T.), the Ministry of Science and Higher Education of the Russian Federation within the framework of the Federal Scientific and Technical Program for the Development of Genetic Technologies for 2019-2027 (agreement no. 075-15-2021-1345, unique identifier RF-193021X0012), the Gordon and Betty Moore Foundation (to P.J.K., https://doi.org/10.37807/GBMF9201), GenomeBC and the Natural Sciences and Engineering Research Council of Canada (to P.J.K., grant no. 2019-03994), and was carried out within the framework of state assignment no. 121051100102-2.

Author contributions D.V.T., K.V.M., R.M.R.G. and P.J.K. designed the study. D.V.T. and A.P.M. discovered the organisms and isolated the cultures. D.V.T. generated material for sequencing. A.O.B., S.A.K., D.G.Z., A.S.B., K.I.P. and D.V.T. performed light and electron microscopy and cultured the cells. K.V.M. and R.M.R.G. performed transcriptomic analyses and phylogenetic analyses. V.M. and V.V.A. performed the environmental distribution analysis and phylogenetic analysis of the SSU rRNA. D.V.T., K.V.M., R.M.R.G. and P.J.K. wrote the manuscript with input from all of the authors.

Competing interests The authors declare no competing interests.

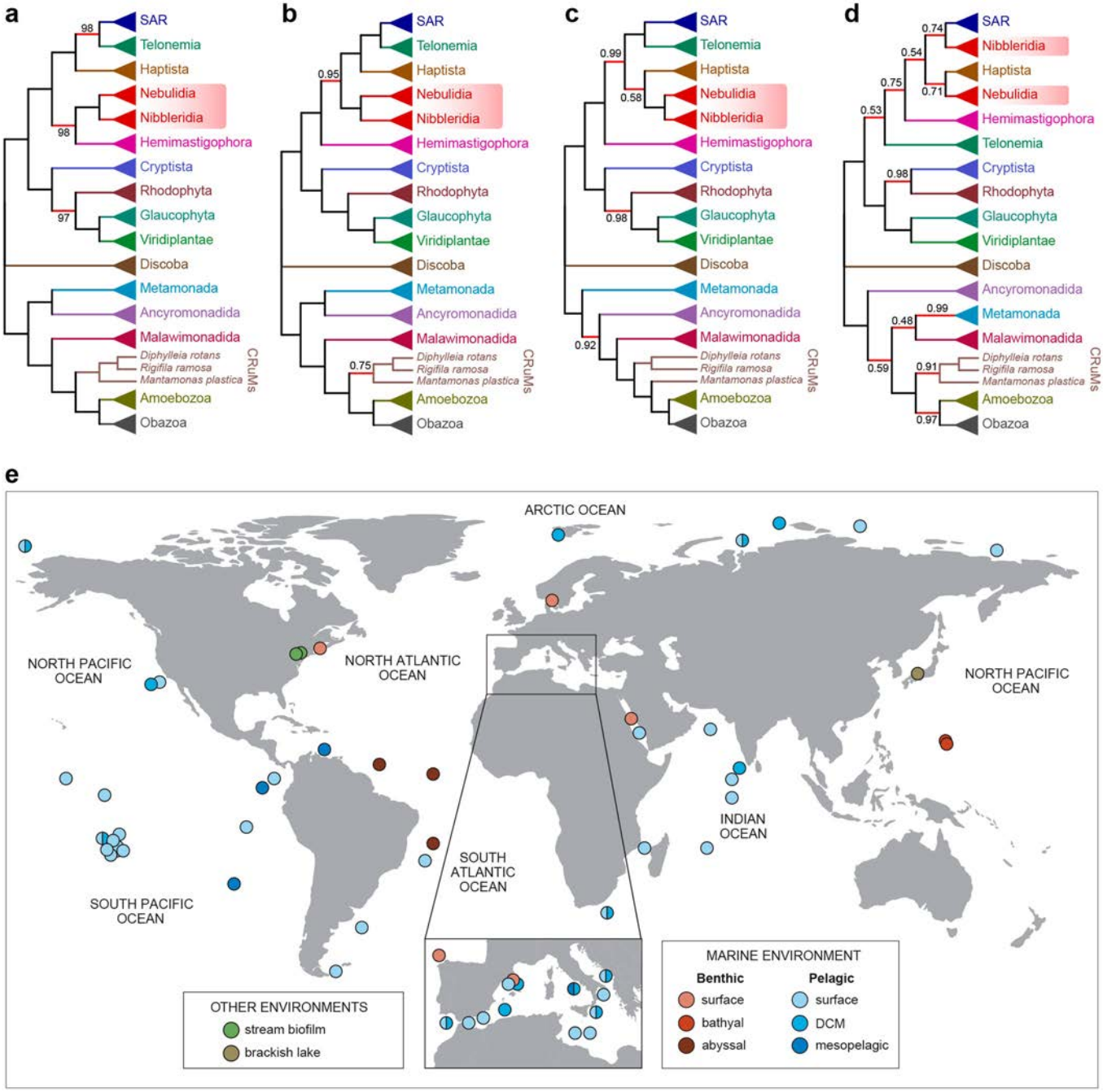
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-022-05511-5.

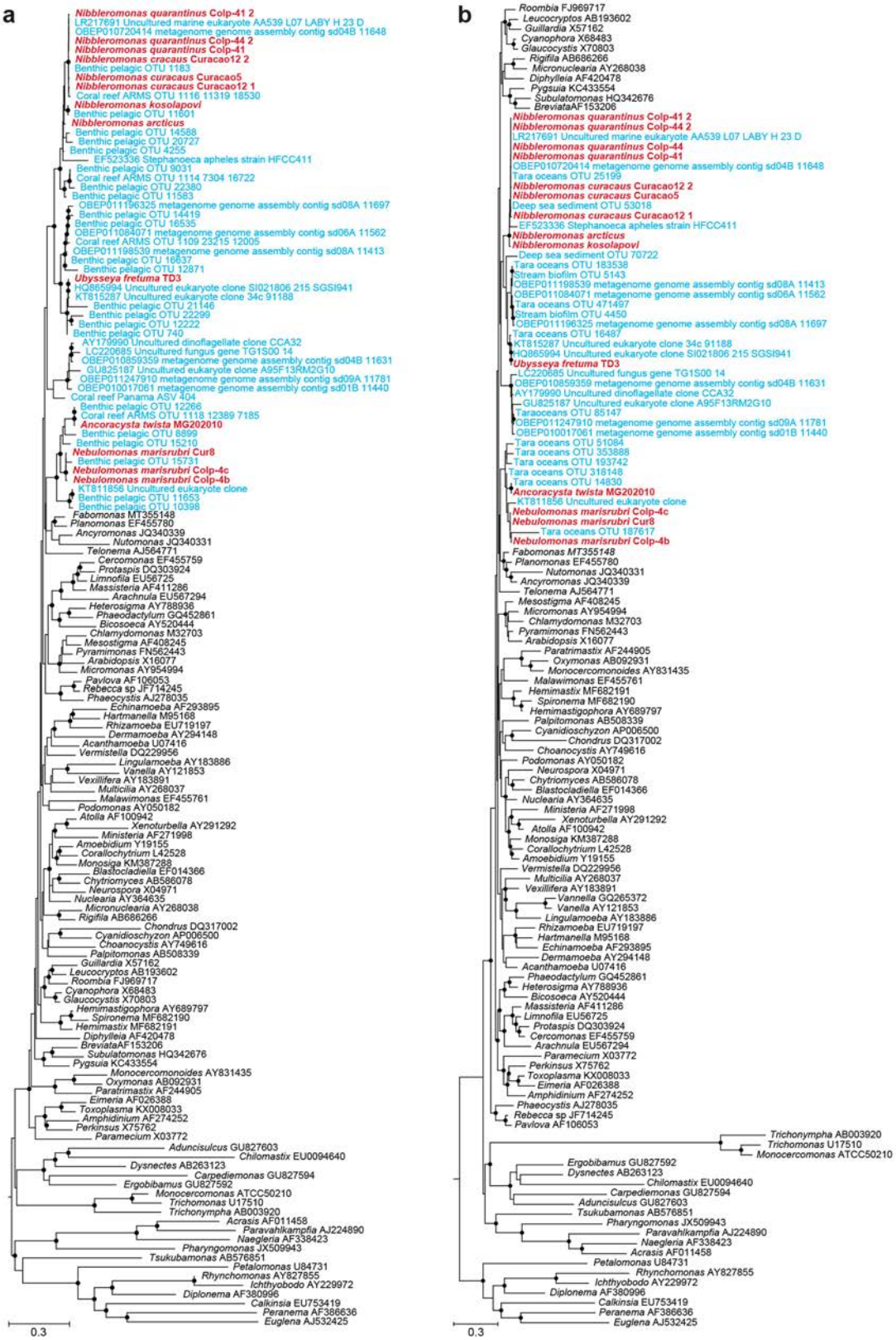
Correspondence and requests for materials should be addressed to Denis V. Tikhonenkov.

Peer review information Nature thanks Thijs Ettema, James McInerney and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

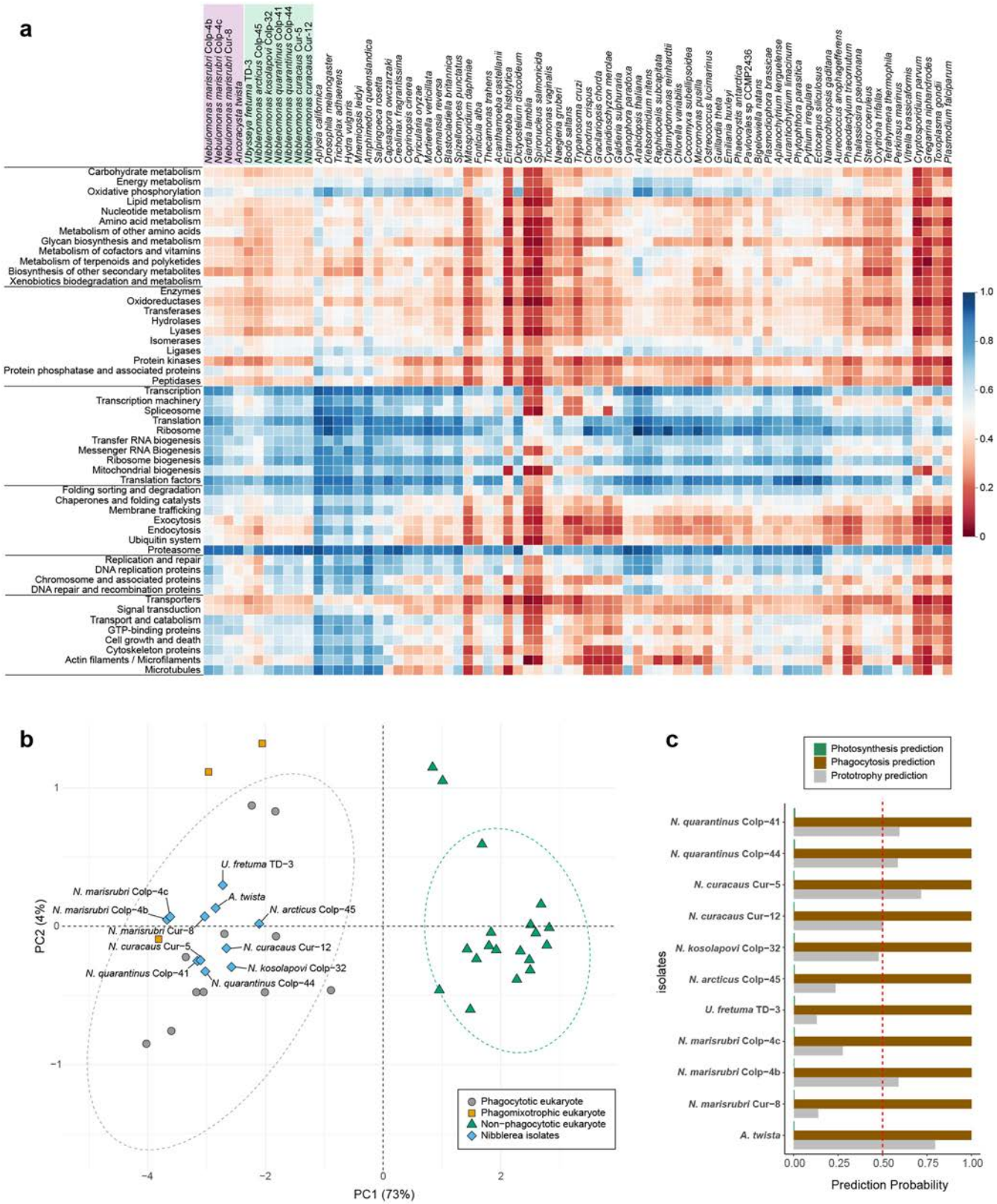
Reprints and permissions information is available at http://www.nature.com/reprints.



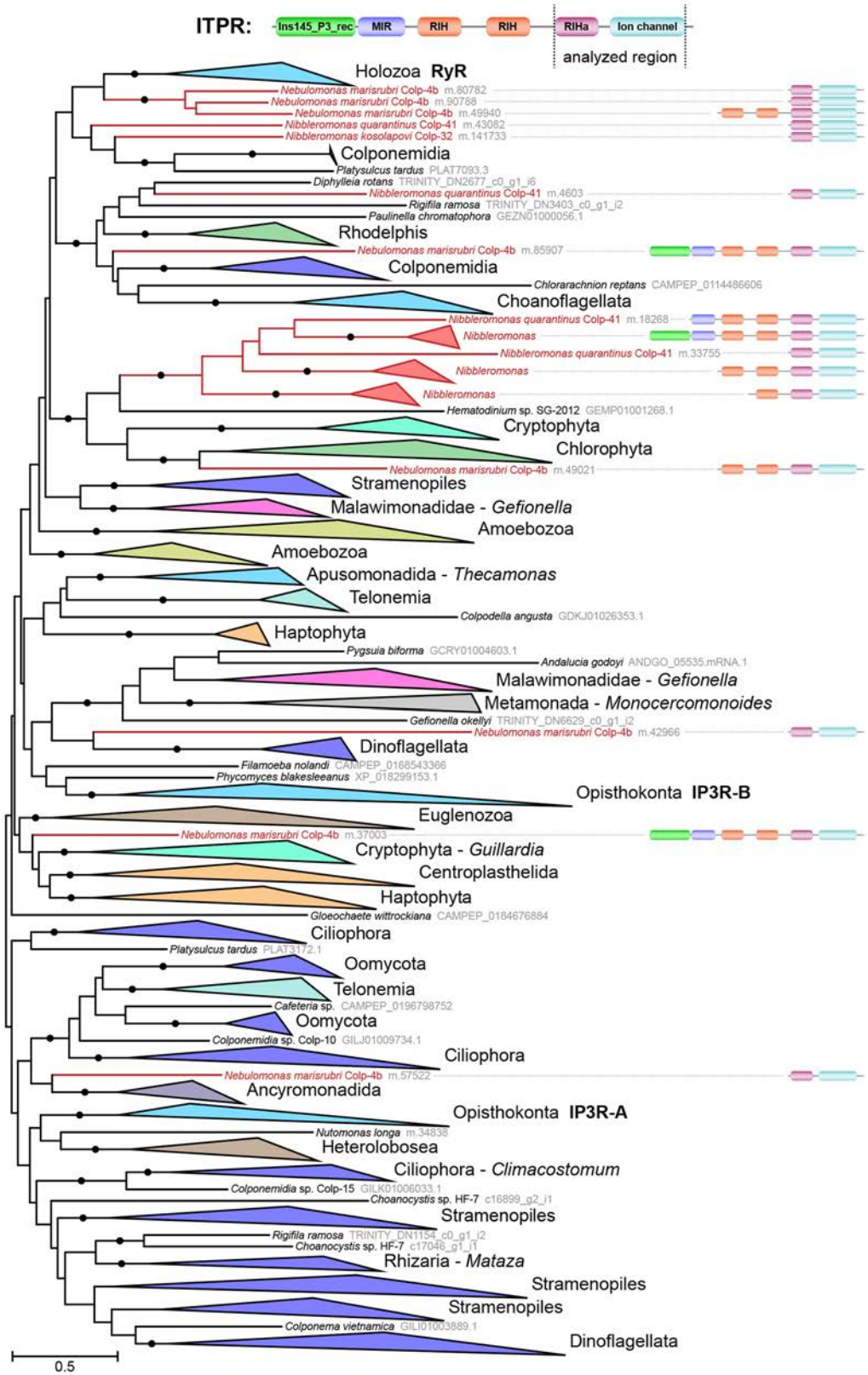
Extended Data Fig. 1 | Outline of tree topologies obtained in the phylogenomic analyses and the geographical distribution of Provora. (a) Maximum-likelihood tree topology obtained with the 320-gene dataset; nodes with support values below 100% (PMSF model, 100 replicates) are labelled red, and the corresponding values are provided next to the tree nodes; established eukaryotic groups with full support in the analysis are collapsed and shown in the tree schematically with triangles. **(b) PhyloBayes consensus tree topology obtained using four analysis chains with the native 320-gene dataset**; posterior probabilities are shown for tree nodes that fail to achieve full support in the analysis. **(c) PhyloBayes consensus tree topology obtained with the Dayhoff 6-recoded 320-gene dataset**; the low posterior probability (0.58 pp) for the union of Provora and Haptista reflects the marginal support for this group in all four analysis chains, rather than the lack of convergence between the chains (maxdiff = 0.27). **(d) PhyloBayes consensus tree topology obtained with the SR4-recoded 320-gene dataset**. **(e) Geographical distribution of environmental sequences of 18S rRNA belonging to Provora**.



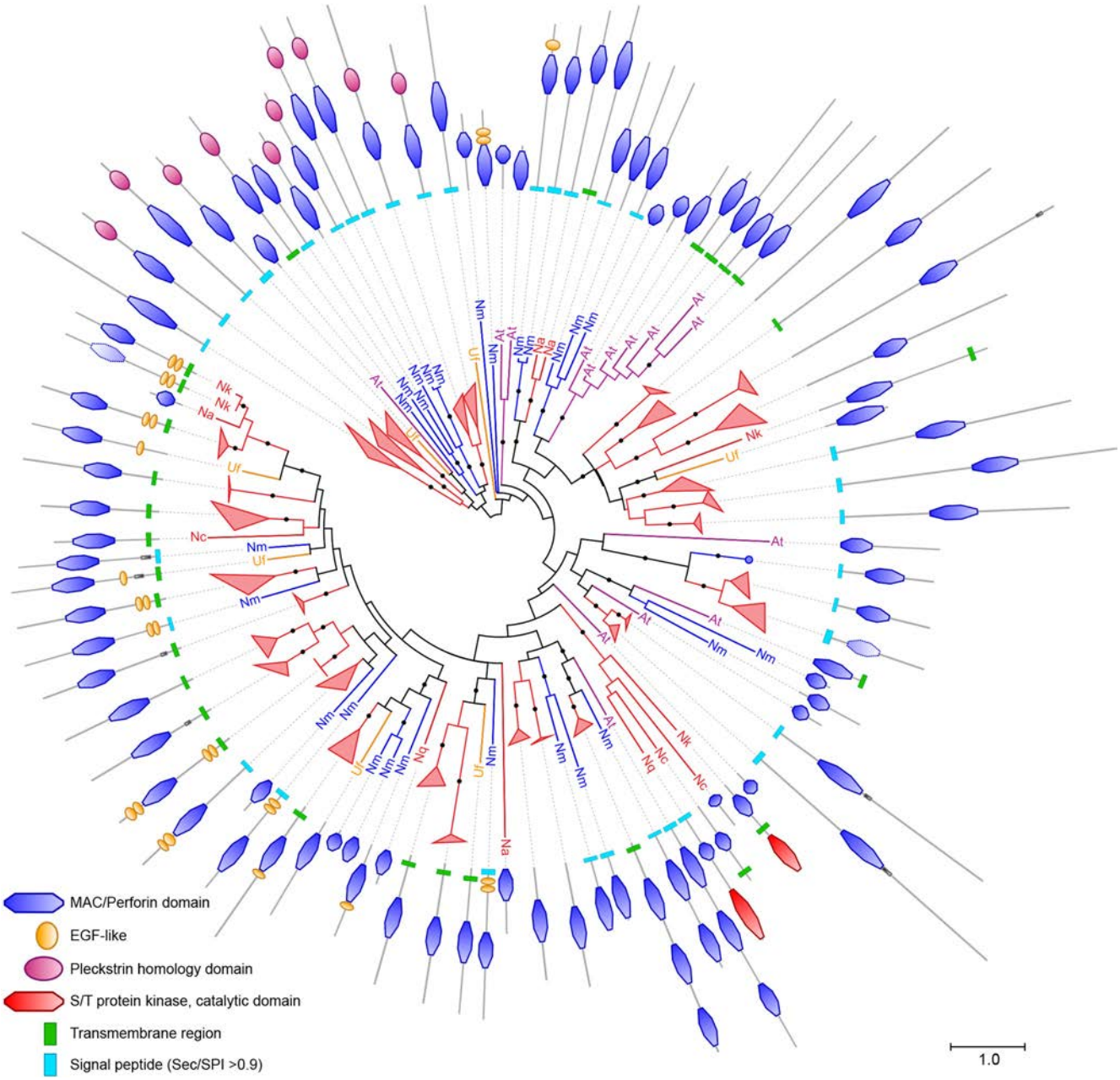
Extended Data Fig. 2 | Phylogenies with variable regions of 18S rRNA featuring identified environmental sequences belonging to Provora. (a) Phylogenetic tree based on the V4 region of the 18S rRNA gene showing the diversity of environmental lineages of Provora. (b) Phylogenetic tree based on the V9 region of the 18S rRNA gene. The 18S rRNA of Provora described in this paper are shown in red. Environmental sequences related to the members of Provora are labelled in blue. Bootstrap values ≥ 90% are indicated with black circles at the tree nodes.



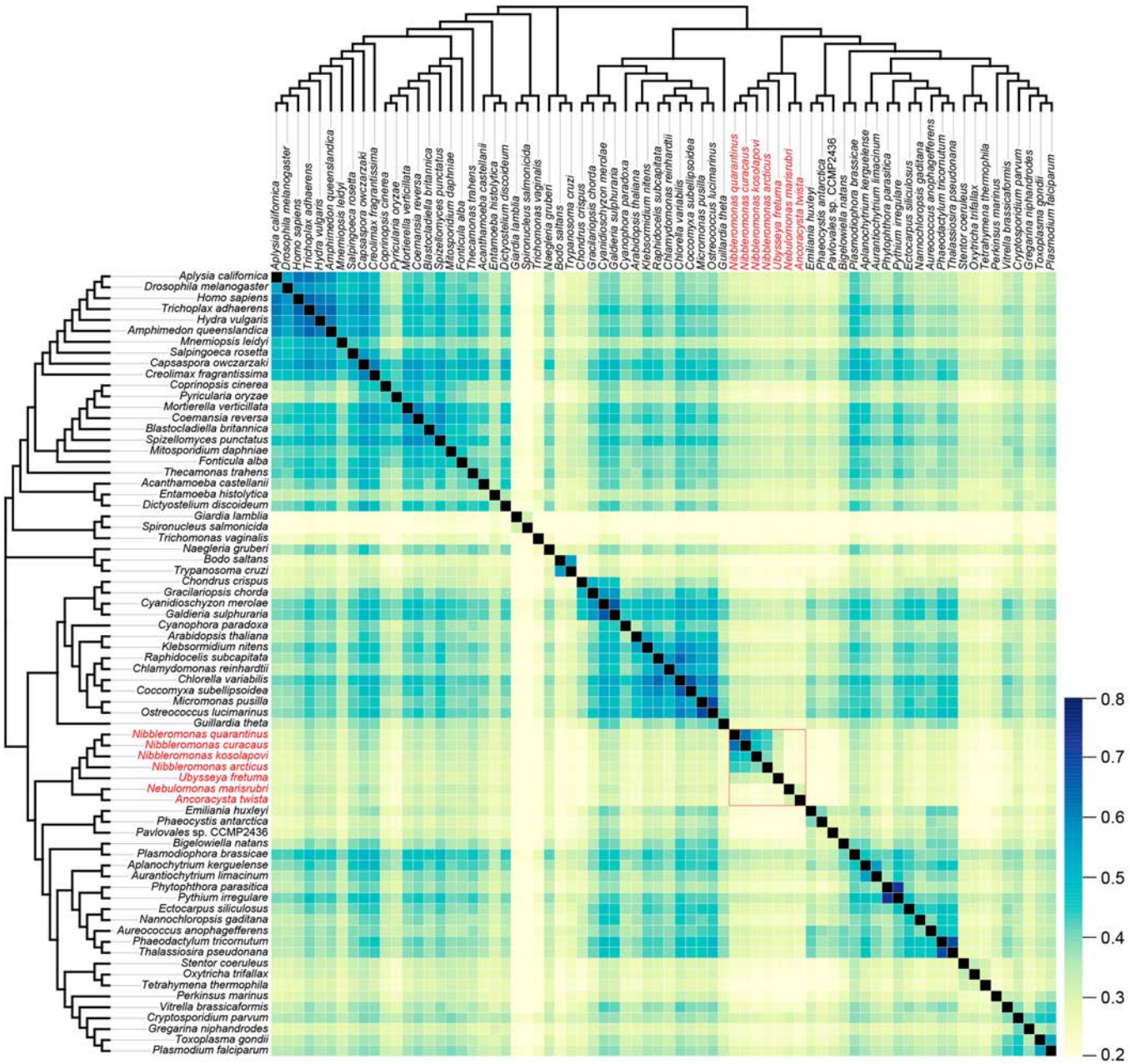
Extended Data Fig. 3 | Conservation of functional categories and trophic mode prediction for the transcriptomes of Provora. (a) Heatmap of annotated KEGG orthology entry counts (presence/absence data) for functional categories defined by BRITE in the transcriptomic data of Provora isolates and the genomic data of eukaryotic organisms; the counts only include entries inferred to be ancestral for eukaryotes by the Dollo parsimony principle: entries that only have hits in one of the major eukaryotic subdivisions (Diaphoretickes, Discoba or Amorphea) were excluded; the counts were normalized to the inferred ancestral eukaryotic KEGG orthologs. **(b) Principal component analysis plot with gene ontology category scores for categories associated with free-living phagocytic organisms**; **(c) Prediction probabilities of trophic modes (phagocytosis, prototrophy, photosynthesis) in Provora isolates, conducted by the Trophic Mode Prediction Tool**.



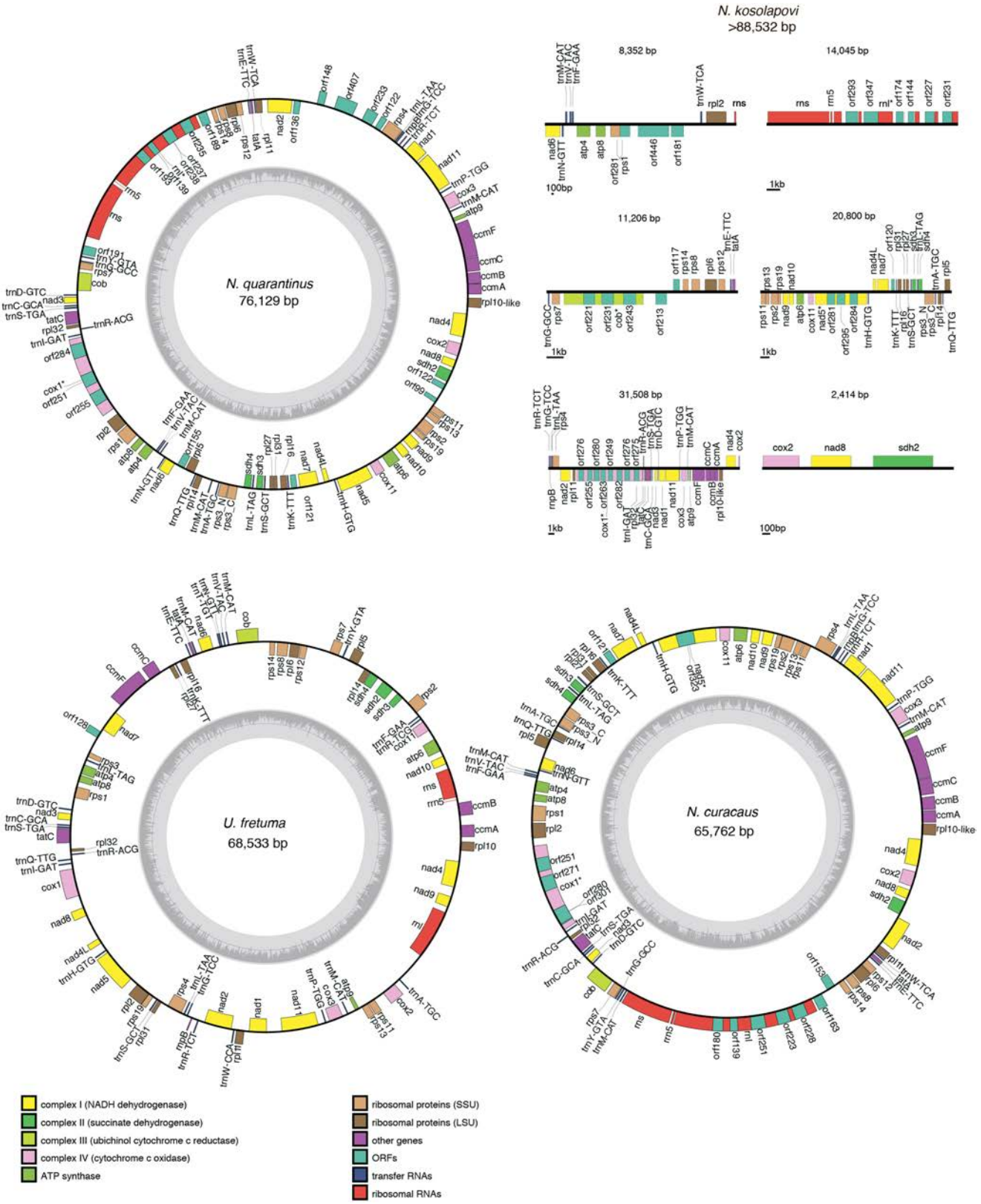
Extended Data Fig. 4 | Maximum likelihood phylogenetic tree with eukaryotic members of the inositol trisphosphate receptor family, identified by the presence of a RyR and IP3R homology associated domain (RIHa, PF08454) and an ion channel domain (PF00520). The phylogeny was reconstructed by IQ-TREE using an alignment with 396 eukaryotic sequences, spanning the RIHa and ion channel regions of the proteins; reconstruction was done under the best-fitting LG+F+R10 model of evolution, and node support was evaluated with 1000 UFBoot replicates; nodes with over 95% support are marked with black circles; clades uniting members of a single taxon are collapsed in the tree and labelled in accordance with their taxonomy; branches that belong to Provora are coloured red; protein domain architectures are displayed for the IP3R family sequences in Provora: Ins145_P3_rec (PF08709), MIR (PF02815), RIH (PF01365), RIHa (PF08454), Ion channel (PF00520).



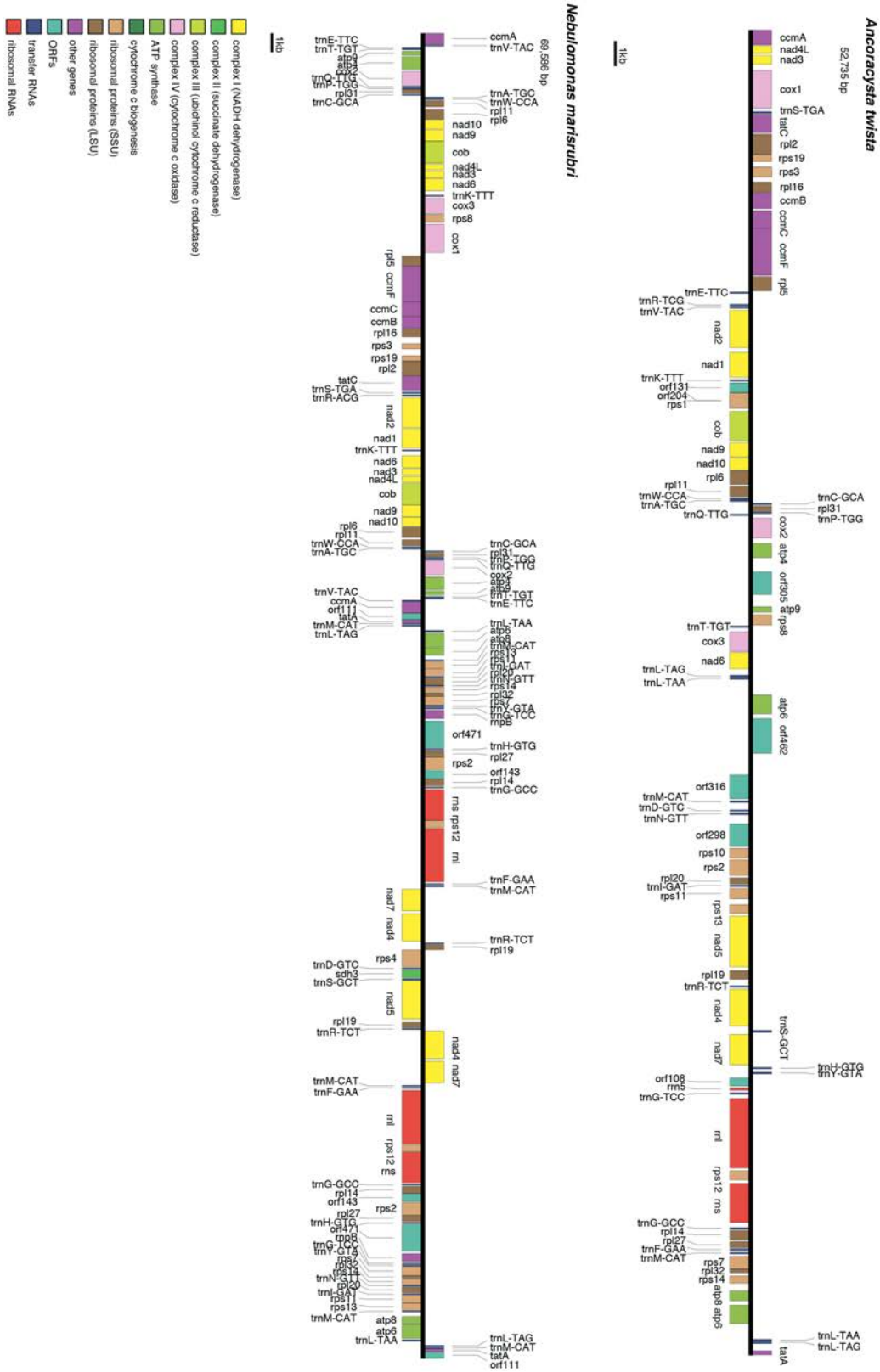
Extended Data Fig. 5 | Maximum likelihood phylogenetic tree with MACPF domain-containing proteins in Provora. The phylogeny was reconstructed by IQ-TREE with the best-fitting WAG+F+R5 model of evolution; node support was evaluated with 1000 UFBoot replicates, and nodes with over 95% support are marked with black circles; clades uniting putatively orthologous MACPF sequences in *Nibbleromonas* species are collapsed; species name abbreviations: At – *Ancoracysta twista*, Nm – *Nebulomonas marisrubri*, Uf – *Ubysseya fretuma*, Na – *Nibbleromonas arcticus*, Nk – *Nibbleromonas kosolapovi*, Nc – *Nibbleromonas curacaus*, Nq – *Nibbleromonas quarantinus*; the domain architectures of MACPF proteins identified using SMART searches are shown; MACPF domains outlined with dotted lines correspond to findings below the default detection threshold.



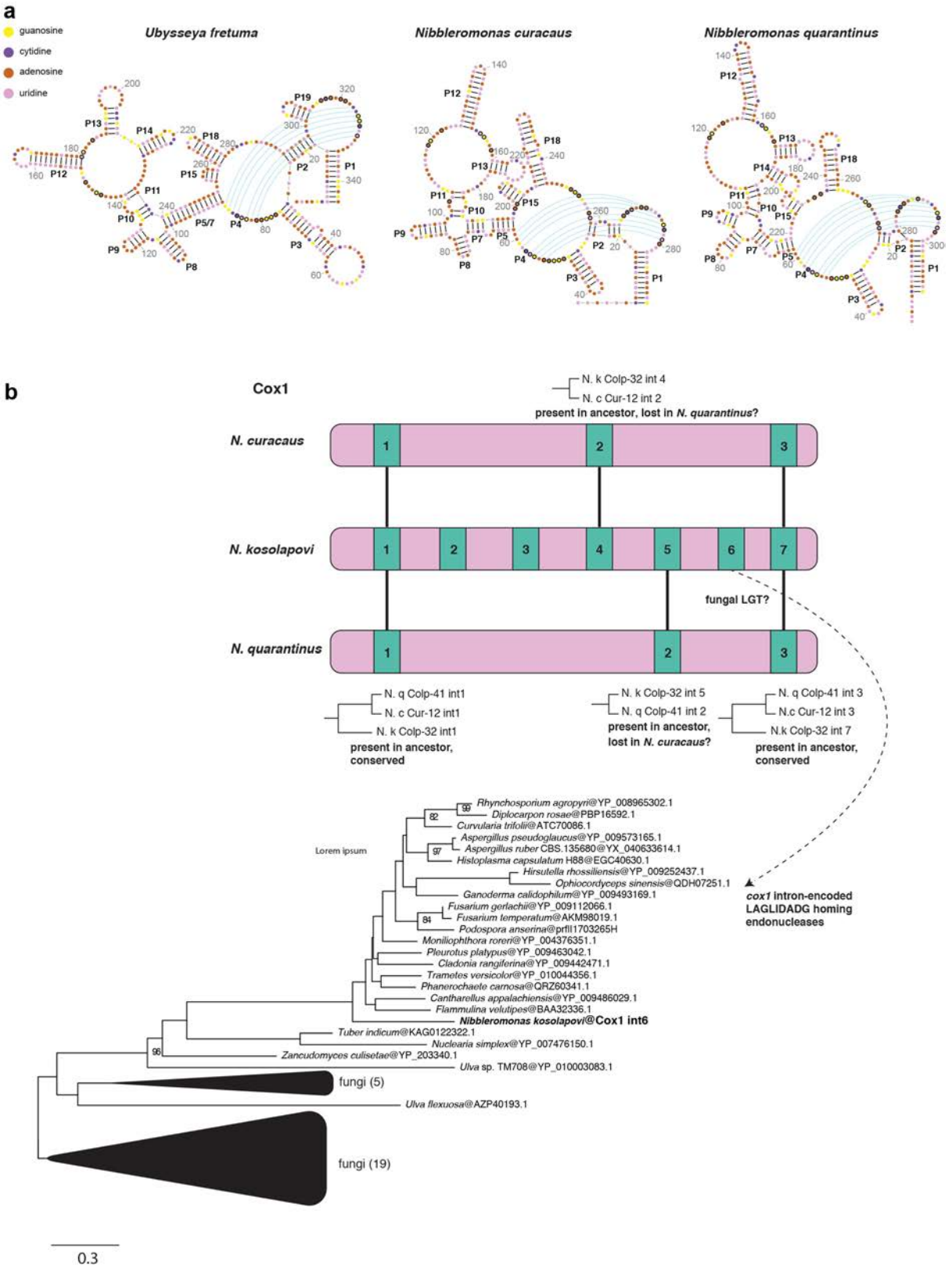
Extended Data Fig. 6 | Proportions of shared to total orthogroup counts in pairwise comparisons of eukaryotic organisms. Arithmetic means of the proportions of shared orthogroups between pairs of genomes or transcriptomes are shown using a heatmap; the organisms are grouped using a tree, which summarizes the current concept of eukaryotic phylogeny; orthogroup inference for members of the Provora lineage relied on the transcriptomic data; the Provora species are labelled in red, and the corresponding intragroup comparisons are outlined with a red square in the heatmap.



Extended Data Fig. 7 | Mitochondrial genome maps of nibblerids. Nibblerid mitochondrial genomes are typically circular-mapping, and gene-rich. All maps were edited to arbitrarily start at the *ccmA* gene. Genes are colour-coded according to their functional classification, as shown in the legend.

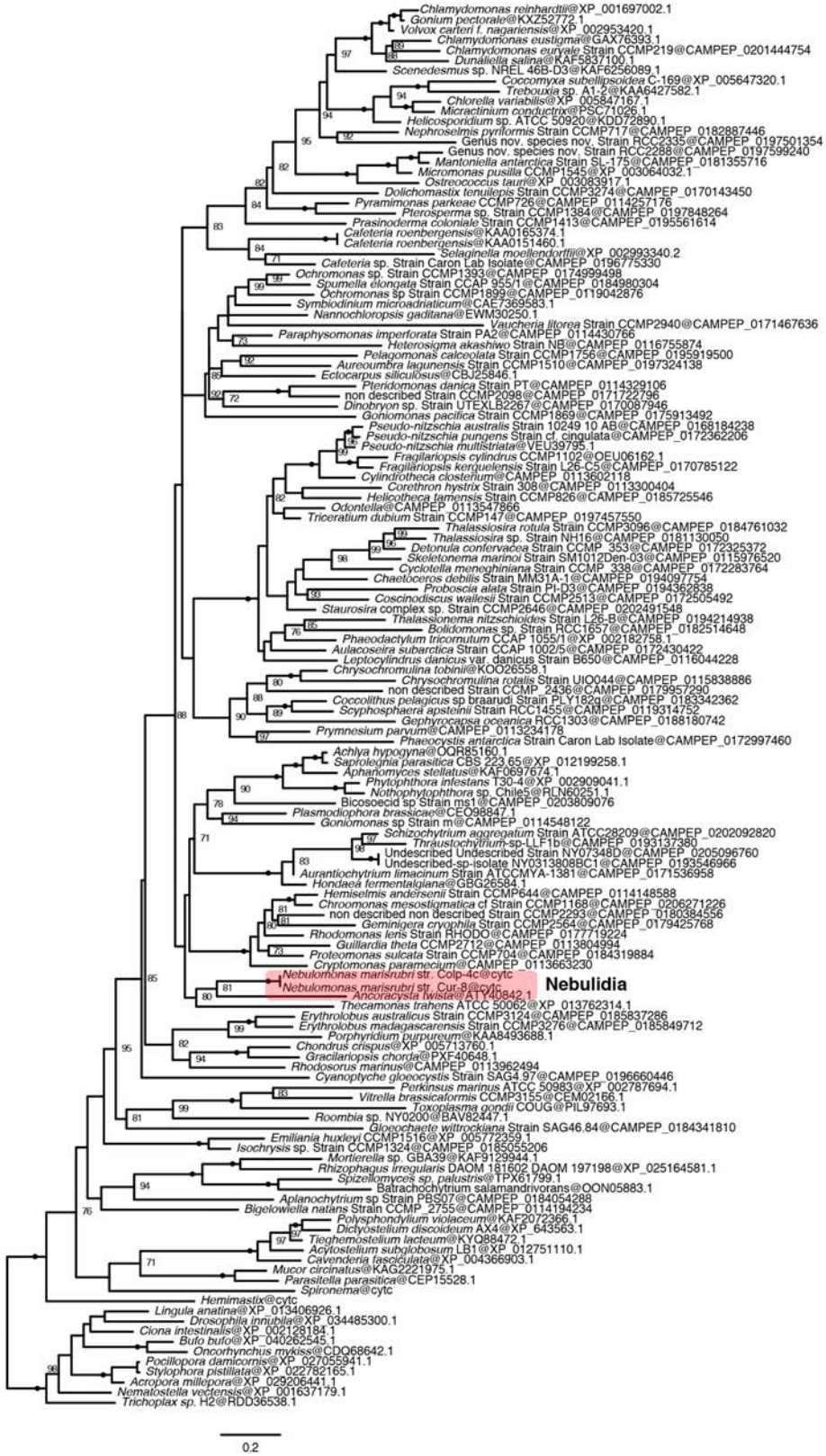


Extended Data Fig. 8 | Mitochondrial genome maps of the nebulids, *Ancoracysta twista* and *Nebulomonas marisrubri*. Nebulid mitochondrial genomes are circular-mapping, but are presented in a linear format to facilitate comparison of gene order. Mitochondrial genomes of *A. twista* (NC_036491.1) and *N. marisrubri* each contain duplications due to the presence of inverted repeats. All maps were edited to arbitrarily start at the *ccmA* gene. Genes are colour-coded according to their functional classification, as shown in the legend.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Provoran mitochondrial genomes retain ancestral features, but their sizes are variable due to group-I intron accumulation. (a) Secondary structure predictions of mitochondrion-encoded RNAse P RNAs from *Ubysseya fretuma*, *Nibbleromonas quarantinus*, and *N. curacaus*; genes encoding rnpB have been identified in a small and phylogenetically disparate collection of eukaryotes, and are often very dissimilar from their counterparts in Alphaproteobacteria. All nibblerid mitochondrial genomes described here encode rnpB, and bear a strong resemblance to bacterial and jakobid rnpB homologs. Nucleotides with black borders indicate positions that are found in eubacterial consensus and jakobid rnpB homologs, and conserved helices are noted (P1-19). **(b) Group-I introns that encode LAGLIDADG homing endonucleases are present in mitochondrial genomes in the genus *Nibbleromonas*;** phylogenetic relationships between intron-encoded homing endonucleases of *cox1* are shown as an exemplar of introns presence in nibblerid mitochondrial genomes. Some homologous homing endonucleases are present in the same position of *N. kosolapovi*, *N. quarantinus* and *N. curacaus cox1* (*e.g.*, intron 1 of each species), indicating that they were present in their common ancestor and have been broadly retained. Other introns are found in only *N. kosolapovi*, and one of *N. quarantinus* or *N. curacaus*, suggesting lineage-specific intron loss. In contrast, the endonuclease encoded in intron 6 of *N. kosolapovi cox1* was likely gained via lateral transfer from fungi, where the endonuclease is also encoded by *cox1* introns.



Extended Data Fig. 10 | Maximum likelihood phylogenetic tree of nucleus-encoded holocytochrome *c* synthase (HCCS) from diverse eukaryotes (140 sites, LG+R7 model, 1000 ultrafast bootstraps). A prior report demonstrated that the nebulid *Ancoracysta twista* retains both mitochondrion-encoded type-I and nucleus-encoded type-III cytochrome *c* maturation systems. Although nibblerids retain only the former, multiple strains of the newly described nebulid, *Nebulomonas marisrubri*, also have both types of cytochrome *c* maturation systems. In our phylogenetic reconstruction, *N. marisrubri* and *A. twista* HCCS proteins are monophyletic, though with only moderate statistical support. One thousand ultrafast bootstrap replicates were performed as a measure of statistical support. For clarity, bipartitions receiving full statistical support are represented by black circles and values less than 70 are not presented.

nature portfolio

Corresponding author(s): Denis Tikhonenkov

Last updated by author(s): Oct 25, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Sequence quality: FastQC v0.10.1 Transcriptome assembly: Trinity v2.4.0; TransDecoder v5.5.0; Genome assembly: SPAdes v3.14.1, NOVOPlasty v4.3 Sequencing read processing: Trimmomatic v0.36; PEAR v0.9.6; BBMap v37.36
Data analysis	Transcriptome and genome analysis: BLAST v2.2.30+; CD-HIT v4.6; HMMER3.1 (hmmer.org); OGDRAW v1.3.1; RNA2Drawer; BUSCO v3.0.0; PfamScan v1.6 (http://ftp.ebi.ac.uk/pub/databases/Pfam/Tools); SMART v9.0; SignalP v5.0; Trophic Mode Prediction Tool v1.0.0; MFannot (https://megasun.bch.umontreal.ca/apps/mfannot/); NCBI Genome Workbench v3.6.0 Transcriptomic data filtering: BlobTools; DIAMOND v0.9.24; TaxonKit v0.3.0 Phylogenomic and phylogenetic analysis: MAFFT v7.222; MAFFT v7.313; trimAL v1.2; trimAl v1.4; BMGE v1.1.2; SCaFoS v1.2.5; PhyloSuite v1.2.2; IQ-TREE v1.6.8; IQ-TREE v2.0.7; IQ-TREE v1.6.10, IQ-TREE v1.6.12; PhyloBayes MPI v1.8c; PREQUAL v1.02; BaCoCa v1.105.r Comparative genomics: OrthoFinder v2.5.4; KEGG Automatic Annotation Server (KAAS) v2.1 Visualization: Seaborn v0.8.1; MEGA v7.0.21; BioEdit v7.2.5

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Raw transcriptome reads from Provora are deposited in GenBank (PRJNA866092), along with the SSU rRNA gene sequences of species (OP101998-OP102010). Assembled transcriptomes, mitochondrial genomes, materials of orthogroup and phylogenetic analyses, along with individual gene alignments, concatenated and trimmed alignments, and maximum-likelihood and Bayesian tree files for the phylogenomic dataset are available at figshare with the identifier doi.org/10.6084/m9.figshare.20497143. The following databases were used in this study: NCBI nt database (https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nt.gz), NCBI non-redundant database (https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz), Swiss-Prot database (https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.fasta.gz), EukProt database (https://figshare.com/articles/dataset/EukProt_a_database_of_genome-scale_predicted_proteins_across_the_diversity_of_eukaryotic_life/12417881/2), KEGG database (https://www.genome.jp/kegg/), Pfam database (http://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam32.0/). Environmental sequencing datasets were used for 18S rRNA gene analysis: Tara Oceans (https://zenodo.org/record/3768510#.Y1ZtKuzMI1I), Protists in European coastal waters and sediments (https://doi.org/10.1111/1462-2920.12955), Autonomous Reef Monitoring Structures (ARMS) in Red Sea (https://doi.org/10.1038/s41598-018-26332-5), Stream biofilm eukaryotic assemblages (https://doi.org/10.1016/j.ecolind.2020.106225), Deep sea basin sediments (https://doi.org/10.1038/s42003-021-02012-5), Eukaryotic plankton in reef environments in Panama (https://doi.org/10.1007/s00338-020-01979-7), Eukaryote communities in a high-alpine lake (https://doi.org/10.1007/s12275-019-8668-8), Mountain lake microbial communities (https://doi.org/10.1111/mec.15469), Microbial eukaryotes in lake Baikal (https://doi.org/10.1093/femsec/fix073); 320-gene dataset was used for constructing alignments for phylogenomic analyses (https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-021-22044-z/MediaObjects/41467_2021_22044_MOESM5_ESM.zip). The novel taxa have been registered with the Zoobank database (http://zoobank.org/) urn:lsid:zoobank.org:act: 9EE01A01-E294-415B-A36F-0FB4373183D0, urn:lsid:zoobank.org:act:A54BD0FB-7FA3-42CB-9D3D-2211FA657DC0, urn:lsid:zoobank.org:act: F6395E20-7BDF-4CBE-95FB-E4CE1E7B8185, urn:lsid:zoobank.org:act:F1E8545D-BAC1-44FF-9B6B-8FEE4AC028BB, urn:lsid:zoobank.org:act:66A5C066-890F-4F25-AAB6-5CDCE2028034, urn:lsid:zoobank.org:act:830A4372-62D9-4CE1-BFD8-9FE9EED67FED, urn:lsid:zoobank.org:act:DFE7080B-6201-455A-99CE-903103CBB049, urn:lsid:zoobank.org:act:A230EC14-DC4B-4F05-8D69-8FE0BAB3DE09, urn:lsid:zoobank.org:act:B8894608-40D4-4D16-A4D9-6F448614F22C, urn:lsid:zoobank.org:act:97B89F6F-72D6-482A-9EA7-88E5C63E6EB6.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	This study did not involve human research
Population characteristics	This study did not involve human research
Recruitment	This study did not involve human research
Ethics oversight	This study did not involve human research

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	In this study, we describe ten new strains of microbial predators, which collectively form a diverse new supergroup of eukaryotes Provora. We performed detailed ultrastructural, transcriptomic/genomic, and phylogenomic analyses, and showed that Provora is genetically and morphologically distinct from all other eukaryotes
Research sample	This research describes three new genera and five new species from a new phylum of predatory eukaryotic microbes that is the sister lineage of the Haptista+TSAR assemblage, possibly also including Hemimastigophora. The organisms were collected from marine habitats, including coral reefs, nearshore sediments, and the water column.
Sampling strategy	Sample size is not relevant to the present study.
Data collection	Samples were collected from marine sediments, water column, and corals, and the new organisms were subsequently grown in the

Data collection	laboratory. Microscopic data were recorded by Denis Tikhonenkov. Sequencing data were generated by D. Tikhonenkov. Transcriptome and genome data were assembled by K. Mikhailov and R. Gawryluk.
Timing and spatial scale	Sampling relevant to the present study was carried out seven times: in the Strait of Georgia, British Columbia, June 13, 2017; Arctic waters of the Kara Sea, September 19, 2015; Arctic waters of the East Siberian Sea, September 5, 2017; shoreland of Quarantine Bay, Black Sea, May 13, 2017; sea waters of the Curaçao island, April 24, 2018; Red Sea, Sharm El Sheikh, April 2015; Kazachya Bay, Black Sea, September 1, 2018. We had no reason to expect to find the organisms that we did, so there is no specific rationale to sampling sites.
Data exclusions	Sequencing data from prey organisms were excluded from the analyses for studied predatory protists. To do this, we subtracted transcripts derived from prey (kinetoplastids) and any non-eukaryotic transcripts from the total datasets. The raw data associated with this will be accessible in the raw read files deposited in the NCBI SRA database.
Reproducibility	Microscopic analyses were conducted several times. Phylogenomic analyses were carried out with a number of different approaches (maximum likelihood, Bayesian etc.) and all associated datasets have been made available.
Randomization	Randomization is not relevant to the present study because organisms were not allocated into groups.
Blinding	Blinding was not relevant to the present study.

Did the study involve field work? ☒ Yes ☐ No

Field work, collection and transport

Field conditions	Climatic conditions in the field were not recorded and are not relevant to the study.
Location	1) Strait of Georgia, British Columbia, Canada (49°10'366'' N, 123°28'50'' W) 2) Arctic waters of the Kara Sea (75°53'16.8'' N, 89°30'28.8'' E) 3) Arctic waters of the East Siberian Sea (71°27'59.8'' N, 152°53'59.3'' E) 4) Shoreland of Quarantine Bay, Black Sea (44°36'41.4'' N, 33°30'6.2'' E) 5) Eastern point of the Curaçao island (12°12'32.3'' N, 68°48'58.8'' W) 6) Red Sea, Sharm El Sheikh (27°50'50.5'' N, 34°18'59.4'' E) 7) Kazachya Bay, Black Sea (44°34'18.8"N 33°24'40.2"E) 8) Curaçao island (12°12'32.3'' N, 68°48'58.8'' W)
Access & import/export	Habitats were accessed via a research vessel (locations 1, 2, 3), a car (locations 4-7), and a diving boat (lication 8). No permissions were required for sampling in the selected sampling sites.
Disturbance	No disturbances to the sites were caused; we sampled a small amount of water and surface sediment from marine habitats.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	Ten clonal cultures of protists were isolated from marine habitats
Authentication	Phase and DIC contrast light microscopy and 18S rRNA gene sequencing was used for authentication.

Mycoplasma contamination	This is not relevant to protist cell culture.
Commonly misidentified lines (See ICLAC register)	This is not relevant to protist cell culture.

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	The study did not involve laboratory animals.
Wild animals	The study did not involve wild animals (or any animals).
Reporting on sex	This is not relevant to protist cell culture.
Field-collected samples	Cultures of predatiry protists were established by isolating cells with a glass micropipette. Cultures were maintained at room temperature and at +4C. Cultures were propagateed using the kinetoplastid protist Procryptobia sorokini B-69 as prey. The kinetoplastid was grown in marine Schmalz-Pratt's medium or artificial marine medium (RS-R11040, Red Sea) and preyed upon Pseudomonas fluorescens.
Ethics oversight	No ethical approval was required. The organisms described here are novel eukaryotic microbes (protists) that feed on other protists and pose no risk.

Note that full information on the approval of the study protocol must also be provided in the manuscript.