Genomics of Preaxostyla Flagellates Illuminates Evolutionary Transitions and the Path Towards Mitochondrial Loss

Lukáš V. F. Novák*[1,2], Sebastian C. Treitli*[1], Jan Pyrih[1], Paweł Hałakuc[3], Shweta V. Pipaliya[4,5], Vojtěch Vacek[1], Ondřej Brzoň[1], Petr Soukal[1], Laura Eme[2], Joel B. Dacks[4,6], Anna Karnkowska[3], Marek Eliáš[7] & Vladimír Hampl[1$]

[1] Charles University, Faculty of Science, Department of Parasitology, BIOCEV, Vestec, Czech Republic.

[2] Ecology, Systematics, and Evolution Unit, Université Paris-Saclay, CNRS, Orsay, France.

[3] Institute of Evolutionary Biology, Faculty of Biology, Biological and Chemical Research Centre, University of Warsaw, Poland.

[4] Division of Infectious Diseases, Department of Medicine, University of Alberta, Edmonton, Canada.

[5] School of Life Sciences, Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland; Swiss Institute of Bioinformatics, Lausanne, Switzerland.

[6] Institute of Parasitology, Biology Centre, Czech Academy of Sciences, České Budějovice, Czechia.

[7] University of Ostrava, Faculty of Science, Department of Biology and Ecology, Ostrava, Czech Republic.

* Contributed equally

[$] Corresponding author

ABSTRACT

Until recently, mitochondria were considered essential organelles impossible to truly lose in a lineage. This view changed in 2016, with the report that the oxymonad *Monocercomonoides exilis*, was the first known eukaryote without any mitochondrion. Questions remain, however, about whether this extends to the entire lineage and how this transition took place. Oxymonadida are a group of gut endobionts of insects, reptiles, and mammals. They are housed in the Preaxostyla (Metamonada), a protistan group that also contains free-living flagellates of genera *Trimastix* and *Paratrimastix*. These latter two taxa harbor conspicuous mitochondrion-related organelles (MROs), while no mitochondria were reported for any oxymonad. Here we report genomic data sets of two Preaxostyla representatives, the free-living *Paratrimastix pyriformis* and the oxymonad *Blattamonas nauphoetae*. We note that *P. pyriformis* possesses a set of unique or ancestral features among metamonads or eukaryotes, e.g., *p*-cresol synthesis, UFMylation system, NAD^+ synthesis, selenium volatilization, or mercury methylation, demonstrating the biochemical versatility of this protist lineage. We performed thorough comparisons among all available genomic and transcriptomic data of Preaxostyla to corroborate both the absence of MRO in Oxymonadida and the nature of MROs present in other Preaxostyla and to decipher the evolutionary transition towards amitochondriality and endobiosis. Our results provide insights into the metabolic and endomembrane evolution, but most strikingly the data confirm the complete loss of mitochondria and every protein that has ever participated in the mitochondrion function for all three oxymonad species (*M. exilis*, *B. nauphoetae*, and *Streblomastix strix*) extending the amitochondriate status to the whole Oxymonadida.

INTRODUCTION

Multiple eukaryotic lineages have adapted to low-oxygen and/or endobiotic lifestyles by modifying their mitochondria into a wide range of mitochondrion-related organelle (MRO) types via the process of reductive and adaptive evolution (Roger et al. 2017). The most radically modified MROs are traditionally categorized as hydrogenosomes, producing ATP by extended glycolysis, and as mitosomes with no role in energy metabolism (Müller et al. 2012). However, exploration of a broader diversity of MRO-containing lineages makes it clear that such categorization is artificial, and MROs of various organisms form a functional continuum (Stairs et al. 2015; Klinger et al. 2016; Leger et al. 2017; Brännström et al. 2022).

The extreme outcome of mitochondrial reductive evolution is the complete loss of the organelle, but so far only one organism has been conclusively shown to have reached this point, the chinchilla gut symbiont *Monocercomonoides exilis* (Oxymonadida, Preaxostyla, Metamonada). A genomic project has thoroughly corroborated the amitochondriate status of *M. exilis*, in which it failed to identify any mitochondrion-associated genes while showing multiple other eukaryotic cellular systems to be well represented (Karnkowska et al. 2016; Karnkowska et al. 2019; Treitli et al. 2021). The existence of such an organism implies that mitochondria are not necessary for the thriving of complex eukaryotic organisms, which also has important bearings to our thinking about the origin of eukaryotes (Hampl et al. 2018).

Oxymonadida contain approximately 140 species of morphologically divergent and diverse flagellates exclusively inhabiting digestive tracts of metazoans, of which none has been shown to possess a mitochondrion by cytological investigations (Hampl 2017). It is, therefore, possible that the entire Oxymonadida group is an amitochondriate lineage, and that the mitochondrion was lost in its common ancestor. Oxymonads belong to Preaxostyla, one of the five primary clades of the phylum Metamonada consisting of exclusively anaerobes or microaerophiles; the other clades being represented by Fornicata (e.g., *Giardia intestinalis*), Parabasalia (e.g., *Trichomonas vaginalis*), barthelonids (Yazaki et al. 2020) and Anaeramoebidae (e.g., *Anaeramoeba flamelloides*; Stairs et al. 2021). The two additional known branches of Preaxostyla, classified as the genera *Trimastix* and *Paratrimastix*, split out at two successive points off the trunk leading to Oxymonadida. They are both comprised of free-living, bacterivorous flagellates exhibiting

a typical "excavate" morphology and ultrastructure and thriving in low oxygen environments (Zhang et al. 2015).

Paratrimastix pyriformis has been shown to possess an MRO morphologically resembling a hydrogenosome. However, it is likely not involved in the ATP-generating extended glycolysis but plays a role in one-carbon metabolism of the cell (Hampl et al. 2008; Zubáčová et al. 2013; Zítek et al. 2022). Putative MROs have also been observed in electron microscopy studies of a *Trimastix* representative, *T. marina* (Zhang et al. 2015), and their nature was illuminated by transcriptomic data (Leger et al. 2017). As a group with at least two MRO-bearing lineages and an amitochondriate species/clade nested within, the Preaxostyla clade provides a promising model system to study causes, conditions, and consequences of the loss of this ubiquitous cellular compartment.

To start answering questions about the timing and circumstances of this single evolutionary experiment in which the mitochondrion was lost from the cell, a denser sampling of "omics", functional and microscopic data from Preaxostyla is needed. Currently, the available data encompasses a genome draft of *M. exilis*, transcriptomes of *P. pyriformis* and *T. marina* with variable completeness, and a fragmentary single-cell genome assembly of an oxymonad *Streblomastix strix*. As a key step towards this end, we present high-quality genomic assemblies for another oxymonad, *Blattamonas nauphoetae*, and for the free-living *P. pyriformis*. The former has enabled us to test the hypothesis of amitochondriality as a shared feature of oxymonads in general, while the latter provided deeper insights into the metabolic capabilities of free-living Preaxostyla represented by *P. pyriformis*. The distribution of mitochondrion hallmark proteins and comparisons of gene repertoires and metabolic functions among five Preaxostyla species of various ecology and MRO status were used to illuminate the adaptations connected to amitochondriality and the origin of endobiotic lifestyle within the group.

RESULTS AND DISCUSSION

De novo genome assemblies are highly contiguous and complete

We employed a combination of Oxford Nanopore and Illumina technologies to obtain genome assemblies of two species of Preaxostyla. As both species are grown in polyxenic cultures where the eukaryotes represent a minority of cells, we employed multiple rounds of decontaminations. Prior to the genomic DNA isolation, enrichment of the sample for the eukaryotic cells was achieved using filtration, and after sequencing the data were carefully decontaminated bioinformatically, resulting in two highly contiguous eukaryotic genome assemblies (see Materials and Methods). The basic characteristics of these genomic assemblies and their comparison to the previously published assemblies of Preaxostyla taxa are given in Table 1.

The *B. nauphoetae* genome was assembled into 879 contigs spanning 88,537,989 bp, with an N50 = 199,589 bp and a GC content of 44.96%. Automatic and manual gene prediction resulted in 25,221 predicted protein-coding genes. Using BUSCO v3 (Simão et al. 2015) with the eukaryota_odb9 dataset, the genome completeness was estimated to be 76.6%. The *P. pyriformis* genome was assembled into 633 scaffolds spanning 56,627,582 bp, with an N50 = 276,605 bp and a GC content of 60.94%. Manual and automatic gene prediction resulted in 13,466 predicted protein-coding genes. Using the same method as for the *B. nauphoetae* genome, we estimated the genome completeness to be 82.1%, which is the highest value among all Preaxostyla genomes analyzed (Table 1). It should be noted that in the case of protist genomes, the estimates are not expected to reach 100% simply due to high divergence or true absence of some of the marker genes. For example, various Euglenozoa representatives, including *Trypanosoma brucei*, have completeness estimates in the range of 71-88% (Butenko et al. 2020) and metamonads in the range 60-91% (Salas-Leiva et al. 2021), both similar to the values for the two Preaxostyla genomes reported here.

Comparison of gene repertoires within Preaxostyla and contribution of LGT

The data generated in this study together with the previously published data from *M. exilis* (Karnkowska et al. 2019), *S. strix* (Treitli et al. 2019), and *T. marina* (Leger et al. 2017) were exploited to define the common genetic toolkit of Preaxostyla and specific toolkits of its subgroups. We note that the analyses described below rely on a slightly outdated version of the *M. exilis*

genome, since they had been initiated before a more complete version (Treitli et al. 2021) became available, and we also acknowledge the limitations of employing the inherently incomplete data from *S. strix* (single-cell genome assembly) and *T. marina* (transcriptome assembly only). We created orthologous groups (OGs) from the protein-coding gene inventories of the five Preaxostyla species and 14 other metamonads (Supplementary file 1) and estimated the core gene set for Metamonada and its clade-specific increments in Preaxostyla and Oxymonadida (Fig. 1, Supplementary file 2). Please note that due to the incomplete nature of all data sets we use the term core gene set in a relaxed sense and included genes present in at least one representative of all main subclades of the taxon (e.g., Fornicata, Parabasalia, and Preaxostyla in the case of Metamonada core gene set), see Material and Methods for details.

Each core set can in principle be composed of three different categories of genes: 1) genes inherited by the respective group from its earlier eukaryotic ancestors outside Metamonada, i.e., ancestral eukaryotic genes; 2) genes gained by lateral gene transfer (LGT); and 3) group-specific genes that originated in the respective group, i.e., truly novel genes or extremely diverged descendants of ancient genes, the homology of which was not discerned. To roughly distinguish between these categories within each ancestral set, we have searched for protein homologs (e-value < 10^{-15}) in the NCBI non-redundant protein sequence database and in the EukProt database (Richter et al. 2022), manually investigated phylogenies of proteins with prokaryotic hits, and in parallel employed an LGT detection pipeline. The latter revealed hundreds of putative LGT candidate genes in Preaxostyla (Supplementary file 3), which probably originated from at least 190 LGT events involving prokaryotic donors and affecting different branches of the Metamonada phylogeny (Fig. 1). Of these, 19, 15, 19, 1, and 3 LGT events were mapped to the common ancestor of Metamonada, the Fornicata+Preaxostyla clade, Preaxostyla, the *P. pyriformis*+Oxymonadida clade, and Oxymonadida, respectively, i.e., contributed to the core gene sets of these clades. The numbers of the ancestral LGTs might be underestimated, as they require sufficient sequence similarity between extant proteins to form a clade in a phylogenetic tree and this condition was often not fulfilled.

Of the 1,399 OGs from the Metamonada core gene set, 1,398 have homologues outside Metamonada and just a single orthologous group, q2019831, did not have any significant hit from NCBI nr or EukProt database at e-value < 10^{-15}. Interestingly, this protein identified by

InterProScan as a member of Major Facilitator Superfamily of transporters is present only in the data set of the parabasalid *Tritrichomonas foetus*, the oxymonad *S. strix* (Streblo_st30458), and the fornicates *Chilomastix cuspidata* and *Dysnectes brevis*. Although it may represent an innovation arising in the common ancestor of Metamonada, its function is clearly not essential. We identified 19 LGT events that contributed to the ancestral gene set of Metamonada and further 15 LGT events specific for the Preaxostyla+Fornicata clade (resolved in a recent phylogenomic analysis and at odds with the traditional view placing Preaxostyla as a sister lineage of all other metamonads; Stairs et al. 2021). These include metabolic enzymes used in glycolysis, amino acid metabolism and other functionally unrelated pathways. Some of these cases, e.g., enzymes involved in the arginine deiminase pathway, have been investigated in detail elsewhere (Liapounova et al. 2006; Novák et al. 2016; Karnkowska et al. 2019), while others are proposed here for the first time (Supplementary file 3).

11,854 OGs of all investigated metamonads are present only in Preaxostyla and 516 of these are widespread in this group and represent a putative Preaxostyla specific increment to the Metamonada core gene set. 71 OGs seem to be the group-specific genes common for the Preaxostyla lineage or genes too divergent, as they have no close eukaryotic BLAST hits (e-value $< 10^{-15}$) in any other organism. The vast majority of these have vague annotations and their function is simply not known. The remaining 445 OGs have homologs in eukaryotes outside Metamonada. We identified 19 LGT events that contributed to the core gene set of Preaxostyla (Fig. 1, Supplementary file 2 and 3).

A well-known example of LGT into Preaxostyla is represented by the proteins involved in the SUF pathway for FeS cluster synthesis (Vacek et al. 2018). Another potential case worth mentioning represents the divergent paralogue of the glycine cleavage system (GCS) L protein (GCS-L). Five GCS-L homologs were identified in the Preaxostyla investigated here: two in *P. pyriformis* and one in *T. marina*, *M. exilis*, and *B. nauphoetae* each. Given the fact that the GCS is a mitochondrion-specific pathway, the presence of the isolated L protein in the amitochondriate oxymonad *M. exilis* draws attention. The identified GCS-L sequences formed two clusters in the phylogenetic analysis (Supplementary file 4). *T. marina* and one of the *P. pyriformis* sequences branch together with sequences from Parabasalia and *Carpediemonas membranifera* with moderate statistical support, suggesting a common metamonad origin. These proteins were

previously hypothesized to function inside the MRO (Zubáčová et al. 2013). The other *P. pyriformis* sequence (PAPYR_1328) and the two oxymonad sequences branch with high statistical support together with sequences from two Archamoebae (*Mastigamoeba balamuthi* and *Pelomyxa schiedti*) and another eukaryotic anaerobe, *Breviata anathema* (Breviatea), and are here referred to as GCS-L2. Interestingly, in *M. balamuthi* it has been suggested that GCS-L2 protein could function outside the context of the glycine cleavage system (Nývltová et al. 2015), which might be the case also for the three related Preaxostyla sequences, potentially explaining the presence of GCS-L2 protein in the amitochondriate oxymonads. It should be noted that, besides GCS, GCS-L is known to be part of three other protein complexes, pyruvate dehydrogenase, branched-chain amino acid dehydrogenase, and 2-oxoglutarate dehydrogenase (Spalding and Prigge 2010), none of them being present in Preaxostyla. Outside of these typical roles or under specific conditions, GCS-L was shown to have a moonlighting proteolytic activity (Babady et al. 2007) or a diaphorase activity by which it oxidizes NADH using labile ferric iron (Petrat et al. 2003), nitric oxide (Igamberdiev et al. 2004), or ubiquinone (Xia et al. 2001). It is, therefore, possible that one of these moonlighting activities may represent the primary role of GCS-L2 in Preaxostyla. The well-supported relationship between the GCS-L2 sequences from Archamoebae and Preaxostyla may be explained by a eukaryote-to-eukaryote LGT.

Putting aside the 19 putative Preaxostyla-specific LGTs, the remaining 425 Preaxostyla-specific increments of the core gene set were likely inherited vertically from the eukaryotic common ancestor but lost or not detected in the other metamonad lineages, and as such they represent an interesting addition to the complexity of the common ancestor of Metamonada revealed thanks to the sampling of Preaxostyla genomes. To provide a single specific example, we focus on the small GTPase Rsg1 (Supplementary file 5). This protein was previously characterized in mammalian cells as a ciliogenesis factor (Agbu et al. 2018; Langousis et al. 2022), but its evolutionary origin and taxonomic occurrence have not been investigated. Having noticed its putative homolog in *P. pyriformis*, we searched publicly available sequence databases to demonstrate that Rsg1 orthologs indeed occur outside Metazoa, exhibiting a patchy distribution across distantly related protist lineages (Supplementary file 5). In Metamonada, Rsg1 was found in *P. pyriformis* and *T. marina*, while no orthologs were identified in oxymonads, fornicates, *Barthelona* sp., parabasalids and *Anaeramoeba* spp., apparently due to multiple independent losses of the Rsg1 gene. Notably, Rsg1 orthologs occur only in eukaryotes that can build cilia (Supplementary file 5),

indicating that the cilium-associated role defined for the mammalian Rsg1 is likely a general and ancestral functional attribute of the protein. The occurrence of Rsg1 in the free-living Preaxostyla thus points to specific differences in the functioning of their flagella as compared to other metamonads. Furthermore, considering the unsettled position of Metamonada with respect to the root of the eukaryote phylogeny (Derelle et al. 2015), the identification of Rsg1 in metamonads, along other phylogenetically diverse eukaryotes, strengthens the case that the protein is an ancestral component of the eukaryotic cell functioning already in the last eukaryote common ancestor (LECA).

Of the 9,513 OGs present only among oxymonads, 8,271 seem to be species-specific and 1,242 represent a putative oxymonad-specific increment to the Preaxostyla core gene set (Fig. 1, Supplementary file 2). Of these, 630 OGs seem to be oxymonad-specific or divergent, without close BLAST hits (e-value < 10^{-15}), and may thus represent novel genes evolved in this group providing new functionalities. Quite expectedly, 580 of them, although supported by transcriptome, have no or vague annotation and possibly represent genetic dark matter. The remaining 50 received annotations usually from KEGG or InterProScan, which were likely based on higher e-values than we used as a threshold for our BLAST and comprise a mixture of putative functions noticeably enriched for Fe-S cluster containing proteins and proteins involved in endomembrane transfer. The other 612 OGs have a homolog in eukaryotes outside Metamonada in the NCBI non-redundant protein sequence database or EukProt database (e-value < 10^{-15}). We identified three LGT events that contributed to the ancestral gene set of oxymonads and brought proteins with unclear or mutually unrelated functions. One of them is UTP—glucose-1-phosphate uridylyltransferase involved in glycogenesis and cell wall synthesis (Fig. 1, Supplementary file 3).

Based on orthology grouping, most of the genetic innovation is observed on the terminal branches of Preaxostyla rather than in the ancestral nodes. Putting aside the highly incomplete transcriptome of *T. marina* and the incomplete and fragmented genome of *S. strix*, we see that the number of species-specific singletons or OGs varies between 906 in *P. pyriformis* to 2,493 in *B. nauphoetae*. Also, 46 LGT events were mapped to the terminal branches and 2 to internal nodes of Preaxostyla, suggesting that the acquisition of genes is an ongoing phenomenon, and it might be adaptive to particular lifestyles of the species.

***P. pyriformis* possesses unexpected metabolic capacities**

The relatively complete and contiguous genomic and transcriptomic datasets for *P. pyriformis* allowed to compose a more complete and more accurate inventory of the *P. pyriformis* genetic toolkit than the previous studies (Hampl et al. 2008; Zubáčová et al. 2013). Given the fact that this species represents the first free-living representative of the Preaxostyla lineage studied at the genome level, it is not that surprising that our exploration of the *P. pyriformis* genome data revealed a number of unusual or unexpected features of its metabolism, some of which are presented in detail below.

P. pyriformis encodes a complete pathway required for the biosynthesis of *p*-cresol from tyrosine (Supplementary file 6), which consists of three steps of the Ehrlich pathway (Hazelwood et al. 2008) converting tyrosine to 4-hydroxyphenyl-acetate and the final step catalyzed by a fusion protein comprised of 4-hydroxyphenylacetate decarboxylase (HPAD) and its activating enzyme (HPAD-AE). The *P. pyriformis* HPAD-AE protein is closely related to its homolog in the free-living archamoebid *M. balamuthi* (Supplementary file 4), the only eukaryote reported so far to be able to produce *p*-cresol (Nývltová et al. 2017). Interestingly, we identified additional eukaryotic homologs of this fusion protein in the free-living fornicates *Kipferlia bialata* and *Trepomonas* sp. This distribution may be explained by LGT from prokaryotes followed by one or more LGTs between the eukaryotic lineages facilitated by a shared anaerobic habitat. Interestingly, the close endobiotic relatives of these taxa (oxymonads, *Entamoeba histolytica*, parasitic diplomonads such as *Giardia intestinalis* or *Spironucleus salmonicida*) lack this protein. Biosynthesis of *p*-cresol by *M. balamuthi* was hypothesized to confer a competitive advantage against prokaryotes inhabiting the same environment, because this compound inhibits the growth of certain prokaryotes, while no such inhibition was observed in *M. balamuthi* (Nývltová et al. 2017).

The presence of demethylmenaquinone methyltransferase / 2-methoxy-6-polyprenyl-1,4-benzoquinol methylase (UbiE) in *P. pyriformis* (Supplementary file 6) is surprising, since the organism lacks any other enzymes involved in the ubiquinone and menaquinone biosynthesis. Interestingly, this enzyme has been shown to have an alternative function, a capability to mediate the synthesis of volatile methylated selenium (Se) compounds, namely dimethyl selenide and dimethyl diselenide, in bacteria and plants (Swearingen et al. 2006; Zhou et al. 2009). Selenium is a micronutrient that becomes toxic in high concentrations. Biological Se volatilization is an

important detoxification process that converts inorganic Se compounds into volatile organic compounds. We propose that Se volatilization might be the physiological function of the *P. pyriformis* UbiE. This enzyme is another noteworthy case of prokaryotic LGT within the *P. pyriformis* genome, most likely from a verrucomicrobial source (Supplementary file 4). Interestingly, UbiE homologs occur also in some additional metamonads, including the oxymonad *B. nauphoetae* and certain fornicates (Supplementary file 4), but these seem to represent independent LGT events from different bacterial sources. The recurrent acquisition of UbiE by different eukaryotic anaerobes attests to the physiological significance of the enzyme.

P. pyriformis has a complete pathway for biosynthesis of NAD^+ from L-aspartate and dihydroxyacetone phosphate consisting of L-aspartate oxidase (NadB), a nicotinate-nucleotide pyrophosphorylase-quinolinate synthase fusion protein (NadC-NadA), nicotinate-mononucleotide adenylyltransferase (NadD), DNA-binding transcriptional repressor/NMN adenylyltransferase (NadR), and NAD^+ synthase (NadE); at least some of these enzymes are also present in *T. marina* (Supplementary file 6). The presence of this pathway in anaerobes is surprising because the standard NadB reaction involves molecular oxygen (O_2). However, NadB can alternatively use fumarate instead of oxygen (Sakuraba et al. 2002). If this is the case also in *P. pyriformis*, it would explain the presence of fumarate hydratase (FH) in this organism: fumarate produced by FH from malate could serve as an electron acceptor for NadB. Phylogenetic analyses of NadA, NadB, NadC, and NadR show the Preaxostyla sequences branch separately from other eukaryotic sequences and have an affinity to Chloroflexi (NadA, NadC), and Lentisphaerae (NadB) hinting at LGT origin from prokaryotes. In the case of NadB, this LGT is likely shared with *M. balamuthi*, which occupies a similar position inside Lentisphaerae. Only the Preaxostyla sequences of NadD and NadE branch together with other eukaryotes (Supplementary file 4).

Uniquely among eukaryotes, *P. pyriformis* may possess enzymes for mercury (Hg) methylation. Production of highly toxic methyl mercury is known in some Bacteria and Archaea (Podar et al. 2015; Cooper et al. 2020). The exact reason why these microorganisms methylate Hg is unknown, however, it is widely accepted that it is not a detoxification process as these are not less susceptible to Hg (Gilmour et al. 2011). In prokaryotes, the Hg methylation is performed by two enzymes HgcA and HgcB encoded by two independent genes, although they are often found adjacent to each other in a single operon (Christensen et al. 2016). In contrast, *P. pyriformis* contains a single

fusion protein HgcAB (Supplementary file 6). As the order of the two parts reflects the arrangement of the genes in bacteria, both were likely acquired *en bloc* by lateral transfer of a whole operon.

The structure of the prokaryotic HgcAB complex was recently determined (Cooper et al. 2020). Comparison of PAPYR_7512 to HgcA and HgcB from *Pseudodesulfovibrio mercurii* and a selection of other prokaryotes (Fig. 2) revealed remarkable similarity and conservation in the HgcA part, which is predicted to contain a corrinoid binding domain (CMD) including the well-conserved corrinoid binding alpha-helix and cobalamin binding residues (Thr60, Thr66, Val91, Cys93, and Ala153 in *P. mercurii*), an N-terminal transmembrane domain (with four transmembrane domains in contrast to five in *P. mercurii*), and a conserved motif (NVWCAAGKG) at the positions 97 – 105. The *P. mercurii* HgcB shares with the *P. pyriformis* sequence the same twofold pseudosymmetry as the bacterial 4Fe-4S ferredoxin and two strictly conserved CxxCxxCxxxCP 4Fe-4S binding motifs, conserved C-terminal cysteine residues, and a cysteine residue (Cys73 in *P. mercurii*) at the beginning of an alpha-helix. Overall PAPYR_7512 shows high similarity to HgcAB complex and contains all important residues necessary for methylation of Hg, a function that has not been reported in any other eukaryote (Fig. 2).

Another intriguing gene fusion was found between genes for two histidine-processing enzymes: histidine ammonia-lyase and histidyl-tRNA synthase (HAL-HARS). HAL is responsible for the first step in the catabolic pathway converting histidine to glutamine. Histidyl-tRNA synthases with an N-terminal extension homologous to HAL have been previously found in land plants and green algae (Saga et al. 2020). However, the plant HAL-HARS sequences lack a crucial catalytic site of the HAL enzyme (MIO cofactor formed from three conserved amino acids) and are therefore considered non-functional. The *P. pyriformis* sequence (Supplementary file 6), on the other hand, has these three amino acids conserved. The existence of this fusion of the two enzymes that both have histidine as substrate in two distant eukaryotic lineages suggests there might be a functional connection between histidine ammonia-lyase and histidyl-tRNA synthase, which deserves further investigation.

Altogether, the gene inventory of *P. pyriformis* demonstrates that this protist possesses several metabolic capabilities which are very rare or so far unseen in eukaryotes and that LGT is a common means of their acquisition.

Membrane transporter complement may reflect adaptation to an endobiotic lifestyle in oxymonads

Though most attention is paid to the divide between MRO-possessing and putatively MRO-lacking Preaxostyla, at least two other transitions can be examined using this set of species. The first is from a free-living to a symbiotic lifestyle. As an assessment of this, we examined proteins responsible for the transport of metabolites and other chemicals across the plasma membrane and other cell membranes, as changes in their repertoire and capabilities may reflect evolutionary transitions in both metabolic capabilities and the lifestyle of the organism. We searched for a broad spectrum of transmembrane transporters (except for transporters of MROs) using homology-based methods in order to compare the repertoire of functional types as well as the number of paralogues between the five species of Preaxostyla (Supplementary file 7).

Just like in *T. vaginalis*, the most gene-rich group of membrane transporters identified in Preaxostyla is the ATP-binding cassette (ABC) superfamily represented by MRP and pATPase families. Altogether, representatives of 19 transporter families have been identified in Preaxostyla. All of them are present in *P. pyriformis* and *T. marina*, while four families (PotE, SPNS, RFC, and TauE) are missing in all three oxymonad species investigated. On the other hand, transporters of nucleosides, sugars, amino acids, choline, and phospholipids have consistently higher numbers of paralogues in the genomic datasets of the oxymonads *M. exilis* and *B. nauphoetae* than in *P. pyriformis*. These two observations parallel findings of functional domain loss and expansion by gene duplication of transporter families in Microsporidia (Nakjang et al. 2013), possibly hinting at a broader evolutionary pattern at the transition to the endobiotic lifestyle.

Evidence for the Golgi complex in all Preaxostyla is consistent with the cisternal adhesion model of Golgi stacking

Another evolutionary transition addressable by our results is the presence of a morphologically identifiable Golgi body in *P. pyriformis* + *T. marina* versus the lack in oxymonads (O'Kelly et al. 1999; Zhang et al. 2015; Treitli et al. 2018). Nonetheless, a substantial complement of proteins associated with Golgi membrane-trafficking and transport has been previously reported, confirming at least the genomic signal for a Golgi body, in the oxymonad *M. exilis* (Karnkowska et al. 2016; Karnkowska et al. 2019). Therefore, our dataset represents one of the tightest samplings

to date where genomic data is available for the closest known species on both sides of the divide between stacked and unstacked Golgi morphology. Consequently, we searched in our datasets to assess the Golgi complement in the additional Preaxostyla representatives, particularly to see whether the complement was more extensive in the organisms possessing stacked Golgi bodies.

The scope of our examination included proteins involved in vesicle formation, vesicle fusion, and the golgin proteins implicated in Golgi structure (Boncompain and Weigel 2018; Ahat et al. 2019; Kulkarni-Gosavi et al. 2019; Li et al. 2019; Park et al. 2021; Aridor 2022). We observed near complete complements of the COPI, AP1, AP3, AP4, and Retromer vesicle coats, as well as the GARP complex, Trs120, and syntaxins 5 and 16 (Fig. 3, Supplementary file 6). We also noted at least one golgin protein in each of the organisms. Indeed, we observed additional paralogues of the vesicle trafficking machinery (e.g., AP1, Retromer, GARP/EARP) in oxymonads compared to *P. pyriformis* and *T. marina*, (Fig. 3, Supplementary file 6). These data, together with previously published observations (Vargová et al. 2021), are indicative of Golgi bodies with multiple anterograde and retrograde pathways entering and exiting the organelle present in all Preaxostyla species.

We did observe two clear differences in the sets of Golgi-associated proteins between the stacked and unstacked possessing organisms. Firstly, *P. pyriformis* encodes seven of the eight Conserved Oligomeric Golgi (COG) complex proteins, while only a sparse representation of the COG complex was seen in the oxymonads. Secondly, the golgin CASP was found in both *P. pyriformis* and *T. marina*, but in none of the oxymonad genomes. Golgin 84 was also found in *P. pyriformis* alone. This marks the first report of CASP or golgin 84 from a metamonad (Barlow et al. 2018) suggesting independent losses of these proteins in the Oxymonadida, Parabasalia, and Fornicata lineages. While caution is warranted when reporting the absence of any given single protein from any given genome, our observations do show a greater number of encoded Golgi-structure implicated proteins in the stacked-Golgi possessing lineages than in the oxymonads (4, 7 vs 1, 4, 3 respectively; Fig. 3). Though expression levels would need to be taken into account, this observation is nonetheless consistent with the "cisternal adhesion" model (Lee et al. 2014), i.e., that it is the amount of adhesive golgin-type proteins that regulate stacking rather than the identity of any given Golgi-stacking protein (Barlow et al. 2018).

Carbon and energy metabolism adaptation in the common Preaxostyla ancestor

The most prominent evolutionary transition addressed in our dataset of course is that of the MROs. Before examining this directly, we delved into the associated cellular metabolism in the preaxostylan species in order to establish a baseline metabolic context for function of the different organelles. Anaerobically living eukaryotes often use a modified glycolytic pathway that incorporates pyrophosphate (PPi)-dependent instead of ATP-dependent enzymes. This alternative form of glycolysis can produce three ATP molecules and five high-energy phosphate bonds per one glucose molecule, in comparison to only two ATP and two high-energy phosphate bonds in the "classical" pathway known from aerobes (Mertens 1993). Multiple alternative glycolytic enzymes have been previously identified in *P. pyriformis*, *M. exilis*, and *S. strix* (Liapounova et al. 2006; Slamovits and Keeling 2006; Stechmann et al. 2006).

Here, we report the identification of both phosphofructokinase (PFK) and diphosphate-fructose-6-phosphate 1-phosphotransferase (PFP) catalyzing the 3rd step of glycolysis in *P. pyriformis* and *T. marina*, while only PFP is present in the three oxymonad species (Fig. 4, Supplementary file 6), suggesting that the alternative enzyme PFP has been acquired by a common ancestor of Preaxostyla. Both pyruvate kinase (PK) and pyruvate-phosphate dikinase (PPDK) catalyzing the last step have been identified in all five species studied. In the case of the 4th and 8th steps, alternative versions of the enzymes catalyzing these reactions, namely Class II fructose biphosphate aldolase (FBA) and 2,3-bisphosphoglycerate independent phosphoglycerate mutase (iPGM), respectively, were identified in all five studied species, while no "classical" enzymes for these reactions have been found. This further emphasizes that the glycolytic pathway in Preaxostyla is an evolutionary mosaic composed of enzymes of different origins (Fig. 4), likely acquired in adaptation to the anaerobic lifestyle.

Many eukaryotic anaerobes, including Preaxostyla, generate ATP using the extended glycolysis pathway, which produces acetate, CO_2, and H_2 from pyruvate while performing substrate-level phosphorylation of ADP to ATP (Lindmark and Müller 1973). The pathway uses pyruvate as a substrate, which is either directly sourced from the cytosolic glycolysis or produced by decarboxylation of malate through the activity of malic enzyme (ME). This enzyme has been identified in all five Preaxostyla species. All Preaxostyla apparently rely on oxidative decarboxylation of pyruvate to acetyl coenzyme A (acetyl-CoA) and CO_2 in a reaction catalyzed

by pyruvate:ferredoxin oxidoreductase (PFO; Williams et al. 1987), owing to the identification of three to six PFO isoforms in each species analyzed (Fig. 4, Supplementary file 6). On the other hand, none of the species encodes any of the two other alternative enzymes mediating the conversion of pyruvate to acetyl-CoA in various other eukaryotic anaerobes, pyruvate:NADP+ oxidoreductase (PNO) and pyruvate formate lyase (PFL). Both the decarboxylation of malate by ME and of pyruvate by PFO are oxidative processes that release electrons, producing NADH and reduced ferredoxin, respectively.

The final fate of the electrons carried by ferredoxin often lies in the reduction of protons to molecular hydrogen through the activity of [FeFe] hydrogenases (HydA; Payne et al. 1993). In addition to the "simple" hydrogenases, which are present in all species of Preaxostyla, [FeFe] hydrogenases with N-terminal homology to the NuoG subunit of NADH-quinone oxidoreductase and a C-terminal homology to NADPH-dependent sulfite reductase (CysJ), were identified in *T. marina* and *P. pyriformis* (Fig. 5). Similar "fused" hydrogenases (Fig. 5) have been previously reported in other eukaryotic anaerobes, including *T. vaginalis* (Tachezy and Doležal 2007), the breviate *Pygsuia biforma* (Stairs et al. 2014), and the jakobid *Stygiella incarcerata* (Leger et al. 2016). Although they do not belong to the group of A3 trimeric hydrogenases (Greening et al. 2016) known to be capable of NADH oxidation via electron confurcation (Schut and Adams 2009), they were hypothesized to catalyze NAD(P)H-dependent formation of H_2 (Tachezy and Doležal 2007). However, experimental evidence is still missing. Ferredoxin (Fd) serves as an electron intermediate between PFO and HydA. Both Fd and unique flavodoxin-ferredoxin fusion proteins (Fld-Fd) were identified in all five species (Supplementary file 6).

The second reaction of the extended glycolysis, which yields ATP, acetate, and CoA, can be catalyzed either by a two-enzyme system consisting of acetate:succinate CoA-transferase (ASCT) and succinyl CoA synthetase (SCS) like in *T. vaginalis*, or by a single enzyme acetyl-CoA synthetase (ACS) like in *G. intestinalis*. All five Preaxostyla species contain ACS, while neither ASCT nor SCS were identified. ACS has a complicated evolutionary history in Metamonada characterized by multiple LGT events and gene losses (Leger et al. 2017; Yazaki et al. 2020; Vargová et al. 2022). The majority of ACS homologs in Metamonada are predicted to function in the cytosol, with the only exception of the ACS2 isoform from *S. salmonicida* which functions in

the MRO (Jerlström-Hultqvist et al. 2013; Leger et al. 2017). Phylogenetic analysis of Preaxostyla ACSs (Supplementary file 4) shows four unrelated clades, none in close relationship to the *S. salmonicida* MRO homolog, suggesting a cytosolic localization of these enzymes in Preaxostyla.

Amino acid metabolism is more complete in *P. pyriformis* and *T. marina* than in oxymonads

While some aspects of the core organellar metabolism seemed similar across Preaxostyla, some clear differences were observed, particularly in amino acid metabolism. *P. pyriformis* predicted proteome contains 53 enzymes putatively involved in amino acid metabolism and seven more enzymes putatively involved in folate metabolism and one-carbon pool, which are closely connected to the amino acid metabolism (Ducker and Rabinowitz 2017). This is a higher number than in *T. vaginalis*, which has the most complex predicted amino acid metabolism among metamonads studied so far, containing 39 enzymes (Carlton et al. 2007). The *T. marina* predicted proteome contains 31 enzymes of amino acid metabolism and six of folate and one-carbon metabolism, but the lower number compared to *P. pyriformis* may at least partly reflect the different nature of the sequence data available (transcriptome assembly only). The number of enzymes involved in the amino acid metabolism in oxymonads is 31 in *M. exilis*, 24 in *B. nauphoetae*, and 22 in *S. strix* (Supplementary file 6).

Hypothetical metabolic maps of the amino acid metabolism were reconstructed based on the metabolic maps in the KEGG database (Kanehisa et al. 2014) and catalytic activities of enzymes reported from other metamonads (Carlton et al. 2007; Xu et al. 2014). The reconstructed metabolism (Supplementary figs. 1-3) shows a possibility for *de novo* biosynthesis from glycolytic intermediates of at least five protein-building amino acids in *P. pyriformis* (cysteine, serine, glycine, threonine, and selenocysteine), only one in *T. marina* (cysteine), which is probably caused by incomplete data, and three in *M. exilis* (cysteine, serine, and selenocysteine).

Methionine (in the form of S-adenosylmethionine, SAM) consumed in reactions catalyzed by SAM-dependent methyltransferases can be regenerated in *P. pyriformis* and *T. marina* thanks to the methionine cycle, which is absent in oxymonads. The selenocysteine biosynthesis pathway present in *P. pyriformis*, *T. marina*, and *M. exilis* is notable, as the capacity to synthesize this amino acid has been reported only in *S. salmonicida* among other metamonads studied so far (Xu et al. 2014).

Like many other metamonads (Novák et al. 2016), *M. exilis* has been suggested to utilize arginine for ATP production via the arginine deiminase pathway consisting of three enzymes: arginine deiminase (ADI), carbamate kinase (CK), and ornithine transcarbamylase (OTC). This important metabolic capability has been probably formed in the common Metamonada ancestor by the acquisition of genes for ADI and OTC via LGT (Supplementary file 2). Here we show the presence of the complete arginine deiminase pathway also in *T. marina* and *B. nauphoetae*, while *P. pyriformis* and *S. strix* lack ADI and CK, respectively. This suggests that ATP production via arginine catabolism is widespread among Preaxostyla, and it is present in both free-living and endobiotic representatives. Other amino acids can be used in energy metabolism as well: cysteine, serine, and tryptophan can be converted to pyruvate, while methionine can be converted to α-keto-butyrate. Both products can be used by PFO and ACS to generate ATP (Anderson and Loftus 2005).

The presence of a complete glycine cleavage system (GCS) in *P. pyriformis* and *T. marina* is connected to the presence of complete folate and methionine cycles (Zítek et al. 2022). The methyl residue liberated from glycine by the activity of GCS enters the connected folate and methionine cycles and can be later utilized in a multitude of metabolic pathways requiring one-carbon units, e.g., remethylation of homocysteine to form the amino acid methionine via MetH (Ducker and Rabinowitz 2017). The transsulfuration pathway, associated with the folate and methionine cycles in mammals (Stipanuk 1986), was not found in any Preaxostyla species. It is possible that the loss of the folate and methionine cycles in oxymonads is related to the loss of mitochondria because it is connected with mitochondria via GCS, an exclusively mitochondrial pathway, which was indeed found to be localized in the MRO of *P. pyriformis* (Zubáčová et al. 2013; Zítek et al. 2022).

Oxymonads predominantly use 4Fe-4S clusters in their iron-sulfur clusters containing proteins

Fe-S clusters are one of the most ancient and versatile inorganic cofactors and are present in virtually all living organisms. Thanks to their ability to delocalize electrons over the Fe-S bond (Noodleman and Case 1992; Glaser et al. 2000) they are ideally suited for mediating electron transfer and therefore are part of important biological pathways, such as the respiratory chain, photosynthesis, DNA metabolism, ribosome biogenesis, and many others. In a typical eukaryotic cell, Fe-S clusters are synthesized by a combination of the mitochondrial ISC and the cytosolic

CIA pathway, but members of Preaxostyla uniquely combine the CIA pathway with the SUF pathway secondarily acquired from bacteria to perform the same task (Vacek et al. 2018; Braymer et al. 2021). Although functional details of the Fe-S cluster biogenesis in this group are mostly unknown, it was shown to run wholly in the cytosol (Zítek et al. 2022), and this major shift was likely a preadaptation for the loss of the mitochondrion in the lineage leading to *M. exilis* (Karnkowska et al. 2016).

To assess how much the change of the Fe-S cluster assembly pathway affected the inventory of Fe-S cluster-containing proteins in Preaxostyla, we predicted these proteins from *in silico* proteomes of *P. pyriformis*, *B. nauphoetae, T. marina,* and *S. strix* and compared them with the previously predicted (Karnkowska et al. 2019) Fe-S proteins of *M. exilis* (Supplementary file 8). The numbers of Fe-S cluster-containing proteins identified in individual species varied from 48 in *T. marina* to 93 in *P. pyriformis* and are thus not decreased in comparison to other heterotrophic eukaryotes (Andreini et al. 2016). Predicted Fe-S proteins fell into distinct 164 OGs. Numbers of these OGs shared among Preaxostyla species are shown in Supplementary Fig. S4.

Most of the predicted Fe-S proteins in oxymonads contain 4Fe-4S clusters, with the exception of xanthine dehydrogenase (XDH), which contains both 2Fe-2S and 4Fe-4S clusters and is present in all Preaxostyla in two (*S. strix, B. nauphoetae, T. marina*) or three copies (*M. exilis* and *P. pyriformis*). The free-living *P. pyriformis* furthermore contains several proteins with 2Fe-2S clusters only, such as [FeFe] hydrogenases and 2Fe-2S ferredoxin. The higher number of 2Fe-2S proteins in *P. pyriformis* compared to the oxymonads may reflect the presence of the MRO in this organism.

The most widespread types of Fe-S cluster-containing proteins in Preaxostyla fall into expected functional groups acting in processes of extended glycolysis and electron transfer (PFO, HydA, ferredoxins, and flavodoxin-ferredoxin domains containing proteins), DNA replication and repair, transcription and translation (e.g., DNA and RNA polymerases, Rli1p), Fe-S cluster assembly itself (SufDSU, SufB, Nbp35, and NAR-1), nucleotide (XDH), and amino acid metabolism (L-serine dehydratase) etc. An exception represents a radical SAM enzyme (OG q2001674) that is homologous to a member of a three-gene system AmmeMemoRadiSam (Balaji and Aravind 2007; Burroughs and Aravind 2014; Burroughs et al. 2019). Notably, homologs of the other two members of this system are also present in all Preaxostyla and most metamonads, suggesting a functional

significance. Each AmmeMemoRadiSam protein of Preaxostyla is affiliated with homologs from other eukaryotes (Supplementary file 4). Given the fact that the function of these genes in prokaryotes is unclear, we do not know what function they could perform in metamonads.

A systematic search for proteins associated with the mitochondrion fails to identify convincing candidates in oxymonads

The uniqueness of Preaxostyla lies in the fact that it includes species harboring MROs, exemplified by *T. marina* and *P. pyriformis* (Hampl et al. 2008; Zubáčová et al. 2013; Leger et al. 2017), and at least one species lacking any form of a mitochondrion, i.e., *M. exilis* (Karnkowska et al. 2016). The hypothesis on the absence of the organelle in *M. exilis* was postulated based on the absence of mitochondrion-related proteins in its genome and transcriptome and led to the twist in the paradigm of mitochondria as ubiquitous among eukaryotes (Karnkowska et al. 2016). Obviously, the amitochondriate status of *M. exilis* is immediately falsifiable by finding any evidence of a putative organelle in this species. Careful inspection of the predicted proteomes of other oxymonads for the presence of mitochondria is an obvious next step that may further support, or weaken, this hypothesis.

We used several different methods to identify candidates for mitochondrial proteins (results for all searches are shown in Supplementary file 9 and summarized in Fig. 6). In all cases, proteins identified in the first step ("hits") were reciprocally searched against the MitoMiner (Smith et al. 2012; Smith and Robinson 2018) and NCBI databases. We considered proteins yielding results of reciprocal blasts as "candidate" mitochondrial proteins, which were further manually verified with extensive searches (see Materials and Methods). Proteins passing this third step were classified as validated mitochondrial proteins.

We searched all five investigated species for homologs of nuclear genome-encoded proteins typically associated with mitochondria or MROs in other eukaryotes. Initially, we used profile Hidden Markov Models (HMMs) to search for components of the mitochondrion protein import and maturation machinery, considered one of the most conserved mitochondrial features (Pfanner et al. 2019). Homology searches resulted in 22 (*T. marina*) up to 63 (*M. exilis*) hits, which were further evaluated by reciprocal searches against the MitoMiner and NCBI databases resulting in 57 candidates. Manual verification indicated that all the candidates recovered in oxymonad data sets

are false positives (Fig. 6, Supplementary file 9). For *P. pyriformis* and *T. marina*, the situation was different. Most of the previously identified components of the *P. pyriformis* mitochondrion protein import and maturation machinery (Zubáčová et al. 2013; Zítek et al. 2022) were found in the predicted proteomes: the β-barrel outer membrane translocases Tom40 and Sam50, the inner membrane translocase Tim17 and its associated protein Pam18, the α and β subunit of the mitochondrial processing peptidase (MPP), and the chaperone protein Cpn60 and Cpn10. Tim17 and Hsp70, were also identified in the *T. marina* dataset, corroborating previous findings (Leger et al. 2017). The successful identification of protein translocon components in these two species validated our approach.

MROs export or import ATP and other metabolites typically using transporters from the mitochondrial carrier family (MCF) or sporadically (Tsaousis et al. 2008; Major et al. 2017) by the bacterial-type (NTT-like) nucleotide transporters. We did not identify any homolog of genes encoding known mitochondrial metabolite transport proteins in any of the three oxymonads investigated. In contrast, the mitochondrial carrier (MCF) proteins were recovered in the number of four for each *P. pyriformis* and *T. marina* (Supplementary file 6).

The sensitive homology-based HMM searches were complemented by an extensive search for putative homologues of known mitochondrial proteins using a pipeline based on the MitoMiner database (Supplementary file 9), which was enriched with identified mitochondrial proteins of diverse anaerobic eukaryotes with MROs (see Materials and Methods). As already shown for *M. exilis*, the specificity of the pipeline in organisms with divergent mitochondria is low (Karnkowska et al. 2016). Many of the selected candidates were assigned as proteins that are obviously not mitochondrial (e.g., histones or ribosomal proteins) or belonging to very general GO categories (e.g., protein-binding). The search recovered a comparable number of proteins as candidates for functions in a putative mitochondrion in all investigated species, from 636 candidates in *S. strix* up to 1,025 candidates in *P. pyriformis*. While no reliable mitochondrial protein was found among the oxymonad candidates, for *P. pyriformis* and *T. marina* we identified a list of proteins previously known to be localized in their MROs as well as some new candidates, such as mitochondrial fission protein 1 (Fis1; Gene.12785::gnl|Trimastix_PCT|4043), which is a model TA protein mediating mitochondrial fission (van der Bliek et al. 2013). The phylogenetic pattern of dynamin-related proteins corroborates the finding of Karnkowska et al. (2019) showing a specific

Paratrimastix+Trimastix clade potentially mediating the division of MROs in these species (PAPYR_3413 and Gene.668::gnl|Trimastix_PCT|268, Supplementary file 4). In addition, both *T. marina* (Gene.3674::gnl|Trimastix_PCT|1191) and *P. pyriformis* (PAPYR_2826) but none of the oxymonads exhibit an ortholog of the mitochondrial outer membrane-anchored protein MIRO ("mitochondrial Rho"; Eberhardt et al. 2020) broadly conserved in eukaryotes but rare in MRO-containing taxa (Vlahou et al. 2011; Gentekaki et al. 2017); recently it was confirmed by proteomics to be MRO-associated in the latter species (Zítek et al. 2022).

As an alternative to homology searches, we also searched for several types of signature sequences typical for mitochondrion-targeted proteins. The matrix proteins of mitochondria and MROs are expected to contain characteristic N-terminal targeting signals (NTS) needed for the targeted import into mitochondrion-related organelles (Dolezal et al. 2006). As we have previously shown, even in the amitochondriate *M. exilis*, a fraction of proteins (though less than 1%) contains NTS recognized by the subcellular localization predicting tools (Karnkowska et al. 2016). It has been previously recognized that the presence of a predicted NTS by itself does not prove the targeting, as such amino acid sequences can also appear by chance (Lucattini et al. 2004). Here we used up-to-date tools for targeting signal prediction and identified from two hits in *S. strix* up to 248 hits in *B. nauphoteae* (Supplementary file 9). Oxymonad species contain a lower proportion of proteins with the putative mitochondrial NTS (always below 1%) than *Trimastix* and *Paratrimastix* and most of them gave no hits in MitoMiner and the NCBI nr database; the few candidates with good hits matched to proteins unlikely to function in mitochondria. If the NTS in the protein is functional and not formed by chance, it is expected that a NTS will be detected also in its orthologues from other species. With this assumption, we have focused on candidates of which more Preaxostyla orthologues were predicted to contain NTS. Only four proteins from oxymonads (ribosomal protein L21, L23a, and L34e, and tRNA-dihydrouridine synthase) fulfilled this criterion, none of them representing a reasonable mitochondrial protein in an organism that for sure lacks a mitochondrial genome. Based on those results, we assume that all NTS predictions on oxymonad proteins are false positives. In *P. pyriformis* and *T. marina* the proportion of proteins with predicted NTS is higher (1.5% and 2.7% respectively), which might reflect the presence of the organelle. Indeed, we identified among them, for example, aminoadipate-semialdehyde dehydrogenase and the GCS-H protein in *T. marina*, both previously suggested to localize to the MRO (Leger et al. 2017).

The outer mitochondrial membrane accommodates two special classes of proteins, tail-anchored (TA) proteins and mitochondrial β-barrel membrane proteins (MBOMPs), the former using specific C-terminal signals (Borgese et al. 2007; Denic 2012; Rada et al. 2019). We identified up to 475 TA hits in the predicted proteome of *S. strix*, and around 100 for the rest of the species, with only 25 for *T. marina* (Supplementary file 9). Depending on the species only five to twelve were considered as candidates with homologs in the MitoMiner database and the majority of these were Golgi apparatus/ER-related. A subset of the TA proteins corresponded to 22 OGs that were present in an oxymonad and at least one more species of Preaxostyla. These were mostly without MitoMiner hits and in most cases annotated as SNARE or other proteins involved in vesicle trafficking. This suggests that the TA prediction did produce true positives, but they all likely represent non-mitochondrial TA proteins. The only probable mitochondrial candidate is Fis1 identified in the *T. marina* predicted proteome.

Our search for proteins with MBOMP characteristics revealed six and 23 hits in *T. marina* and *P. pyriformis*, respectively, but only three of them have a hit in the MitoMiner database, including the *P. pyriformis* Tom40, the detection of which validates our approach. Among the 26 to 36 candidate proteins from each oxymonad genome, only four hit a protein in the MitoMiner database, none of the hits being a known MBOMP (Supplementary file 9). Similarly to the previous protein categories, only four OGs predicted as MBOMPs were present in an oxymonad and at least one other Preaxostyla species, indicating that the candidates were selected from the species pools nonspecifically and likely do not contain a putative new MBOMP. The only overlapping OGs represent clathrin heavy chain (two OGs), kelch-type beta propeller, and EF-hand domain-containing protein, which are again rather false positives than mitochondrial proteins.

Finally, given the limitations of the previous searches, namely, high false-positive rates, uncertainty about mitochondrial localization of proteins included in the MitoMiner database, and the inability to distinguish between the cytosolic and mitochondrial isoforms, we have used also a more curated approach and searched for orthologues of proteins from the experimentally well-established mitochondrial proteome of *T. brucei* (Panigrahi et al. 2009; Dean et al. 2017; Peikert et al. 2017). Reciprocal BLAST searches of the *T. brucei* proteins against the predicted proteins of Preaxostyla revealed ~200 putative groups of orthologues and those were investigated manually. Careful inspection of the raw localization data in *T. brucei* and protein phylogenies (Supplementary file 10)

rejected most cases by disputing either the mitochondrial localization in *T. brucei* or the orthology. The gene orthology was considered uncertain if *T. brucei* and Preaxostyla proteins were separated by prokaryotic or non-mitochondrial eukaryotic homologs in the phylogenetic trees. The 30 cases surviving the manual scrutiny (Supplementary file 9) were divided into two groups.

The high confidence group comprises 17 proteins that likely share mitochondrial origin as they form a monophyletic cluster with other mitochondrial homologs. These were present only in *P. pyriformis* and/or *T. marina* and never in oxymonads. Nine out of these 17 proteins were already predicted to be mitochondrial previously (Zubáčová et al. 2013; Zítek et al. 2022). The remaining eight proteins, such as L-threonine 3-dehydrogenase (TDH) or 2Fe-2S ferredoxin (Fd) thus have a high potential for being novel mitochondrial proteins in those two protists.

The 13 putative mitochondrial proteins of the low confidence group cluster with mitochondrial proteins in other eukaryotes; however, the mitochondrial origin is not well supported. They either cluster also with cytosolic and peroxisomal isoforms, or the enzymes are known to be dually localized. An example of the former is isocitrate dehydrogenase, for which the yeast cytosolic, peroxisomal, and mitochondrial proteins form a clade in the tree. Examples of the latter are aconitase and alanine aminotransferase, for which a dual cytosolic/mitochondrial location was reported for various eukaryotes including trypanosomatids (Duschak and Cazzulo 1991; Saas et al. 2000). This low confidence group contains nine proteins present also in oxymonads (Supplementary file 9), and these are discussed in more detail below.

Aspartate and alanine aminotransferases and the two tRNA synthetases in this category are all dually localized in the mitochondrion and cytosol of *T. brucei*. In the case of tRNA synthetases, the dual localization arose specifically in trypanosomatids (Peikert et al. 2017) and, consistently with this hypothesis, Preaxostyla and *T. brucei* proteins cluster together with cytosolic tRNA synthetases of other eukaryotes. In the case of aspartate aminotransferase, the dual localization seems to be of deeper evolutionary origin (Yagi et al. 1990). The two mevalonate pathway enzymes (hydroxymethylglutaryl-CoA synthase and 3-hydroxy-3-methylglutaryl-CoA reductase) are localized in the mitochondrion of kinetoplastids (Peña-Diaz et al. 2004), however, this is again a specialty of this lineage, as eukaryotes typically run this pathway in the cytosol and the ER. In all these cases, we conservatively assume that these enzymes were localized in the cytosol in the common ancestor of Preaxostyla and *T. brucei* and were partially or fully moved to the

mitochondrion in the lineage leading to *T. brucei*. The last candidate is the malic enzyme. The evolution of this enzyme family is complicated, and there are two homologs of the enzyme in kinetoplastids, one cytoplasmic and one mitochondrial, which cluster together (Supplementary file 4). Again, it cannot be inferred what was the situation in the common ancestor with Preaxostyla.

In summary, our systematic searches for MRO proteins allowed us to update the predicted set of MRO-localized proteins of *P. pyriformis* and *T. marina* and to provide updated predictions of their MRO proteomes (Supplementary file 6). Notably, no reliable candidate for an MRO protein was detected in any of the oxymonad data sets, supporting the hypothesis of the absence of the mitochondrion in oxymonads as a group (Fig. 6).

No evidence for subcellular retargeting of ancestral mitochondrial proteins in oxymonads

Several studies demonstrated the presence of genes of alphaproteobacterial origin in non-mitochondrial compartments of eukaryotes. For example, it was proposed that more than 50% of the peroxisomal proteome is of alphaproteobacterial origin (Bolte et al. 2015). Similarly, some cytosolic enzymes such as the cytosolic uroporphyrinogen synthase participating in heme synthesis, or several glycolytic enzymes are also proposed to be of alphaproteobacterial origin (Martin and Müller 1998). A widely accepted explanation for these observations is that these genes originated by endosymbiotic gene transfer from an ancestor of mitochondria during eukaryogenesis (Martin 2010). This assumes that relocation of proteins from endosymbiont/mitochondria into other cell compartments is a simple process, in which the loss of protein-localization information either by the gene transfer to the nucleus or by the removal of the targeting peptide on the nuclear-encoded mitochondrial protein represents the only obstacle.

While a substantial effort was made to assess the degree of the "early" endosymbiotic gene transfer before last eukaryotic common ancestor (Thiergart et al. 2012; Pittis and Gabaldón 2016) relocation of proteins from established mitochondria into the cytosol in the individual eukaryote crown lineages has not been systematically studied. A careful literature review has revealed only a single case of a protein for which the relocation from mitochondria into the cytosol has been sufficiently documented, mHCF101. This protein was recently shown to be ubiquitous in mitochondria of Cryptista, Haptista, and the SAR clade. However, in the SAR member *Toxoplasma gondii* (Apicomplexa), it has been relocated to the cytosol (Barylyuk et al. 2020; Pyrih et al. 2021).

Amitochondriate oxymonads represent a suitable model group for the detection of nucleus-encoded proteins that were secondarily displaced out of the mitochondrion in the course of evolution. We argue that the analyses in the previous section suggest that not only do oxymonads contain no mitochondrion, but they also lost all the protein associated with this organelle in their metamonad ancestors, i.e., none was relocated. Firstly, oxymonads contain no clear orthologues of proteins known to localize to the MROs of *P. pyriformis* and *T. marina* (Leger et al. 2017; Zítek et al. 2022). The closest case is the second copy of glycine cleavage system L-protein (GCS-L2), which likely represents a cytosolic paralogue or xenologue with an unknown function in all Preaxostyla and it is, moreover, suspected to have undergone LGT in its history. Noteworthy is also the situation with PFO and [FeFe] hydrogenases. They are present in multiple paralogues in both oxymonads and their free-living relatives, but none of them was localized into the MRO of *P. pyriformis* (Zítek et al. 2022), so cytosolic localization of these proteins is assumed in the common ancestor of *P. pyriformis* and oxymonads, while their localization in the metamonad common ancestor cannot be inferred from their distribution and phylogenies. Finally, the comparison with the experimentally verified *T. brucei* proteome indicated several partial or full relocations of originally cytosolic proteins into the mitochondrion of trypanosomatids but not a single mitochondrion-to-cytosol protein relocation in the lineage leading to oxymonads.

The fact that we do not detect any clear protein retargeted outside mitochondria in oxymonads is surprising. Why not even a single of the mitochondrial proteins present in the common ancestor of oxymonads and trypanosomatids has been retained and repurposed in the oxymonad cell? Is there a constrain that prevents mitochondrial proteins' relocation during the reductive evolution of these organelles, or are the functions of mitochondrial proteins outside the organelle in anaerobes and microaerophiles useless?

Similar observations were made in other protists with highly reduced mitochondria, such as *G. intestinalis* or *E. histolytica*, but the topic has never been addressed systematically, and the evidence was not as clear-cut, since the proteomes of their MROs are still not robustly established. Mitosomes of *G. intestinalis* represent an interesting and extreme example. They harbour a handful of proteins and fulfil likely only one essential biological role, Fe-S cluster synthesis. To run this pathway in the organelle, the cell maintains additional pathways for mitosomal biogenesis and the import of proteins and metabolites. Although the relocation of the “last-standing” pathway into the

cytosol seems a move to a much more efficient arrangement, the evolution has never achieved it by retargeting the proteins. In the few organisms that have managed to dispense with the mitochondrial part of the Fe-S cluster synthesis pathway, the proteins were replaced by their functional xenologues acquired by LGT: SUFs in Preaxostyla and NIFs in Archamoebae (Ali et al. 2004; Gill et al. 2007; Karnkowska et al. 2016; Zítek et al. 2022). Although the protein relocations are generally considered relatively simple and probable events, relocations outside mitochondria may not be, as our data suggest.

Conclusions

In this manuscript we report the genome sequences of *P. pyriformis* and *B. nauphoetae*, adding to the previously published genomes of *M. exilis* and *S. strix*, thus doubling the number of available annotated genomes of Preaxostyla protists. By comparing the proteomes predicted from these data to available genomes and transcriptomes of metamonads we were able to define the core gene inventories of Metamonada, Preaxostyla, and Oxymonadida, and to detect LGT contributions to them. Based on its genetic toolkit, *P. pyriformis* is endowed with several functions unique among, or rare in, eukaryotes – synthesis of *p*-cresol, mercury methylation, and Se volatilization, or preforms some of the common functions in an unusual way, exemplified by the predicted fumarate-dependent NAD^+ synthesis. We identified differences in inventories of membrane transporters between oxymonads and free-living *P. pyriformis* and *T. marina* potentially reflecting the transition from the free-living to the endobiotic lifestyle at the origin of oxymonads. Proteins involved in the formation and regulation of the Golgi structure have a patchy distribution and show a trend towards loss in oxymonads, which is consistent with the lack of ultrastructural evidence for the presence of a stacked Golgi in oxymonads, but also with an emerging cell biological model for how Golgi maintain their hallmark morphology of stacked cisternae. The switch to the SUF pathway in these species has apparently not affected the number of Fe-S-containing proteins but led to a decrease in the usage of 2Fe-2S clusters.

The findings of multiple exclusively mitochondrion-associated proteins in the two studied free-living species, *T. marina* and *P. pyriformis*, corroborate the presence of a unique types of MROs in these organisms. Contrary, the thorough search for mitochondrion-associated proteins has failed to uncover any convincing candidates in the studied oxymonads, further supporting the hypothesis that not only *M. exilis* but at least a large part of oxymonads has completely lost the mitochondrion.

This fact moves this unique loss to at least 100 MYA deep past, when oxymonads had been already diversified (Poinar 2009), and shows that a eukaryotic lineage without mitochondria can thrive for eons and undergo pronounced morphological evolution. Thorough searches revealed that oxymonads not only lost the organelle but did not retain any protein acting within the mitochondrion of their ancestors, indicating that the transformation from aerobic mitochondria to microaerophilic MROs and further to the amitochondriate cell was not accompanied by the relocations of proteins to other cell compartments, but by gradual loss of all proteins concerned.

MATERIALS AND METHODS

Cell culture, DNA and RNA isolation

Monoeukaryotic, xenic culture of *P. pyriformis* (strain RCP-MX, ATCC 50935) was maintained in the Sonneborn's Paramecium medium ATCC 802 at room temperature by serial transfer of 1 ml of well-grown culture (approximately 5×10^4 cells/ml) into a 15 ml test tube containing 10 ml of fresh, bacterized medium every week. The medium was bacterized 24 hours before the transfer. *B. nauphoetae* strain NAU3 (Treitli et al. 2018) was cultured in a similar way as described above but using a modified TYSGM media (Diamond 1982).

For DNA isolation, 15 liters of *P. pyriformis* and 32 liters of *B. nauphoetae* culture were used. To remove most of the bacterial contamination, the cells were filtered as described previously (Karnkowska et al. 2016). After filtration, the cells were collected at 1200×g for 10 min at 4ºC. The DNA was isolated using two different kits. The gDNA samples for PacBio, Illumina HiSeq, and Illumina MiSeq sequencing were isolated using the Qiagen DNeasy Blood & Tissue Kit (Qiagen). The isolated gDNA was further ethanol-precipitated to increase the concentration and remove any contaminants. For Nanopore sequencing DNA was isolated using Qiagen MagAttract HMW DNA Kit (Qiagen) according to the manufacturer's protocol.

RNA for Illumina transcriptome sequencing of *P. pyriformis* was isolated from 10 liters of a well-grown culture using TRI reagent (Sigma-Aldrich) The eukaryotic mRNA was selected using the Dynabeads mRNA Purification Kit for mRNA Purification from Total RNA preps (Thermo Fisher Scientific). cDNA was synthesized by the SMARTer PCR cDNA Synthesis Kit (Takara Bio Group) and sequenced using the Illumina HiSeq 2000 sequencer at the Beijing Genomics Institute (BGI).

For RNA isolation of *B. nauphoetae* we used 500 ml of a well-grown culture. Prior to RNA isolation the culture was filtered as described previously (Karnkowska et al. 2016), and the cells were harvested by centrifugation at 1200×g for 10 min at 4ºC. The total RNA was isolated using TRI reagent (Sigma-Aldrich) according to the manufacturer's protocol. From the total RNA, mRNA was isolated using two rounds of Dynabeads Oligo(dT) beads (ThermoFisher) according to the manufacturer's protocol. The purified mRNA was used for de novo whole transcriptome sequencing performed at the Beijing Genomics Institute (BGI).

Library preparation, Illumina, and Nanopore sequencing

For *P. pyriformis* PacBio sequencing was performed at SEQme s.r.o, using a PacBio RSII sequencer, whereas Illumina sequencing was performed using Illumina HiSeq and MiSeq sequencers at the Institute of Molecular Genetics of the ASCR, v. v. i. For *B. nauphoetae*, one pair-end and two mate-pair libraries were prepared and sequenced on Illumina MIseq PE 2x300bp at Beijing Genomics Institute (BGI).

Libraries for Nanopore sequencing were prepared from 4 µg of gDNA for each library. First, DNA was sheared at ~20kbp using Covaris g-TUBES (Covaris Ltd, UK). After shearing, the library was prepared using the ligation sequencing kit from Oxford Nanopore technologies (SQK-LSK108), according to the manufacturer's protocol. The prepared libraries were loaded onto a R9.4 Spot-On Flow cell (FLO-MIN106). Sequencing was performed on a MinION Mk1B machine for 48 hours using MinKNOW 2.0 software with live base calling enabled. In total we prepared four libraries and used four flow-cells for sequencing, two for *P. pyriformis,* and two for *B. nauphoetae.*

Genome assembly binning and decontamination

The quality of sequencing data was assessed with FastQC (Andrew 2010). For the Illumina data, adapter and quality trimming was performed using Trimmomatic 0.36 (Bolger et al. 2014), with a quality threshold of 15. For the Nanopore data, trimming and removal of chimeric reads was performed using Porechop v0.2.3 (https://github.com/rrwick/Porechop).

The initial assembly of the genomes was made only with the Nanopore and PacBio generated reads using Canu v1.7.1 assembler (Koren et al. 2017), with corMinCoverage and corOutCoverage set to 0 and 100000 respectively. After assembly, the data was binned using tetraESOM (Haddad et al. 2009). The resulting eukaryotic bins were also checked using a combination of BLASTn and a

scoring strategy based on the identity and coverage of the scaffold as described in (Treitli et al. 2019). After binning, the resulted genomic bins were polished in two phases. In the first phase, the scaffolds were polished with Nanopolish (Loman et al. 2015) using the raw reads generated by Nanopore. In the second phase, the resulting scaffolds generated by Nanopolish were further corrected using Illumina short reads with Pilon v1.21 (Walker et al. 2014). Finally, the genome assembly of *P. pyriformis* was further scaffolded with raw RNA-seq reads using Rascaf (Song et al. 2016).

Repeat masking, prediction, and annotation of the genomes

Repetitive elements in the genomic assembly were identified using RepeatModeler v1.0.11 (Tarailo-Graovac and Chen 2009), and only repeats that are members of known repeat families were used for masking the genome assemblies prior to gene prediction. For the *P. pyriformis* and *B. nauphoetae*, we used Augustus for gene prediction (Stanke and Waack 2003). For *de novo* prediction of genes, Augustus was first re-trained using a manually curated set of gene models. In the next step, intron hints were generated from the RNAseq data and gene prediction was performed on repeat-masked genomes using Augustus 3.2.3 (Stanke and Waack 2003). Next, transcriptome assemblies were mapped to the genomes using PASA (Haas et al. 2003), and the resulting assembled transcripts were used as evidence for gene model polishing with EVM (Haas et al. 2008). The genome completeness for each genome was estimated using BUSCO v3 with the Eukaryota odb9 dataset and the genome completeness was estimated on the sets of Augustus-predicted protein sequences as the input.

Automatic functional annotation of the predicted genes was performed using the KEGG Automatic Annotation Server (Moriya et al. 2007), in parallel to similarity searches against NCBI nr protein database using BLASTp. Targeted analyses of genes and gene families of specific interest were performed by manual searches of the predicted proteomes using BLASTp and HMMER (Eddy 2011), and complemented by tBLASTn searches of the genome and transcriptome assemblies to check for the presence of genes that were potentially missed in the predicted protein sets. Gene models were manually refined for proteins of interest when necessary and possible. The annotated genomes were submitted to NCBI under BioProject PRJNA903905, BioSample SAMN31819772 (*P. pyriformis*) and BioProject PRJNA887011, BioSample SAMN31149887 (*B. nauphoetae*).

Defining of groups-specific gene sets

We combined the gene inventories of five Preaxostyla species (two genomes presented here + *M. exilis* genome under NCBI BioProject PRJNA304271 + *S. strix* genome under NCBI BioProject PRJNA524138 + *T. marina* transcriptome-derived predicted proteome in EukProt EP00771, Richter et al. 2022) with 14 other metamonads for which the genome or transcriptome drafts were available when the study was initiated (Supplementary file 1). The complete set of proteins from the 19 species comprised 337,300 items, of which 256,334 (76%) were assigned to 37,923 orthologous groups (OGs) by OrthoMCL (Supplementary file 1). We run the OrthoMCL pipeline (Li et al. 2003) to determine orthologous groups (OGs) of proteins of selected metamonads. To optimize the inflation parameter value, we ran multiple clusterings with different inflation parameter values in the range 1.01 to 30.0 and calculated the number of OGs containing genes from all Preaxostyla but no other taxa and from all oxymonads but no other taxa. We chose to proceed with the inflation parameter 4, because under this setting the sum of these two numbers was the highest, so the strength of the clustering should be optimal. Proteins belonging to the thus defined OGs were automatically annotated using BLASTp searches against the NCBI nr protein database (Supplementary file 1). Venn diagrams were generated using InteractiVenn (Heberle et al. 2015). Relaxed core gene set for Metamonada and its Preaxostyla and Oxymonadida specific increments were estimated based on the presence/absence of each OG in the analyzed species (Supplementary file 2).

1,399 OGs present in at least one member of each of three main metamonad clades represented in the analysis (Parabasalia, Fornicata, and Preaxostyla) were considered as the relaxed core Metamonada gene set. The Preaxostyla specific increment was determined as a union of two sets determined by two different strategies to minimize the effect of data incompleteness: i) the set of 285 OGs present in representatives of all three lineages of Preaxostyla (at least one oxymonad, *P. pyriformis*, and *T. marina*), and ii) the set of 460 OGs present in any four out of the five Preaxostyla species in the analysis. Out of these, 169 OGs were present in all five Preaxostyla species. The Oxymonadida specific increment was determined as a set of 1,242 OGs present in at least two out of the three oxymonad species yet lacking representative genes in other metamonads analyzed.

Detection of lateral gene transfer candidates

To identify candidate LGT events from prokaryotic donors into Preaxostyla species or any of their metamonad ancestors, we used a phylogenetic approach based on the identification protein phylogenies in which Preaxostyla sequences (and their other metamonad homologs if they exist) appear nested within a clade of prokaryotic counterparts with strong statistical support. We aimed to be strict in our criteria, preferring underdetection to overprediction of LGT cases.

We first discarded predicted oxymonad proteins of size smaller than 80 amino acids, since the corresponding phylogenies are unlikely to contain a strong and reliable signal. Given a large number of remaining proteins (109,359), and in order to reduce computational burden, we carried out a preliminary screen to discard sequences without detectable prokaryotic homologues. For this, each metamonad sequence was used as a query for a similarity search using DIAMOND (Buchfink et al. 2014) against the GTDB database (release 95; Parks et al. 2022) spanning 194,600 proteomes representing known prokaryotic diversity as of July 2020 (e-value cut-off = 1e-10; identity percentage >25%; query coverage >50%; subject coverage >50%). The query and subject coverage criteria were chosen to ensure we did not consider spurious hit over a short part of the protein. Metamonad sequences with significant hits against the GTDB were then used for a more exhaustive homologue search by using them as queries for DIAMOND searches against the NCBI nr database (the same cut-offs as before) in order to retrieve both prokaryotic and eukaryotic homologues. In order to reduce the computational burden, we then reduced the number of sequences retrieved from NCBI nr by running CD-HIT (-c 0.7; Li and Godzik 2006) to keep a single representative sequence for clusters of sequences displaying >70% identity. We then retrieved the NCBI taxonomy associated with each sequence.

Protein superfamilies were assembled by first running DIAMOND searches of all metamonad sequences against all (-e 1e-20 --id 25 --query-cover 50 --subject-cover 50). Reciprocal hits were gathered into a single FASTA file, as well as their NCBI nr homologues. We also added all metamonad sequences belonging to the same OGs into a single FASTA file.

Files with 30 sequences or fewer were discarded in order to guarantee that the directionality of the LGTs can be inferred from the phylogenies. In addition, files with more eukaryotic homologues than prokaryotic ones were also discarded from follow-up analyses, as they were unlikely to

represent a clear case of prokaryote-to-eukaryote gene transfer. The remaining files were individually aligned using MAFFT (Katoh and Standley 2013) and trimmed using BMGE (Criscuolo and Gribaldo 2010), and the corresponding phylogenies were inferred using FastTree v.2 (Price et al. 2010).

To investigate the potential cases of prokaryotes-to-metamonads LGT, we used the R package PhySortR (Stephens et al. 2016) in order to identify well-supported clades comprised of metamonad sequences nested within prokaryotic ones. This non-trivial task was done in a stepwise manner by first identifying clans with a statistical support (Shimodaira-Hasegawa test) >0.8 and containing at least 90% of prokaryotes and metamonads. The reason for the latter threshold is to allow for cases of taxonomic misannotation, or for the presence of a few homologues from other eukaryotes, themselves potentially resulting from LGT. After this first assessment, we checked which metamonad sequences were forming a well-supported monophyletic clan (i.e., a clan with >90% of metamonads with >0.8 support) in order to infer if they likely corresponded to a single LGT even within their last common ancestor, or if they were dispersed in the tree suggesting multiple independent transfers of homologous genes.

However, when metamonad sequences were detected as paraphyletic, and in order to avoid overestimating the number of LGT events, we manually inspected all phylogenies in order to distinguish most accurately between probable independent cases of LGTs and paraphyly resulting from an overall lack of phylogenetic resolution. For this, our manual inspection took into account: 1) the overall number of sequences in the tree (i.e., a small number can be indicative that only a few homologues passed our DIAMOND threshold because this protein family is fast evolving and that, consequently, the corresponding phylogeny might be poorly resolved and/or not contain representatives of the true diversity of homologues); 2) the branch length separating the metamonad sequences (i.e., short branch lengths can be suggestive of an unreliably resolved backbone); 3) the overall taxonomic coherence of the tree (i.e., do the various clades display taxonomic homogeneity or are the lineages interspersed all over the tree, suggestive of a poorly resolved phylogeny?); 4) the taxonomic affiliation of the prokaryotic neighbors of each independent metamonad clade in the tree (i.e., if they are similar, this is more in favor of a single LGT event ancestral to all metamonad represented in the tree, which phylogenetic reconstructions have not resolved properly); 5) the overall branch support across the tree (i.e., some trees might show a few well supported branches

suggestive of paraphyletic metamonads but if the overall backbone of tree is generally poorly supported, we were cautious about the weakness of the phylogenetic signal). Taking all this into account, in case of doubt, we always minimized the number of inferred LGT events by assuming a single event in the last common ancestor of the metamonad sequences in a tree. All phylogenies and FASTA files can be found in Supplementary file 11.

Single gene phylogenies

The phylogenies of the genes of interest were analyzed individually using the methodology described below. Eukaryotic homologues of the Preaxostyla genes were gathered by taxonomically constrained iterative BLAST search against publicly available sequence databases in order to sample as broad eukaryotic diversity as possible. In the cases of overrepresented taxa of low interest (e.g. Metazoa, land plants), only a small number of representative sequences were selected arbitrarily. Prokaryotic homologues were gathered by a BLASTp search with each eukaryotic sequence against the NCBI nr database with an e-value cutoff of 1e-10 and max. 10 target sequences. The sequences were aligned using MAFFT v7.221 (Katoh and Standley 2013) and automatically trimmed using trimAl v1.2 (Capella-Gutiérrez et al. 2009). Phylogenetic analyses were performed simultaneously using IQ-Tree v2.0.5 (Nguyen et al. 2015) and RAxML-HPC2 on v8.2.12 (Stamatakis 2014) on the Cyberinfrastructure for the Phylogenetic Research (CIPRES) Science Gateway (Miller et al. 2010). Substitution models were inferred using IQ-Tree TESTNEW function for IQ-Tree and using ModelTest-NG v0.1.5 (Darriba et al. 2019) on the CIPRES Science Gateway for RAxML.

Analysis of the Rsg1 gene

Candidate Rsg1 orthologues were identified in a variety of databases and other sequence resources (Supplementary file 5) using a combination of BLAST (BLASTp or tBLASTn, depending on the target database) and HMMER searches. The taxon selection essentially followed the list of taxa employed in a recent study of the distribution of the cilium-associated Ras superfamily GTPase Arl16 (Dewees et al. 2022), with further expansion to include additional metamonads and recently sequenced species of isolated phylogenetic position. Additional taxa were added based on searching the newly released EukProt database (Richter et al. 2022) such that representatives of broader taxa consistently possessing or lacking a discernible Rsg1 orthologue were selected and

including the final list. Significant BLAST hits retrieved from searches with reference queries (the human sequence – GenBank accession number NP_112169.1, or another already validated sequence from a given taxon to increase the sensitivity of the search in a taxon-specific manner) were evaluated by BLAST searches against a large custom manually annotated database of Ras superfamily genes to distinguish Rsg1 orthologues from other GTPase types. The discrimination was facilitated by Rsg1 being rather divergent from other Ras superfamily members (such as Rab GTPases). A set of confidently identified Rsg1 sequences was then used for building a custom profile HMM, which was employed in searches with HMMER to identify possible divergent Rsg1 orthologues missed by BLAST searches. The candidates were again evaluated by BLAST-based comparisons against the custom GTPase database.

Phylogenetic analysis of Golgi-related proteins

Comparative genomics was carried out using HMM searches. Pan-eukaryotic query sequences analyzed and curated for previous pan-eukaryotic vesicle coat, multisubunit tethering complexes, and golgins were used to build profile HMMs. Query sequences were obtained from supplementary material of previous papers dealing specifically with Adaptins and COPI (Hirst et al. 2014), multisubunit tethering complexes (Klinger et al. 2013), and golgins (Barlow et al. 2018). Individual components and proteins from each sub-complex were aligned using MUSCLE v.3.8.31 and the resulting alignment files were used to generate HMM matrices using the HMMBUILD option available through the HMMER package (Eddy 2011). HMMER searches were carried out in the predicted proteomes of *P. pyriformis*, *B. nauphoetae*, and *S. strix*, whereas for *T. marina* the transcriptome assembly was first translated in all six open reading frames using the *ab initio* gene prediction program GeneMarkS-T using the default parameters (Tang et al. 2015) and the longest resulting predicted protein sequences were used for the searches. Forward hits meeting an e-value cutoff of 0.05 were subject to reciprocal BLASTP analyses against the *Homo sapiens* and *M. exilis* predicted proteomes as well as the NCBI nr database. Any absent components were also subject to additional tBLASTn searches in the nucleotide scaffolds. Hits were deemed positive if both forward hits and at least two of three reciprocal BLAST results retrieved the correct ortholog with an e-value cutoff of 0.05. Paralogous gene families were subject to further phylogenetic analyses.

Phylogenetic analyses were undertaken for the large, medium and small subunits from identified AP1-4 and CopI families, as well as Syntaxin16 and Syntaxin5. Resolved backbone alignments

from dataset curated for Karnkowska et al. (2019) were used and sequences from *T. marina*, *P. pyriformis*, *B. nauphoetae*, and *S. strix* were iteratively aligned with the backbone alignment using the profile option available through MUSCLE v3.8.31 (Edgar 2004). All alignments were subsequently inspected and manually trimmed using Mesquite v. 3.5 (https://www.mesquiteproject.org/) to remove heterogeneous regions lacking obvious homology and partial or misaligned sequences.

Maximum likelihood analyses were carried out using RAxML-HPC2 on XSEDE v.8.2.10 for non-parametric bootstrap replicates (Stamatakis 2006). The best protein matrix model for RAxML was tested using ProtTest v.3.4.2 (Abascal et al. 2005), set to account for Gamma rate variation (+G), invariant sites (+I), and observed frequency of amino acids (+F) with default tree and a consensus tree was obtained using the Consense package, available through the Phylip v.3.66.

Bayesian inference was carried out using MRBAYES on XSEDE v3.2.6 (Huelsenbeck and Ronquist 2001). Parameters specified included 10 million Markov Chain Monte Carlo generations under a mixed amino acid model with the number of gamma rate categories set a 4. Sampling frequency was set to occur every 1000 generations with a burn-in of 0.25. Tree convergence was ensured with the average standard deviation of split frequency values below 0.01. Both RAxML and MRBAYES analyses were performed on the CIPRES webserver (Miller et al. 2015). RAxML bootstrap values as well as Bayesian posterior probabilities were overlaid on the best Bayes topology with combined values of >50 and >0.80, respectively, indicating branch support. Tree visualization and rooting were carried out in FigTree v.1.4.4. Branch support value overlay and additional annotations were prepared using Adobe Illustrator CS4. All alignments are available upon request.

Prediction of proteins containing Fe-S clusters

Fe-S cluster containing proteins were predicted with MetalPredator webserver (Valasatava et al. 2016). Predicted proteins were functionally annotated by BLAST searches against NCBI nr database and by InterProScan against the InterPro database (Blum et al. 2021). KEGG categories were assigned by GhostKOALA searches against the KEGG database (Kanehisa et al. 2016). Orthologous groups were created with the OrthoFinder 2 software (Emms and Kelly 2019) and the

Venn diagram of OG sharing among Preaxostyla was visualized using InteractiVenn (Heberle et al. 2015).

Search for mitochondrial proteins

The comprehensive search for putative mitochondrial protein was performed for all five Preaxostyla species, including previously analyzed *M. exilis* (Karnkowska et al. 2016) as a control. Predicted proteins of each species were independently searched for similarity to known mitochondrial/MRO-related proteins and the presence of mitochondrial signature sequences. The general design of the search followed the previously described methodology (Karnkowska et al. 2016). Briefly, a custom mitochondrial protein sequence database was established using the MitoMiner v4.0 database (Smith et al. 2012; Smith and Robinson 2018). Experimentally confirmed mitochondrial proteins (at least one GFP tagging experiment or three different MS experiments) coming from *H. sapiens*, *Mus musculus*, *Rattus norvegicus*, *Danio rerio*, *Saccharomyces cerevisiae*, and *Schizosaccharomyces pombe* were used and supplemented with the MROs proteins from sixteen different organisms (Stechmann et al. 2008; Mi-ichi et al. 2009; Barberà et al. 2010; Alcock et al. 2012; Stairs et al. 2014; Noguchi et al. 2015; Nývltová et al. 2015; Leger et al. 2016; Leger et al. 2017; Pyrihová et al. 2018). Redundant homologues (90% similarity threshold) were removed from the database using CD-HIT (Li and Godzik 2006). The resulting non-redundant database contained 6,979 proteins. Reciprocal BLAST analysis was performed for each predicted proteome with an e-value threshold of 0.001. HMM searches were used to identify proteins involved in protein import and translocation, as these were shown to be often divergent (Leger et al. 2016). Searches were done in HMMER 3.1b2 (Eddy 2011) using profile HMMs previously employed in Karnkowska et al. (2016).

Mitochondrial targeting signals were detected using TargetP v1.1 (Emanuelsson et al. 2007) and MitoFates v1.1 (Fukasawa et al. 2015). Both programs indicate a putative probability of mitochondrial localization of the protein. Hits with the probability of mitochondrial localization indicated to be >0.5 by both programs were considered for manual verification. To find tail-anchored proteins, transmembrane domains (TMDs) were predicted using TMHMM2.0 (Krogh et al. 2001) for all analyzed proteins. Hits with a TMD within 32 amino acid residues from the C-terminus were kept for verification. The mitochondrial β-barrel outer membrane proteins (MBOMPs) search has been conducted using the pipeline described by Imai et al. (2011). The

pipeline firstly identifies a β-signal ($P_oxGh_yxH_yxH_y$ motif) in the C-terminus of protein, required for the insertion into the membrane. Subsequently, the secondary structure of the stretch of 300 amino acid residues preceding the β-signal is analyzed using PSIPRED (McGuffin et al. 2000) to check for a typical β-structure. Significant hits, with at least 25% of the sequence predicted to form β-strands, no more than 10% assumed by an α-helical structure, and no more than 50% of the eight residues of the β-signal predicted as an α-helical structure, were further analyzed.

In all methods, the hits were checked with additional attention paid to the proteins with hits in the MitoMiner database, herein considered "candidates". Since divergent MBOMPs have been shown to be missed by homology-based searches, for those we used NCBI nr instead, following the idea that any match with NCBI-nr should allow us to distinguish between unusual and improperly predicted proteins. All "candidates" have been thoroughly manually inspected. All proteins have been BLAST-searched against NCBI-nr and the best hits with the description not including the terms 'low quality protein', 'hypothetical', 'unknown', etc. in the description were kept. For each hit, the Gene Ontology categories were assigned using InterProScan-5.36-75.0 (Jones et al. 2014). The annotation of the *M. exilis* genome (Karnkowska et al. 2019) was transferred to orthologous proteins from the other Preaxostyla species. Based on gathered information, verified "candidates" are considered mitochondrial proteins.

Trypanosoma brucei mitoproteome-guided comparative analyses

Predicted proteomes of *T. marina*, *P. pyriformis*, *S. strix*, *B. nauphoetae,* and *M. exilis* were individually reverse BLAST-searched against the proteome of *T. brucei* (downloaded from Tritrypdb.org; November 2019; Aslett et al. 2010). Only reciprocal best blast hits that were identified in either of the previously published mitochondrial proteomes of *T. brucei* (Panigrahi et al. 2009; Peikert et al. 2017) were subjected to further phylogenetic analysis. For each such protein, the data set for the tree construction was composed of hits from a custom BLAST database of selected protist proteomes (Supplementary file 11, downloaded from UniProt, November 2019; Bateman 2019). Protein sequence sets were automatically aligned with MAFFT v7.453 (Katoh and Standley 2013) using the L-INS-i refinement and a maximum of 1,000 iterations, followed by trimming of sites with >70% gaps.

ML trees (Supplementary file 10) were inferred by IQ-TREE v 1.6.12 using the Posterior Mean Site Frequency (PMSF) empirical model with the ultrafast bootstrapping strategy (1,000 replicates) and a LG4X guide tree. Subcellular targeting of all proteins in the tree was predicted by using TargetP-1.1 (Emanuelsson et al. 2007; https://services.healthtech.dtu.dk/service.php?TargetP-1.1); the presence of a signal peptide, a chloroplast targeting peptide or a mitochondrial targeting peptide in the respective proteins is marked by the letters S, C, or M, respectively, at the very beginning of the sequence names.

ACKNOWLEDGEMENTS

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 771592) and the Centre for research of pathogenicity and virulence of parasites (registration no. CZ.02.1.01/0.0/0.0/16_019/0000759). Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA LM2018140) provided within the program Projects of Large Research, Development and Innovations Infrastructures. Research in Karnkowska lab is supported by EMBO Installation Grant 4150 and Ministry of Education and Science, Poland and the Interdisciplinary Centre for Mathematical and Computational Modelling (ICM) University of Warsaw under computational allocation no G 72-16. Research in the Dacks Lab is supported by grants from the Natural Sciences and Research Council of Canada (RES0021028, RES0043758, and RES0046091) and SVP was supported by an Alberta Innovates Graduate Studentship (Doctoral) and Canadian Institutes of Health Research Canada Graduate Scholarships. ME was supported by the Czech Science Foundation project 22-29633S.

REFERENCES

Abascal F, Zardoya R, Posada D. 2005. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* 21:2104–2105.

Agbu SO, Liang Y, Liu A, Anderson K v. 2018. The small GTPase RSG1 controls a final step in primary cilia initiation. *J Cell Biol* 217:413–427.

Ahat E, Li J, Wang Y. 2019. New Insights into the Golgi Stacking Proteins. *Front Cell Dev Biol* 7:131.

Alcock F, Webb CT, Dolezal P, Hewitt V, Shingu-Vasquez M, Likić VA, Traven A, Lithgow T. 2012. A small Tim homohexamer in the relict mitochondrion of *Cryptosporidium*. *Mol Biol Evol* 29:113–122.

Ali V, Shigeta Y, Tokumoto U, Takahashi Y, Nozaki T. 2004. An intestinal parasitic protist, *Entamoeba histolytica*, possesses a non-redundant nitrogen fixation-like system for iron-sulfur cluster assembly under anaerobic conditions. *J Biol Chem* 279:16863–16874.

Anderson IJ, Loftus BJ. 2005. *Entamoeba histolytica*: Observations on metabolism based on the genome sequence. *Exp Parasitol* 110:173–177.

Andreini C, Banci L, Rosato A. 2016. Exploiting Bacterial Operons to Illuminate Human Iron-Sulfur Proteins. *J Proteome Res* 15:1308–1322.

Andrew S. 2010. FastQC: a quality control tool for high throughput sequence data.

Aridor M. 2022. A tango for coats and membranes: New insights into ER-to-Golgi traffic. *Cell Rep* 38:110258.

Aslett M, Aurrecoechea C, Berriman M, Brestelli J, Brunk BP, Carrington M, Depledge DP, Fischer S, Gajria B, Gao X, et al. 2010. TriTrypDB: a functional genomic resource for the Trypanosomatidae. *Nucleic Acids Res* 38:D457–D462.

Babady NE, Pang YP, Elpeleg O, Isaya G. 2007. Cryptic proteolytic activity of dihydrolipoamide dehydrogenase. *Proc Natl Acad Sci U S A* 104:6158–6163.

Balaji S, Aravind L. 2007. The RAGNYA fold: a novel fold with multiple topological variants found in functionally diverse nucleic acid, nucleotide and peptide-binding proteins. *Nucleic Acids Res* 35:5658–5671.

Barberà MJ, Ruiz-Trillo I, Tufts JY a, Bery A, Silberman JD, Roger AJ. 2010. *Sawyeria marylandensis* (Heterolobosea) has a hydrogenosome with novel metabolic properties. *Eukaryot Cell* 9:1913–1924.

Barlow LD, Nývltová E, Aguilar M, Tachezy J, Dacks JB. 2018. A sophisticated, differentiated Golgi in the ancestor of eukaryotes. *BMC Biol* 16:27.

Barylyuk K, Koreny L, Ke H, Butterworth S, Crook OM, Lassadi I, Gupta V, Tromer E, Mourier T, Stevens TJ, et al. 2020. A Comprehensive Subcellular Atlas of the Toxoplasma Proteome via hyperLOPIT Provides Spatial Context for Protein Functions. *Cell Host Microbe* 28:752-766.e9.

Bateman A. 2019. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* 47:D506–D515.

van der Bliek AM, Shen Q, Kawajiri S. 2013. Mechanisms of mitochondrial fission and fusion. *Cold Spring Harb Perspect Biol* 5:a011072.

Blum M, Chang H-Y, Chuguransky S, Grego T, Kandasaamy S, Mitchell A, Nuka G, Paysan-Lafosse T, Qureshi M, Raj S, et al. 2021. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res* 49(D1): D344–D354.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120.

Bolte K, Rensing SA, Maier UG. 2015. The evolution of eukaryotic cells from the perspective of peroxisomes. *BioEssays* 37:195–203.

Boncompain G, Weigel A v. 2018. Transport and sorting in the Golgi complex: multiple mechanisms sort diverse cargo. *Curr Opin Cell Biol* 50:94–101.

Borgese N, Brambillasca S, Colombo S. 2007. How tails guide tail-anchored proteins to their destinations. *Curr Opin Cell Biol* 19:368–375.

Brännström IO, Stairs C, Campos KIA, Ettema TJG, Keeling PJ, Bass D, Burki F. 2022. A Mitosome With Distinct Metabolism in the Uncultured Protist Parasite *Paramikrocytos canceri* (Rhizaria, Ascetosporea). *SSRN Electronic Journal* SSRN:4065074.

Braymer JJ, Freibert SA, Rakwalska-Bange M, Lill R. 2021. Mechanistic concepts of iron-sulfur protein biogenesis in Biology. *Biochim Biophys Acta Mol Cell Res* 1868(1):118863.

Buchfink B, Xie C, Huson DH. 2014. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* 12:59–60.

Butenko A, Opperdoes FR, Flegontova O, Horák A, Hampl V, Keeling P, Gawryluk RMR, Tikhonenkov D, Flegontov P, Lukeš J. 2020. Evolution of metabolic capabilities and molecular features of diplonemids, kinetoplastids, and euglenids. *BMC Biol* 18:1–28.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25:1972–1973.

Carlton JM, Hirt RP, Silva JC, Delcher AL, Schatz M, Zhao Q, Wortman JR, Bidwell SL, Alsmark UCM, Besteiro S, et al. 2007. Draft genome sequence of the sexually transmitted pathogen *Trichomonas vaginalis. Science* 315:207–212.

Christensen GA, Wymore AM, King AJ, Podar M, Hurt RA, Santillan EU, Soren A, Brandt CC, Brown SD, Palumbo A v., et al. 2016. Development and validation of broad-range qualitative and clade-specific quantitative molecular probes for assessing mercury methylation in the environment. *Appl Environ Microbiol* 82:6068–6078.

Cooper CJ, Zheng K, Rush KW, Johs A, Sanders BC, Pavlopoulos GA, Kyrpides NC, Podar M, Ovchinnikov S, Ragsdale SW, et al. 2020. Structure determination of the HgcAB complex using metagenome sequence data: insights into microbial mercury methylation. *Communications Biology* 3:1–9.

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

Darriba D, Posada David, Kozlov AM, Stamatakis A, Morel B, Flouri T. 2019. ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models. *bioRxiv* 612903.

Dean S, Sunter JD, Wheeler RJ. 2017. TrypTag.org: A Trypanosome Genome-wide Protein Localisation Resource. *Trends Parasitol* 33:80–82.

Denic V. 2012. A portrait of the GET pathway as a surprisingly complicated young man. *Trends Biochem Sci* 37:411–417.

Derelle R, Torruella G, Klimeš V, Brinkmann H, Kim E, Vlček Č, Lang BF, Eliáš M. 2015. Bacterial proteins pinpoint a single eukaryotic root. *Proc Natl Acad Sci U S A* 112:E693–E699.

Dewees SI, Vargová R, Hardin KR, Turn RE, Devi S, Linnert J, Wolfrum U, Caspary T, Eliáš M, Kahn RA. 2022. Phylogenetic profiling and cellular analyses of ARL16 reveal roles in traffic of IFT140 and INPP5E. *Mol Biol Cell* 33:ar33.

Diamond LS. 1982. A new liquid medium for xenic cultivation of *Entamoeba histolytica* and other lumen-dwelling protozoa. *J Parasitol* 68:958–959.

Dolezal P, Likic V, Tachezy J, Lithgow T. 2006. Evolution of the Molecular Machines for Protein Import into Mitochondria. *Science (1979)* 313:314–319.

Ducker GS, Rabinowitz JD. 2017. Cell Metabolism Review One-Carbon Metabolism in Health and Disease. *Cell Metab* 25:27–42.

Duschak VG, Cazzulo JJ. 1991. Subcellular localization of glutamate dehydrogenases and alanine aminotransferase in epimastigotes of *Trypanosoma cruzi*. *FEMS Microbiol Lett* 83:131–135.

Eberhardt EL, Ludlam A v., Tan Z, Cianfrocco MA. 2020. Miro: A molecular switch at the center of mitochondrial regulation. *Protein Sci* 29:1269–1284.

Eddy SR. 2011. Accelerated profile HMM searches. *PLoS Comput Biol* 7(10): e1002195.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797.

Emanuelsson O, Brunak S, von Heijne G, Nielsen H. 2007. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc* 2:953–971.

Emms DM, Kelly S. 2019. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol* 20:238.

Fukasawa Y, Tsuji J, Fu SC, Tomii K, Horton P, Imai K. 2015. MitoFates: Improved prediction of mitochondrial targeting sequences and their cleavage sites. *Molecular and Cellular Proteomics* 14:1113–1126.

Gentekaki E, Curtis BA, Stairs CW, Klimeš V, Eliáš M, Salas-Leiva DE, Herman EK, Eme L, Arias MC, Henrissat B, et al. 2017. Extreme genome diversity in the hyper-prevalent parasitic eukaryote *Blastocystis*. *PLoS Biol* 11;15(9):e2003769.

Gill EE, Diaz-Triviño S, Barberà MJ, Silberman JD, Stechmann A, Gaston D, Tamas I, Roger AJ. 2007. Novel mitochondrion-related organelles in the anaerobic amoeba *Mastigamoeba balamuthi*. *Mol Microbiol* 66:1306–1320.

Gilmour CC, Elias DA, Kucken AM, Brown SD, Palumbo A v., Schadt CW, Wall JD. 2011. Sulfate-reducing bacterium *Desulfovibrio desulfuricans* ND132 as a model for understanding bacterial mercury methylation. *Appl Environ Microbiol* 77:3938–3951.

Glaser T, Hedman B, Hodgson KO, Solomon EI. 2000. Ligand K-edge X-ray absorption spectroscopy: A direct probe of ligand - Metal covalency. *Acc Chem Res* 33(12):859–868.

Greening C, Biswas A, Carere CR, Jackson CJ, Taylor MC, Stott MB, Cook GM, Morales SE. 2016. Genomic and metagenomic surveys of hydrogenase distribution indicate H 2 is a widely utilised energy source for microbial growth and survival. *ISME Journal* 10:761–777.

Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, et al. 2003. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* 31:5654–5666.

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Robin CR, Wortman JR. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* 9:R7.

Haddad I, Hiller K, Frimmersdorf E, Benkert B, Schomburg D, Jahn D. 2009. An emergent self-organizing map based analysis pipeline for comparative metabolome studies. *In Silico Biol* 9:163–178.

Hampl V. 2017. Preaxostyla. In: Handbook of the Protists. Cham: Springer International Publishing. p. 1139–1174.

Hampl V, Čepička I, Eliáš M. 2018. Was the mitochondrion necessary to start eukaryogenesis? *Trends Microbiol* 27:1–9.

Hampl V, Silberman JD, Stechmann A, Diaz-Triviño S, Johnson PJ, Roger AJ. 2008. Genetic Evidence for a Mitochondriate Ancestry in the 'Amitochondriate' Flagellate *Trimastix pyriformis*. *PLoS One* 3:9.

Hazelwood LA, Daran JM, van Maris AJA, Pronk JT, Dickinson JR. 2008. The Ehrlich pathway for fusel alcohol production: A century of research on *Saccharomyces cerevisiae* metabolism. *Appl Environ Microbiol* 74:2259–2266.

Heberle H, Meirelles VG, da Silva FR, Telles GP, Minghim R. 2015. InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. *BMC Bioinformatics* 16:169.

Hirst J, Schlacht A, Norcott JP, Traynor D, Bloomfield G, Antrobus R, Kay RR, Dacks JB, Robinson MS. 2014. Characterization of TSET, an ancient and widespread membrane trafficking complex. *Elife* 3:e02866.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17:754–755.

Igamberdiev AU, Bykova N v., Ens W, Hill RD. 2004. Dihydrolipoamide dehydrogenase from porcine heart catalyzes NADH-dependent scavenging of nitric oxide. *FEBS Lett* 568:146–150.

Imai K, Fujita N, Gromiha MM, Horton P. 2011. Eukaryote-wide sequence analysis of mitochondrial β-barrel outer membrane proteins. *BMC Genomics* 12:79.

Jerlström-Hultqvist J, Einarsson E, Xu F, Hjort K, Ek B, Steinhauf D, Hultenby K, Bergquist J, Andersson JO, Svärd SG. 2013. Hydrogenosomes in the diplomonad Spironucleus salmonicida. *Nat Commun* 4:2493.

Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. 2014. InterProScan 5: Genome-scale protein function classification. *Bioinformatics* 30:1236–1240.

Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2014. Data, information, knowledge and principle: Back to metabolism in KEGG. *Nucleic Acids Res* 42:D199-205.

Kanehisa M, Sato Y, Morishima K. 2016. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J Mol Biol*. 428:726-731.

Karnkowska A, Treitli SC, Brzoň O, Novák L, Vacek V, Soukal P, Barlow LD, Herman EK, Pipaliya S V, Pánek T, et al. 2019. The oxymonad genome displays canonical eukaryotic complexity in the absence of a mitochondrion. *Mol Biol Evol* 36:2292–2312.

Karnkowska A, Vacek V, Zubáčová Z, Treitli SC, Petrželková R, Eme L, Novák L, Žárský V, Barlow LD, Herman EK, et al. 2016. A eukaryote without a mitochondrial organelle. *Current Biology* 26:1274–1284.

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

Klinger CM, Karnkowska A, Herman EK, Hampl V, Dacks JB. 2016. Phylogeny and Evolution. In: Molecular Parasitology. Vienna: Springer Vienna. p. 383–408.

Klinger CM, Klute MJ, Dacks JB. 2013. Comparative genomic analysis of multi-subunit tethering complexes demonstrates an ancient pan-eukaryotic complement and sculpting in Apicomplexa. *PLoS One* 8:e76278.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* 27:722–736.

Krogh A, Larsson B, Von Heijne G, Sonnhammer ELL. 2001. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J Mol Biol* 305:567–580.

Kulkarni-Gosavi P, Makhoul C, Gleeson PA. 2019. Form and function of the Golgi apparatus: scaffolds, cytoskeleton and signalling. *FEBS Lett* 593:2289–2305.

Langousis G, Cavadini S, Boegholm N, Lorentzen E, Kempf G, Matthias P. 2022. Structure of the ciliogenesis-associated CPLANE complex. *Sci Adv* 8(15):eabn0832.

Lee I, Tiwari N, Dunlop MH, Graham M, Liu X, Rothman JE. 2014. Membrane adhesion dictates Golgi stacking and cisternal morphology. *Proc Natl Acad Sci U S A* 111:1849–1854.

Leger MM, Eme L, Hug LA, Roger AJ. 2016. Novel hydrogenosomes in the microaerophilic jakobid *Stygiella incarcerata*. *Mol Biol Evol* 33:2318–2336.

Leger MM, Kolisko M, Kamikawa R, Stairs CW, Kume K, Čepička I, Silberman JD, Andersson JO, Xu F, Yabuki A, et al. 2017. Organelles that illuminate the origins of *Trichomonas* hydrogenosomes and *Giardia* mitosomes. *Nat Ecol Evol* 1:0092.

Li Jie, Ahat E, Wang Yanzhuang, Li J, Ahat · E, Wang Y. 2019. Golgi Structure and Function in Health, Stress, and Diseases. *Results Probl Cell Differ* 67:441–485.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 13:2178–2189.

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

Liapounova NA, Hampl V, Gordon PMK, Sensen CW, Gedamu L, Dacks JB. 2006. Reconstructing the mosaic glycolytic pathway of the anaerobic eukaryote Monocercomonoides. *Eukaryot Cell* 5:2138–2146.

Lindmark DG, Müller M. 1973. Hydrogenosome, a cytoplasmic organelle of the anaerobic flagellate *Tritrichomonas foetus*, and its role in pyruvate metabolism. *J Biol Chem* 248:7724–7728.

Loman NJ, Quick J, Simpson JT. 2015. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat Methods* 12:733–735.

Lucattini R, Likić VA, Lithgow T. 2004. Bacterial Proteins Predisposed for Targeting to Mitochondria. *Mol Biol Evol* 21:652–658.

Major P, Embley TM, Williams TA. 2017. Phylogenetic Diversity of NTT Nucleotide Transport Proteins in Free-Living and Parasitic Bacteria and Eukaryotes. *Genome Biol Evol* 9:480.

Martin W. 2010. Evolutionary origins of metabolic compartmentalization in eukaryotes. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365:847–855.

Martin W, Müller M. 1998. The hydrogen hypothesis for the first eukaryote. *Nature* 392:37–41.

Maxwell Burroughs A, Aravind L. 2014. A highly conserved family of domains related to the DNA-glycosylase fold helps predict multiple novel pathways for RNA modifications. *RNA Biol* 11:360–372.

Maxwell Burroughs A, Glasner ME, Barry KP, Taylor EA, Aravind L. 2019. Oxidative opening of the aromatic ring: Tracing the natural history of a large superfamily of dioxygenase domains and their relatives. *Journal of Biological Chemistry* 294:10211–10235.

McGuffin LJ, Bryson K, Jones DT. 2000. The PSIPRED protein structure prediction server. *Bioinformatics* 16:404–405.

Mertens E. 1993. ATP versus pyrophosphate: glycolysis revisited in parasitic protists. *Parasitology Today* 9:122–126.

Mi-ichi F, Yousuf MA, Nakada-Tsukui K, Nozaki T, Abu Yousuf M, Nakada-Tsukui K, Nozaki T. 2009. Mitosomes in *Entamoeba histolytica* contain a sulfate activation pathway. *Proceedings of the National Academy of Sciences* 106:21731–21736.

Miller MA, Pfeiffer W, Schwartz T. 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In: 2010 Gateway Computing Environments Workshop (GCE). IEEE. p. 1–8.

Miller MA, Schwartz T, Pickett BE, He S, Klem EB, Scheuermann RH, Passarotti M, Kaufman S, O'Leary MA. 2015. A RESTful API for Access to Phylogenetic Tools via the CIPRES Science Gateway. *Evolutionary Bioinformatics* 11:EBO.S21501.

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. 2007. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* 35:W182–W185.

Müller M, Mentel M, van Hellemond JJ, Henze K, Woehle C, Gould SB, Yu R-Y, van der Giezen M, Tielens AGM, Martin WF. 2012. Biochemistry and evolution of anaerobic energy metabolism in eukaryotes. *Microbiol Mol Biol Rev* 76:444–495.

Nakjang S, Williams TA, Heinz E, Watson AK, Foster PG, Sendra KM, Heaps SE, Hirt RP, Embley TM. 2013. Reduction and expansion in microsporidian genome evolution: New insights from comparative genomics. *Genome Biol Evol* 5:2285–2303.

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

Noguchi F, Shimamura S, Nakayama T, Yazaki E, Yabuki A, Hashimoto T, Inagaki Y, Fujikura K, Takishita K. 2015. Metabolic Capacity of Mitochondrion-related Organelles in the Free-living Anaerobic Stramenopile Cantina marsupialis. *Protist* 166:534–550.

Noodleman L, Case DA. 1992. Density-functional theory of spin polarization and spin coupling in iron—sulfur clusters. *Adv Inorg Chem* 38:423-458.

Novák L, Zubáčová Z, Karnkowska A, Kolisko M, Hroudová M, Stairs CW, Simpson AGB, Keeling PJ, Roger AJ, Čepička I, et al. 2016. Arginine deiminase pathway enzymes: evolutionary history in metamonads and other eukaryotes. *BMC Evol Biol* 16:197.

Nývltová E, Stairs CW, Hrdý I, Rídl J, Mach J, Paqes J, Roger AJ, Tachezy J. 2015. Lateral gene transfer and gene duplication played a key role in the evolution of *Mastigamoeba balamuthi* hydrogenosomes. *Mol Biol Evol* 32:1039–1055.

Nývltová E, Šut'ák R, Žárský V, Harant K, Hrdý I, Tachezy J. 2017. Lateral gene transfer of p-cresol- and indole-producing enzymes from environmental bacteria to *Mastigamoeba balamuthi. Environ Microbiol* 19:1091–1102.

O'Kelly CJ, Farmer MA, Nerad TA. 1999. Ultrastructure of *Trimastix pyriformis* (Klebs) Bernard et al.: similarities of *Trimastix* species with retortamonad and jakobid flagellates. *Protist* 150:149–162.

Panigrahi AK, Ogata Y, Zíková A, Anupama A, Dalley RA, Acestor N, Myler PJ, Stuart KD. 2009. A comprehensive analysis of *Trypanosoma brucei* mitochondrial proteome. *Proteomics* 9:434–450.

Park K, Ju S, Kim N, Park SY. 2021. The Golgi complex: a hub of the secretory pathway. *BMB Rep* 54:246.

Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil PA, Hugenholtz P. 2022. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res* 50:D785–D794.

Payne MJ, Chapman A, Cammack R. 1993. Evidence for an [Fe]-type hydrogenase in the parasitic protozoan *Trichomonas vaginalis*. *FEBS Lett* 317:101–104.

Peikert CD, Mani J, Morgenstern M, Käser S, Knapp B, Wenger C, Harsman A, Oeljeklaus S, Schneider A, Warscheid B. 2017. Charting organellar importomes by quantitative mass spectrometry. *Nat Commun* 8:15272.

Peña-Diaz J, Montalvetti A, Flores C-L, Constán A, Hurtado-Guerrero R, De Souza W, Gancedo C, Ruiz-Perez LM, Gonzalez-Pacanowska D. 2004. Mitochondrial localization of the mevalonate pathway enzyme 3-Hydroxy-3-methyl-glutaryl-CoA reductase in the Trypanosomatidae. *Mol Biol Cell* 15:1356–1363.

Petrat F, Paluch S, Dogruöz E, Dörfler P, Kirsch M, Korth HG, Sustmann R, de Groot H. 2003. Reduction of Fe(III) Ions Complexed to Physiological Ligands by Lipoyl Dehydrogenase and Other Flavoenzymes in Vitro: implications for an enzymatic reduction of Fe(III) ions of the labile iron pool. *Journal of Biological Chemistry* 278:46403–46413.

Pfanner N, Warscheid B, Wiedemann N. 2019. Mitochondrial proteins: from biogenesis to functional networks. *Nature Reviews Molecular Cell Biology 2018 20:5* 20:267–284.

Pittis AA, Gabaldón T. 2016. Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. *Nature 2016 531:7592* 531:101–104.

Podar M, Gilmour CC, Brandt CC, Soren A, Brown SD, Crable BR, Palumbo A v., Somenahally AC, Elias DA. 2015. Global prevalence and distribution of genes and microorganisms involved in mercury methylation. *Sci Adv* 1(9):e1500675.

Poinar GO. 2009. Description of an early Cretaceous termite (Isoptera: Kalotermitidae) and its associated intestinal protozoa, with comments on their co-evolution. *Parasit Vectors* 2.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS One* 5:e9490.

Pyrih J, Žárský V, Fellows JD, Grosche C, Wloga D, Striepen B, Maier UG, Tachezy J. 2021. The iron-sulfur scaffold protein HCF101 unveils the complexity of organellar evolution in SAR, Haptista and Cryptista. *BMC Ecol Evol* 21:1–19.

Pyrihová E, Motyčková A, Voleman L, Wandyszewska N, Fišer R, Seydlová G, Roger A, Kolísko M, Doležal P. 2018. A single tim translocase in the mitosomes of *Giardia intestinalis* illustrates convergence of protein import machines in anaerobic eukaryotes. *Genome Biol Evol* 10:2813–2822.

Rada P, Makki A, Žárský V, Tachezy J. 2019. Targeting of tail-anchored proteins to *Trichomonas vaginalis* hydrogenosomes. *Mol Microbiol* 111:588–603.

Richter DJ, Berney C, Strassert JFH, Poh Y-P, Herman EK, Muñoz-Gómez SA, Wideman JG, Burki F, Vargas C de. 2022. EukProt: a database of genome-scale predicted proteins across the diversity of eukaryotes. *bioRxiv* 2020.06.30.180687.

Roger AJ, Muñoz-Gómez SA, Kamikawa R. 2017. The origin and diversification of mitochondria. *Current Biology* 27:R1177–R1192.

Saas J, Ziegelbauer K, von Haeseler A, Fast B, Boshart M. 2000. A developmentally regulated aconitase related to iron-regulatory protein-1 is localized in the cytoplasm and in the mitochondrion of *Trypanosoma brucei*. *J Biol Chem* 275:2745–2755.

Saga Y, Kawashima M, Sakai S, Yamazaki K, Kaneko M, Takahashi M, Sato N, Toyoda Y, Takase S, Nakano T, et al. 2020. Plant-specific domains and fragmented sequences imply non-canonical functions in plant aminoacyl-trna synthetases. *Genes (Basel)* 11:1–13.

Sakuraba H, Satomura T, Kawakami R, Yamamoto S, Kawarabayasi Y, Kikuchi H, Ohshima T. 2002. L-aspartate oxidase is present in the anaerobic hyperthermophilic archaeon *Pyrococcus horikoshii* OT-3: Characteristics and role in the de novo biosynthesis of nicotinamide adenine dinucleotide proposed by genome sequencing. *Extremophiles* 6:275–281.

Salas-Leiva DE, Tromer EC, Curtis BA, Jerlström-Hultqvist J, Kolisko M, Yi Z, Salas-Leiva JS, Gallot-Lavallée L, Williams SK, Kops GJPL, et al. 2021. Genomic analysis finds no evidence of canonical eukaryotic DNA processing complexes in a free-living protist. *Nat Commun* 12:6003.

Schut GJ, Adams MWW. 2009. The iron-hydrogenase of *Thermotoga maritima* utilizes ferredoxin and NADH synergistically: A new perspective on anaerobic hydrogen production. *J Bacteriol* 191(13):4451-4457.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E v., Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31:3210–3212.

Slamovits CH, Keeling PJ. 2006. Pyruvate-phosphate dikinase of oxymonads and Parabasalia and the evolution of pyrophosphate-dependent glycolysis in anaerobic eukaryotes. *Eukaryot Cell* 5:148–154.

Smith AC, Blackshaw JA, Robinson AJ. 2012. MitoMiner: A data warehouse for mitochondrial proteomics data. *Nucleic Acids Res* 40:D1160-7.

Smith AC, Robinson AJ. 2018. MitoMiner v4.0: an updated database of mitochondrial localization evidence, phenotypes and diseases. *Nucleic Acids Res* 47: D1225–D1228.

Song L, Shankar DS, Florea L. 2016. Rascaf: Improving Genome Assembly with RNA Sequencing Data. *Plant Genome* 9:0.

Spalding MD, Prigge ST. 2010. Lipoic acid metabolism in microbial pathogens. *Microbiol Mol Biol Rev* 74:200–228.

Stairs CW, Eme L, Brown MW, Mutsaers C, Susko E, Dellaire G, Soanes DM, van der Giezen M, Roger AJ. 2014. A SUF Fe-S cluster biogenesis system in the mitochondrion-related organelles of the anaerobic protist *Pygsuia*. *Current Biology* 24:1176–1186.

Stairs CW, Leger MM, Roger AJ. 2015. Diversity and origins of anaerobic metabolism in mitochondria and related organelles. *Philos Trans R Soc Lond B Biol Sci* 370:20140326.

Stairs CW, Táborský P, Salomaki ED, Kolisko M, Pánek T, Eme L, Hradilová M, Vlček Č, Jerlström-Hultqvist J, Roger AJ, et al. 2021. Anaeramoebae are a divergent lineage of eukaryotes that shed light on the transition from anaerobic mitochondria to hydrogenosomes. *Curr Biol* 31:5605-5612.e5.

Stamatakis A. 2006. The RAxML 7.0.4 Manual. *Bioinformatics* 22(21):2688–2690.

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

Stanke M, Waack S. 2003. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 19:ii215–ii225.

Stechmann A, Baumgartner M, Silberman JD, Roger AJ. 2006. The glycolytic pathway of *Trimastix pyriformis* is an evolutionary mosaic. *BMC Evol Biol* 6:101.

Stechmann A, Hamblin K, Pérez-Brocal V, Gaston D, Richmond GSS, Van Der Giezen M, Clark CGG, Roger AJ, Giezen M Van Der, Clark CGG, et al. 2008. Report Organelles in

Blastocystis that Blur the Distinction between Mitochondria and Hydrogenosomes. *Current Biology* 18:580–585.

Stephens TG, Bhattacharya D, Ragan MA, Chan CX. 2016. PhySortR: a fast, flexible tool for sorting phylogenetic trees in R. *PeerJ* 12;4:e2038.

Stipanuk MH. 1986. Metabolism of sulfur-containing amino acids. *Annu Rev Nutr* 6:179-209.

Swearingen JW, Fuentes DE, Araya MA, Plishker MF, Saavedra CP, Chasteen TG, Vásquez CC. 2006. Expression of the ubiE gene of Geobacillus stearothermophilus V in *Escherichia coli* K-12 mediates the evolution of selenium compounds into the headspace of selenite- and selenate-amended cultures. *Appl Environ Microbiol* 72:963–967.

Tachezy J, Doležal P. 2007. Iron–Sulfur Proteins and Iron–Sulfur Cluster Assembly in Organisms with Hydrogenosomes and Mitosomes. In: Origin of Mitochondria and Hydrogenosomes. Berlin, Heidelberg: Springer Berlin Heidelberg. p. 105–133.

Tang S, Lomsadze A, Borodovsky M. 2015. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res* 43:1–10.

Tarailo-Graovac M, Chen N. 2009. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. In: Current Protocols in Bioinformatics. Vol. 25. Hoboken, NJ, USA: John Wiley & Sons, Inc. p. 4.10.1-4.10.14.

Thiergart T, Landan G, Schenk M, Dagan T, Martin WF. 2012. An evolutionary network of genes present in the eukaryote common ancestor polls genomes on eukaryotic and mitochondrial origin. *Genome Biol Evol* 4:466–485.

Treitli SC., Kolisko M, Husník F, Keeling PJ, Hampl V. 2019. Revealing the metabolic capacity of Streblomastix strix and its bacterial symbionts using single-cell metagenomics. *Proceedings of the National Academy of Sciences* 116:19675–19684.

Treitli SC, Kotyk M, Yubuki N, Jirounková E, Vlasáková J, Smejkalová P, Šípek P, Čepička I, Hampl V. 2018. Molecular and Morphological Diversity of the Oxymonad Genera *Monocercomonoides* and *Blattamonas* gen. nov. *Protist* 169:744–783.

Treitli SC, Peña-Diaz P, Hałakuc P, Karnkowska A, Hampl V. 2021. High quality genome assembly of the amitochondriate eukaryote Monocercomonoides exilis. *Microb Genom* 7(12):000745.

Tsaousis AD, Kunji ERS, Goldberg A v, Lucocq JM, Hirt RP, Embley TM. 2008. A novel route for ATP acquisition by the remnant mitochondria of *Encephalitozoon cuniculi*. *Nature* 453:553–556.

Vacek V, Novák LVF, Treitli SC, Táborský P, Čepička I, Kolísko M, Keeling PJ, Hampl V, Ruiz-Trillo I. 2018. Fe-S Cluster Assembly in Oxymonads and Related Protists. *Mol Biol Evol* 35(11):2712-2718.

Valasatava Y, Rosato A, Banci L, Andreini C. 2016. MetalPredator: A web server to predict iron-sulfur cluster binding proteomes. *Bioinformatics* 32(18): 2850–2852.

Vargová R, Hanousková P, Salamonová J, Žihala D, Silberman JD, Eliáš M, Čepička I. 2022. Evidence for an Independent Hydrogenosome-to-Mitosome Transition in the CL3 Lineage of Fornicates. *Front Microbiol* 13:1229.

Vargová R, Wideman JG, Derelle R, Klimeš V, Kahn RA, Dacks JB, Eliáš M. 2021. A Eukaryote-Wide Perspective on the Diversity and Evolution of the ARF GTPase Protein Family. *Genome Biol Evol* 13(8):evab157.

Vlahou G, Eliáš M, von Kleist-Retzow JC, Wiesner RJ, Rivero F. 2011. The Ras related GTPase Miro is not required for mitochondrial transport in *Dictyostelium discoideum*. *Eur J Cell Biol* 90:342–355.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. 2014. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9:e112963.

Williams K, Lowe PN, Leadlay PF. 1987. Purification and characterization of pyruvate: ferredoxin oxidoreductase from the anaerobic protozoon *Trichomonas vaginalis*. *Biochem J* 246:529–536.

Xia L, Björnstedt M, Nordman T, Eriksson LC, Olsson JM. 2001. Reduction of ubiquinone by lipoamide dehydrogenase. An antioxidant regenerating pathway. *Eur J Biochem* 268:1486–1490.

Xu F, Jerlström-Hultqvist J, Einarsson E, Ástvaldsson Á, Svärd SG, Andersson JO. 2014. The Genome of Spironucleus salmonicida Highlights a Fish Pathogen Adapted to Fluctuating Environments. *PLoS Genet* 10:e1004053.

Yagi T, Shounaka M, Yamamoto S. 1990. Distribution of aspartate aminotrasnsferase activity in yeasts, and purification and characterization of mitochondrial and cytosolic isoenzymes from *Rhodotorula marina*. *J Biochem* 107:151–159.

Yazaki E, Kume K, Shiratori T, Eglit Y, Tanifuji G, Harada R, Simpson AGB, Ishida KI, Hashimoto T, Inagaki Y. 2020. Barthelonids represent a deep-branching metamonad clade with mitochondrion-related organelles predicted to generate no ATP. *Proceedings of the Royal Society B* 287:20201538.

Zhang Q, Táborský P, Silberman JD, Pánek T, Čepička I, Simpson AGB. 2015. Marine Isolates of *Trimastix marina* Form a Plesiomorphic Deep-branching Lineage within Preaxostyla, Separate from Other Known Trimastigids (*Paratrimastix* n. gen.). *Protist* 166:468–491.

Zhou X, Yuan Y, Yang Y, Rutzke M, Thannhauser TW, Kochian L v., Li L. 2009. Involvement of a broccoli COQ5 methyltransferase in the production of volatile selenium compounds. *Plant Physiol* 151:528–540.

Zítek J, Füssy Z, Treitli SC, Peña-Diaz P, Vaitová Z, Zavadska D, Harant K, Hampl V. 2022. Reduced mitochondria provide an essential function for the cytosolic methionine cycle. *bioRxiv* 2022.04.01.486701.

Zubáčová Z, Novák L, Bublíková J, Vacek V, Fousek J, Rídl J, Tachezy J, Doležal P, Vlček Č, Hampl V. 2013. The Mitochondrion-Like Organelle of *Trimastix pyriformis* Contains the Complete Glycine Cleavage System. *PLoS One* 8:e55417.

FIGURE LEGENDS

Figure 1. Distribution of orthologous groups (OGs) and other discussed features among metamonads. **A)** Schematic phylogenetic tree of Preaxostyla within Metamonada showing the distribution of categories of OGs, lateral gene transfer events, and key biological features mapped to nodes. Core set: estimated relaxed core gene set for Metamonada; increment: clade-specific increment of the relaxed core gene set; specific: unique OGs for the external leaves (species); new: novel or extremely divergent genes detected in the species or lineage; LGT: lateral gene transfer events; SUF: iron-sulfur cluster assembly SUF system; L2: glycine cleavage system protein L2; FF: flavodoxin-ferredoxin fusion protein; FH: fused hydrogenase; FC: folate cycle; MC: methionine cycle; PotE: b(0,+)-type amino acid transporter; SPNS: sphingolipid transporter; RFC: Reduced Folate Carrier; TauE: sulfite exporter TauE/SafE family protein; SP: sugar porters; Hg: mercury methylation; Se: selenium volatilization; pC: *p*-cresol synthesis; ADI: arginine deiminase; ADIp: arginine deiminase pathway; CK: carbamate kinase. **B)** Venn diagram showing the distribution of OGs among the three main analyzed lineages of Metamonada. **C)** Venn diagram showing the distribution of OGs among the three lineages of Preaxostyla. **D)** Venn diagram showing the distribution of OGs among the three studied species of Oxymonadida. The IDs and functional annotation of OGs is given in Supplementary file 2. The IDs and functional annotation of LGT events is given in Supplementary file 3.

Figure 2. Schematic representation of the structure of HgcAB protein in *P. pyriformis* and its comparison to the prokaryotic HgcA-HgcB operon. The upper panel gives the overview of the full alignment, the middle panel zooms on the primary structure of the functionally most important regions, and the lower panel displays the complete domain structure.

Figure 3. Complement of Golgi-associated proteins in Preaxostyla. This Coulson plot shows the set of proteins present in the Preaxostyla predicted proteomes. Empty segments denote failure to identity a candidate orthologue, while filled segments denote success, with paralogue numbers inset as relevant. Candidate proteins are identified by homology-searching and verified by phylogenetics as relevant. Details are given in Supplementary file 6.

Figure 4. Hypothetical energy metabolism in Preaxostyla. The glycolysis reactions are simplified. Bold outline indicates alternative glycolytic enzymes. Abbreviations and Enzyme Commission

numbers are given in Supplementary file 6. Presence of the enzymes in Preaxostyla data sets is indicated by a color code.

Figure 5. Phylogenetic relationship among hydrogenases of Preaxostyla. A) Schematic representation of domain architectures of hydrogenases among Preaxostyla species; domain architecture is indicated. HydA: [FeFe] hydrogenase; CysJ: NADPH-dependent sulfite reductase; NuoG: NADH-quinone oxidoreductase. B) Detailed view on the part of the tree comprising the clade of long *P. pyriformis* hydrogenases. The domain architecture of the proteins is indicated by colour bars. The full tree is given in Supplementary file 4.

Figure 6. Summary of the searches for proteins physically associated with MROs. In the upper panel, the numbers of candidates recovered by the four strategies for the five Preaxostyla species (colour-coded) are summarized. Detailed information on the searches and candidates is given in Supplementary file 9. In the lower panel, the mitochondrion hallmark proteins detected in the data sets of *P. pyriformis* and *T. marina* are summarized. No such candidate was recovered for any oxymonad representative.

SUPPLEMENTARY FIGURE LEGENDS

Figure S1. Hypothetical map of amino acid metabolism in *P. pyriformis*. Brown color indicates enzymes possibly involved in amino acid biosynthesis pathways. Red color indicates enzymes possibly involved in ATP production. Note that some of the connections between metabolites correspond to the mere transfer of the amino group rather than conversion of the carbon backbone of the molecule. Abbreviations and Enzyme Commission numbers are given in Supplementary file 6.

Figure S2. Hypothetical map of amino acid metabolism in *T. marina*. Brown color indicates enzymes possibly involved in amino acid biosynthesis pathways. Red color indicates enzymes possibly involved in ATP production. Abbreviations and Enzyme Commission numbers are given in Supplementary file 6.

Figure S3. Hypothetical map of amino acid metabolism in *M. exilis*, *B. nauphoetae*, and *S. strix*. Brown color indicates enzymes possibly involved in amino acid biosynthesis pathways. Red color

indicates enzymes possibly involved in ATP production. Abbreviations and Enzyme Commission numbers are given in Supplementary file 6.

Figure S4. Venn diagram showing the distribution of orthologous groups (OGs) of Fe-S cluster-containing proteins among the species of Preaxostyla. The identity of the OGs and of the component proteins are provided in Supplementary file 8.

TABLES

Table 1. General features of the Preaxostyla genomes discussed in this study

Sample	Scaffolds	Total length (bp)	N50 (kbp)	Completeness (%; BUSCO odb9)	G+C content (%)	Protein-Coding Loci	NCBI BioProject	NCBI BioSample
P. pyriformis	633	56,627,582	276.6	82.1	60.94	13,466	PRJNA903905	SAMN31819772
B. nauphoetae	879	88,537,989	199.5	76.6	44.96	25,221	PRJNA887011	SAMN31149887
M. exilis	2,092	74,712,536	71.44	75.3	36.8	16,768	PRJNA304271	SAMN04297179
S. strix	50,889	152,152,197	5.18	69.6	26.6	56,706	PRJNA524138	SAMN10998475

FIGURES

Figure 1

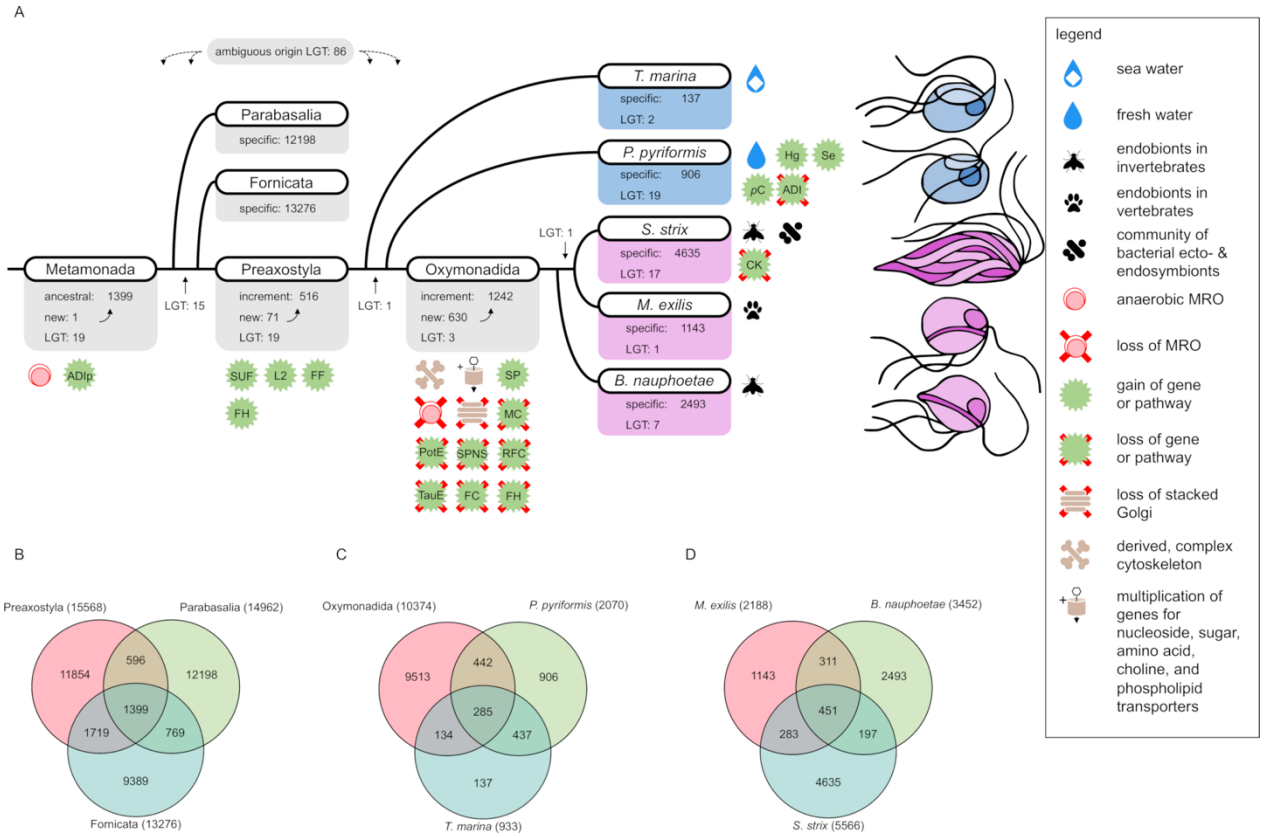


Figure 2

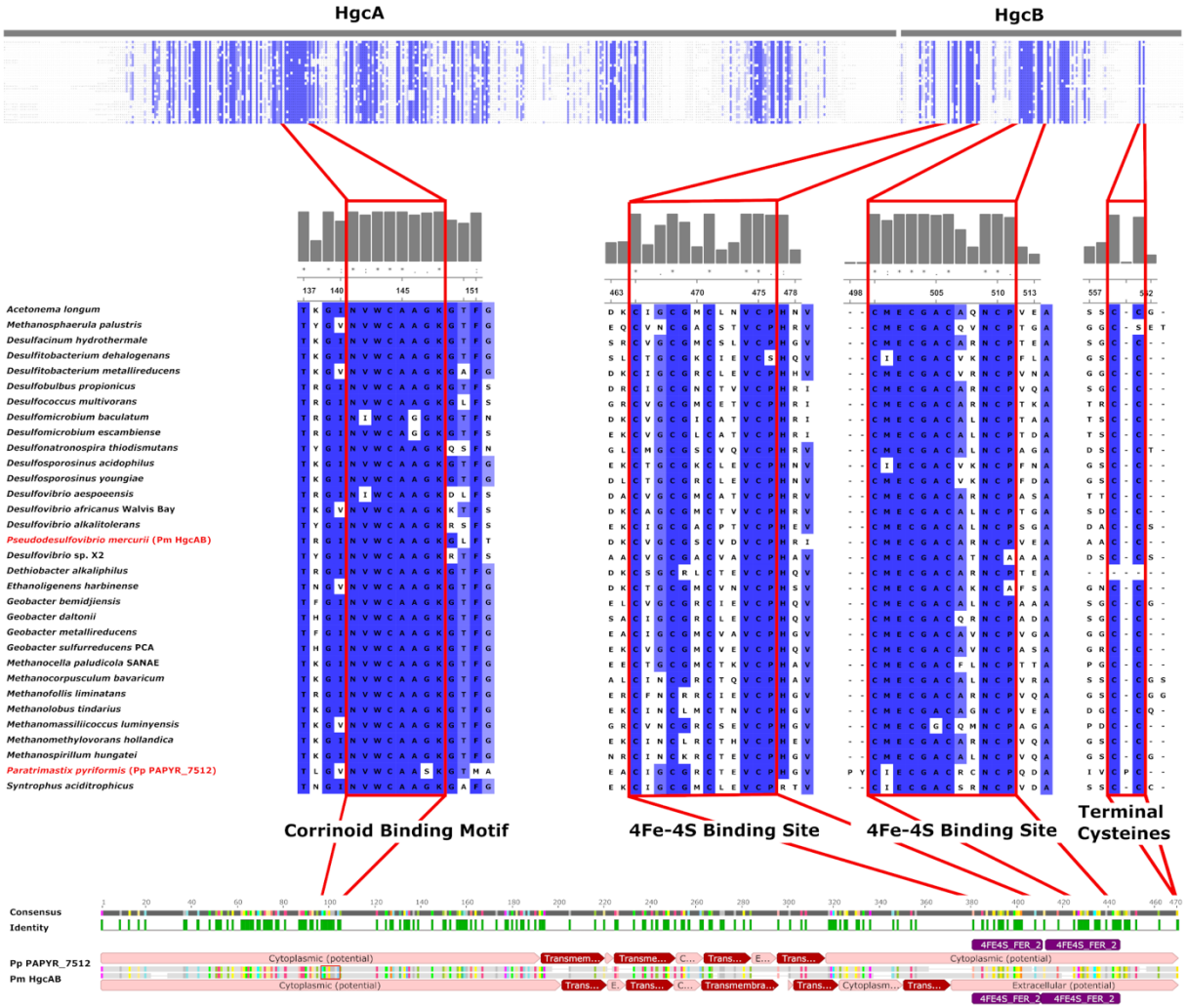

HgcA
HgcB
Acetonema longum
Methanosphaerula palustris
Desulfacinum hydrothermale
Desulfitobacterium dehalogenans
Desulfitobacterium metallireducens
Desulfobulbus propionicus
Desulfococcus multivorans
Desulfomicrobium baculatum
Desulfomicrobium escambiense
Desulfonatronospira thiodismutans
Desulfosporosinus acidophilus
Desulfosporosinus youngiae
Desulfovibrio aespoeensis
Desulfovibrio africanus Walvis Bay
Desulfovibrio alkalitolerans
Pseudodesulfovibrio mercurii (Pm HgcAB)
Desulfovibrio sp. X2
Dethiobacter alkaliphilus
Ethanoligenes harbinense
Geobacter bemidjiensis
Geobacter daltonii
Geobacter metallireducens
Geobacter sulfurreducens PCA
Methanocella paludicola SANAE
Methanocorpusculum bavaricum
Methanofollis liminatans
Methanolobus tindarius
Methanomassiliicoccus luminyensis
Methanomethylovorans hollandica
Methanospirillum hungatei
Paratrimastix pyriformis (Pp PAPYR_7512)
Syntrophus aciditrophicus
Corrinoid Binding Motif
4Fe-4S Binding Site
4Fe-4S Binding Site
Terminal Cysteines
Consensus
Identity
Pp PAPYR_7512
Pm HgcAB
Cytoplasmic (potential)
Extracellular (potential)
4FE4S_FER_2
4FE4S_FER_2

Figure 3

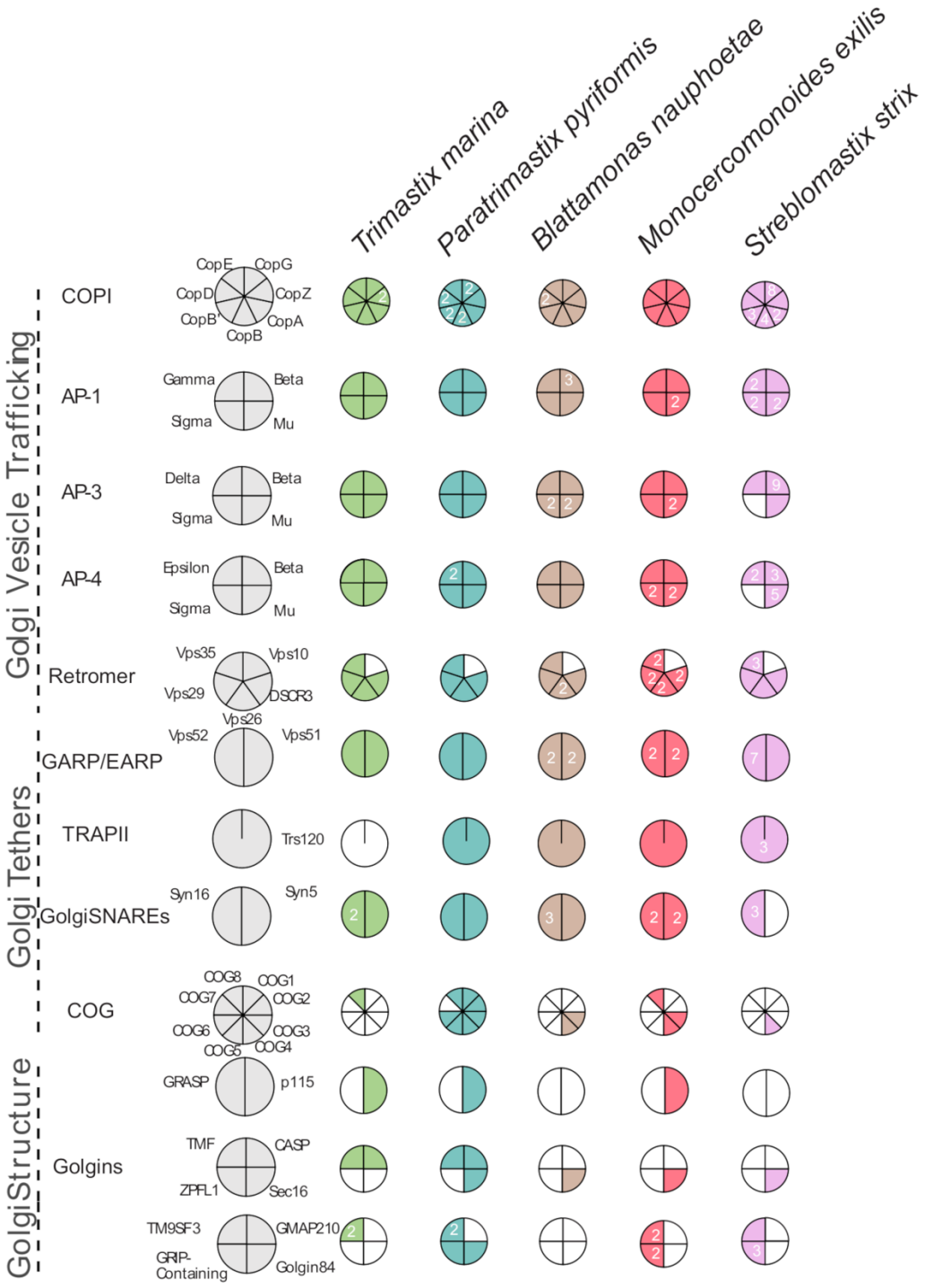

Trimastix marina
Paratrimastix pyriformis
Blattamonas nauphoetae
Monocercomonoides exilis
Streblomastix strix
Golgi Vesicle Trafficking
Golgi Tethers
Golgi Structure
COPI
CopE
CopG
CopD
CopZ
CopB'
CopA
CopB
AP-1
Gamma
Beta
Sigma
Mu
AP-3
Delta
Beta
Sigma
Mu
AP-4
Epsilon
Beta
Sigma
Mu
Retromer
Vps35
Vps10
Vps29
DSCR3
GARP/EARP
Vps26
Vps52
Vps51
TRAPII
Trs120
GolgiSNAREs
Syn16
Syn5
COG
COG8
COG1
COG7
COG2
COG6
COG3
COG5
COG4
GRASP
p115
Golgins
TMF
CASP
ZPFL1
Sec16
TM9SF3
GMAP210
GRIP-Containing
Golgin84

Figure 4

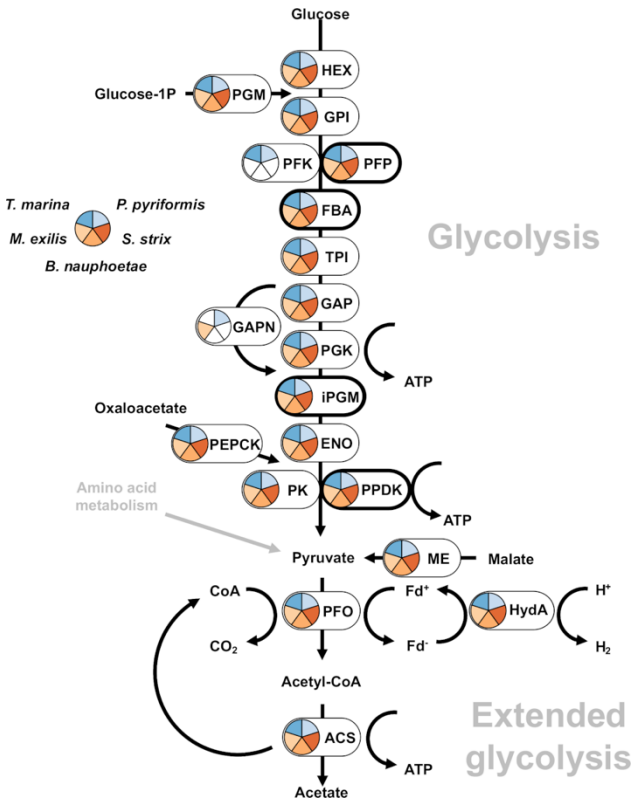
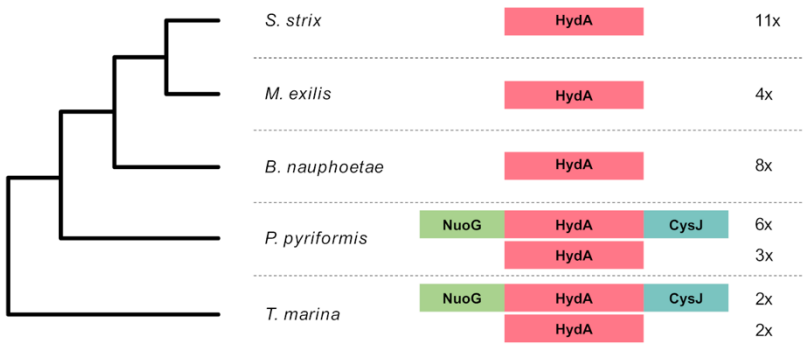

Glucose
HEX
Glucose-1P
PGM
GPI
PFK
PFP
T. marina
P. pyriformis
M. exilis
S. strix
B. nauphoetae
FBA
Glycolysis
TPI
GAP
GAPN
PGK
ATP
iPGM
Oxaloacetate
PEPCK
ENO
Amino acid metabolism
PK
PPDK
ATP
Pyruvate
ME
Malate
CoA
PFO
Fd+
HydA
H+
CO2
Fd-
H2
Acetyl-CoA
ACS
ATP
Acetate
Extended glycolysis

Figure 5

A


S. strix
HydA
11x
M. exilis
HydA
4x
B. nauphoetae
HydA
8x
P. pyriformis
NuoG
HydA
CysJ
6x
HydA
3x
T. marina
NuoG
HydA
CysJ
2x
HydA
2x

B

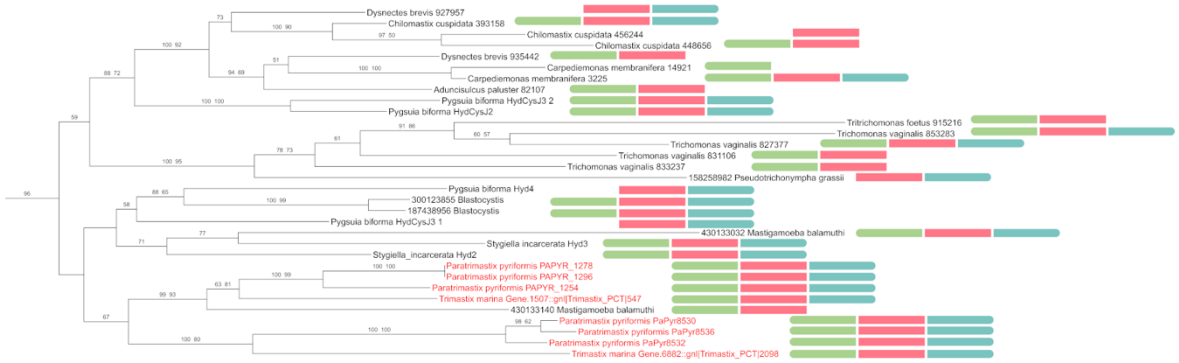
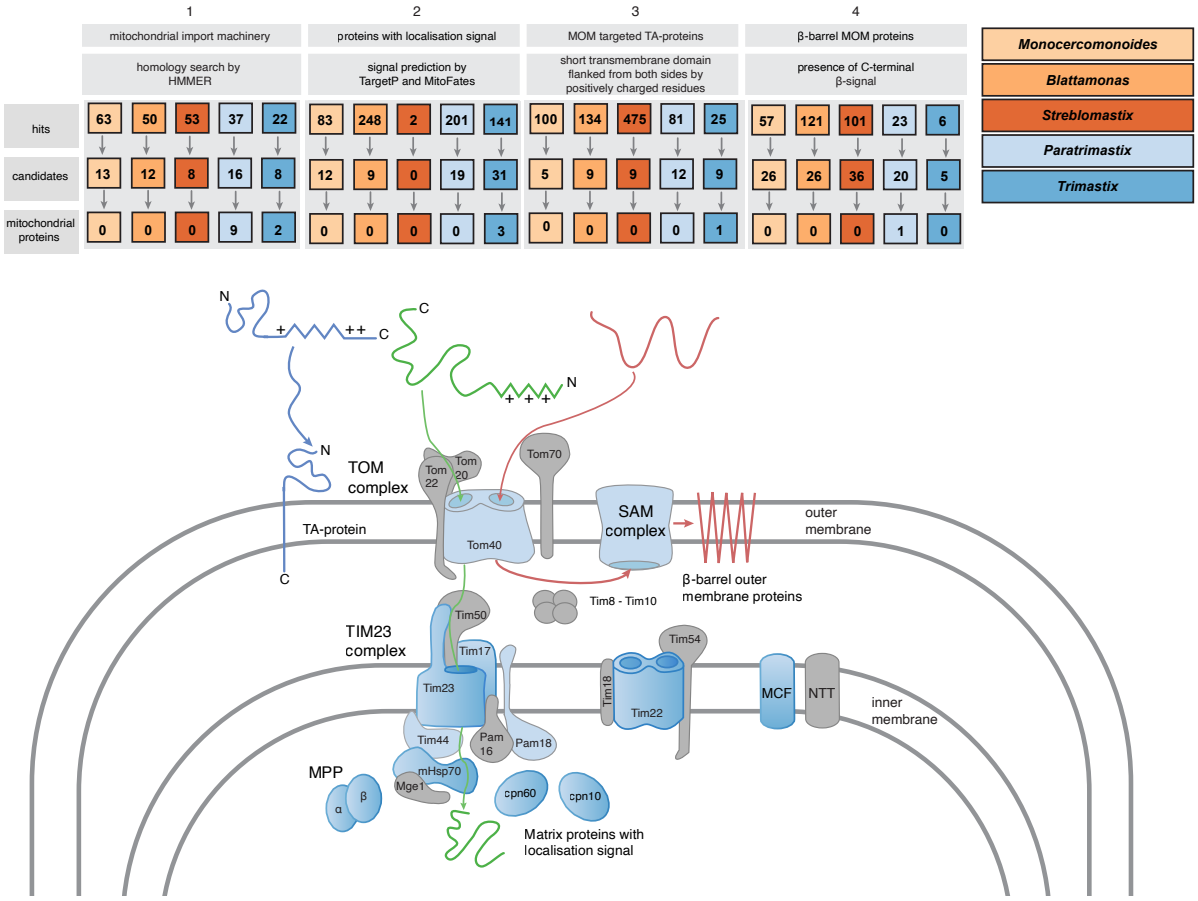

Dysnectes brevis 927957
Chilomastix cuspidata 393158
Chilomastix cuspidata 456244
Chilomastix cuspidata 448656
Dysnectes brevis 935442
Carpediemonas membranifera 14921
Carpediemonas membranifera 3225
Aduncisulcus paluster 82107
Pygsuia biforma HydCysJ3 2
Pygsuia biforma HydCysJ2
Tritrichomonas foetus 915216
Trichomonas vaginalis 853283
Trichomonas vaginalis 827377
Trichomonas vaginalis 831106
Trichomonas vaginalis 833237
158258982 Pseudotrichonympha grassii
Pygsuia biforma Hyd4
300123855 Blastocystis
187438956 Blastocystis
Pygsuia biforma HydCysJ3 1
430133032 Mastigamoeba balamuthi
Stygiella incarcerata Hyd3
Stygiella_incarcerata Hyd2
Paratrimastix pyriformis PAPYR_1278
Paratrimastix pyriformis PAPYR_1296
Paratrimastix pyriformis PAPYR_1254
Trimastix marina Gene.1507::gnl|Trimastix_PCT|547
430133140 Mastigamoeba balamuthi
Paratrimastix pyriformis PaPyr8530
Paratrimastix pyriformis PaPyr8536
Paratrimastix pyriformis PaPyr8532
Trimastix marina Gene.6882::gnl|Trimastix_PCT|2098

Figure 6


1
mitochondrial import machinery
homology search by HMMER
2
proteins with localisation signal
signal prediction by TargetP and MitoFates
3
MOM targeted TA-proteins
short transmembrane domain flanked from both sides by positively charged residues
4
β-barrel MOM proteins
presence of C-terminal β-signal
hits
candidates
mitochondrial proteins
63 50 53 37 22
13 12 8 16 8
0 0 0 9 2
83 248 2 201 141
12 9 0 19 31
0 0 0 0 3
100 134 475 81 25
5 9 9 12 9
0 0 0 0 1
57 121 101 23 6
26 26 36 20 5
0 0 0 1 0
Monocercomonoides
Blattamonas
Streblomastix
Paratrimastix
Trimastix
N
C
TOM complex
Tom 22
Tom 20
Tom40
Tom70
TA-protein
SAM complex
outer membrane
β-barrel outer membrane proteins
Tim8 - Tim10
TIM23 complex
Tim50
Tim17
Tim23
Tim44
Pam 16
Pam18
Tim18
Tim54
Tim22
MCF
NTT
inner membrane
MPP
α
β
mHsp70
Mge1
cpn60
cpn10
Matrix proteins with localisation signal

SUPPLEMENTARY FIGURES

Figure S1

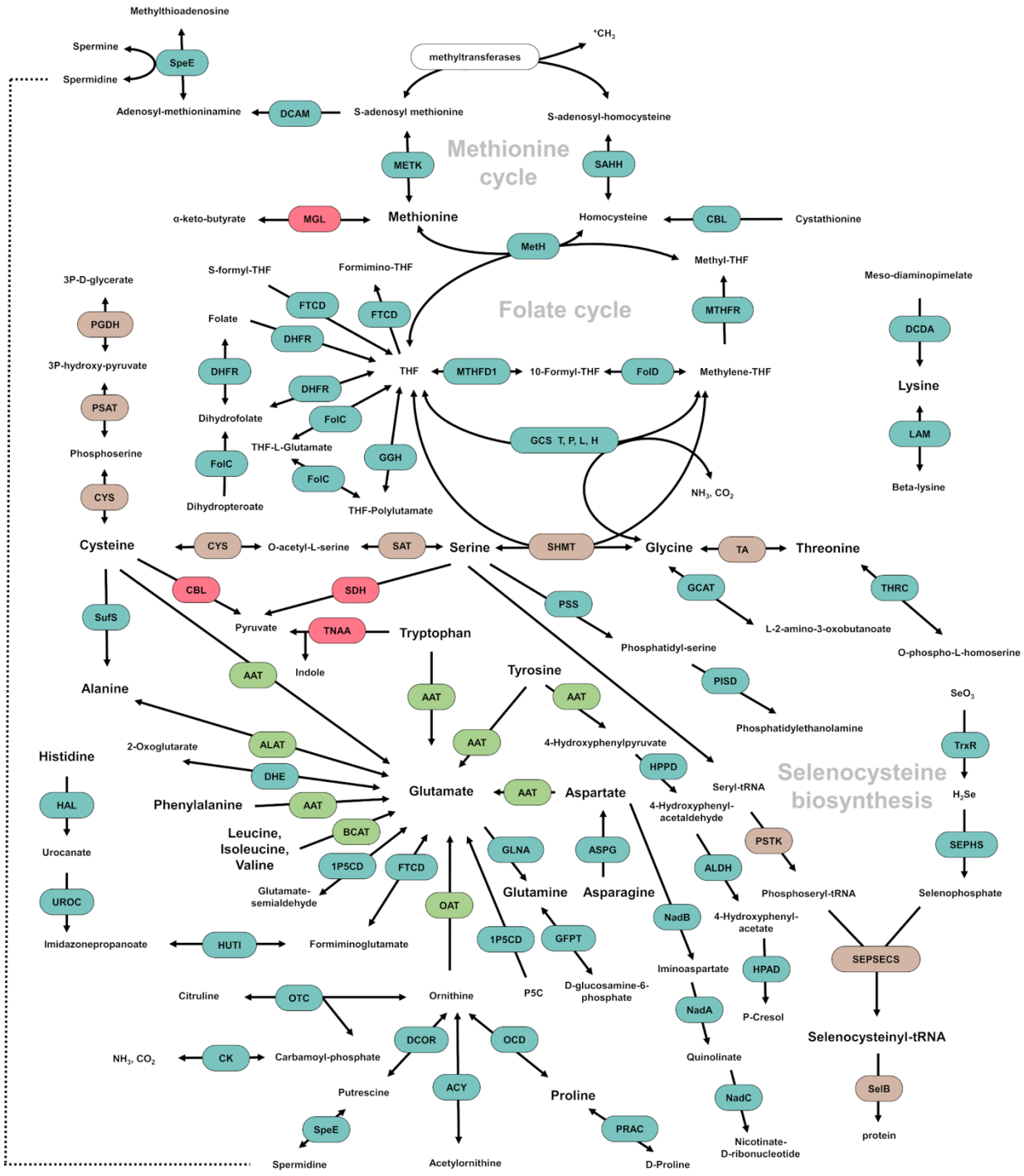


Figure S2

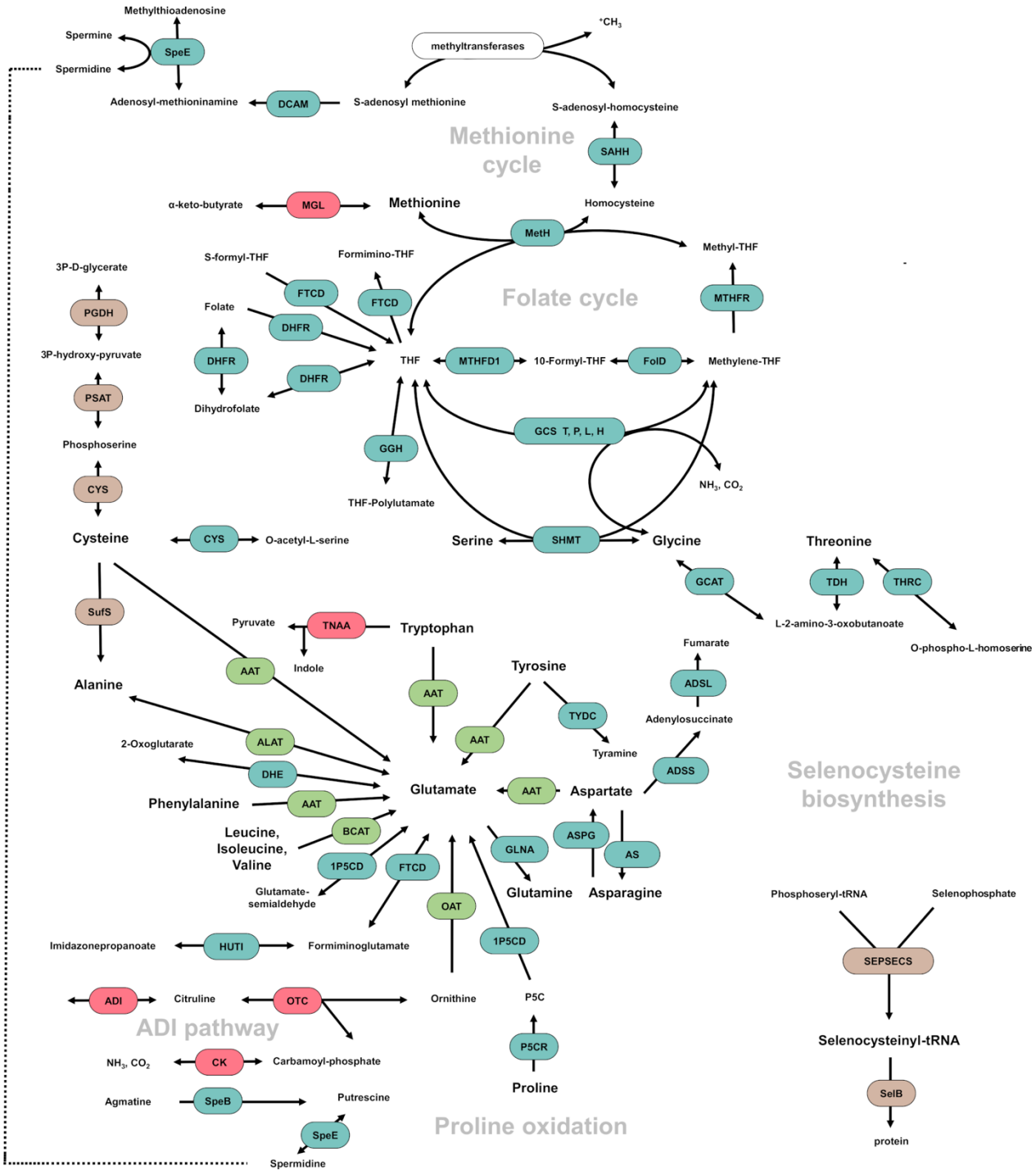

Methylthioadenosine
Spermine
SpeE
Spermidine
methyltransferases
+CH3
Adenosyl-methioninamine
DCAM
S-adenosyl methionine
S-adenosyl-homocysteine
Methionine cycle
SAHH
α-keto-butyrate
MGL
Methionine
Homocysteine
MetH
Methyl-THF
S-formyl-THF
Formimino-THF
3P-D-glycerate
FTCD
FTCD
Folate cycle
MTHFR
PGDH
Folate
DHFR
3P-hydroxy-pyruvate
DHFR
THF
MTHFD1
10-Formyl-THF
FolD
Methylene-THF
DHFR
PSAT
Dihydrofolate
GCS T, P, L, H
Phosphoserine
GGH
NH3, CO2
CYS
THF-Polylutamate
Cysteine
CYS
O-acetyl-L-serine
Serine
SHMT
Glycine
Threonine
GCAT
TDH
THRC
SufS
Pyruvate
TNAA
Tryptophan
L-2-amino-3-oxobutanoate
O-phospho-L-homoserine
Fumarate
AAT
Indole
Tyrosine
ADSL
Alanine
AAT
TYDC
Adenylosuccinate
2-Oxoglutarate
ALAT
AAT
Tyramine
DHE
ADSS
Selenocysteine biosynthesis
Phenylalanine
AAT
Glutamate
AAT
Aspartate
Leucine, Isoleucine, Valine
BCAT
ASPG
AS
GLNA
1P5CD
FTCD
Glutamate-semialdehyde
Glutamine
Asparagine
Phosphoseryl-tRNA
Selenophosphate
OAT
Imidazonepropanoate
HUTI
Formiminoglutamate
1P5CD
SEPSECS
ADI
Citruline
OTC
Ornithine
P5C
ADI pathway
Selenocysteinyl-tRNA
P5CR
NH3, CO2
CK
Carbamoyl-phosphate
Proline
SelB
Agmatine
SpeB
Putrescine
Proline oxidation
SpeE
protein
Spermidine

Figure S3

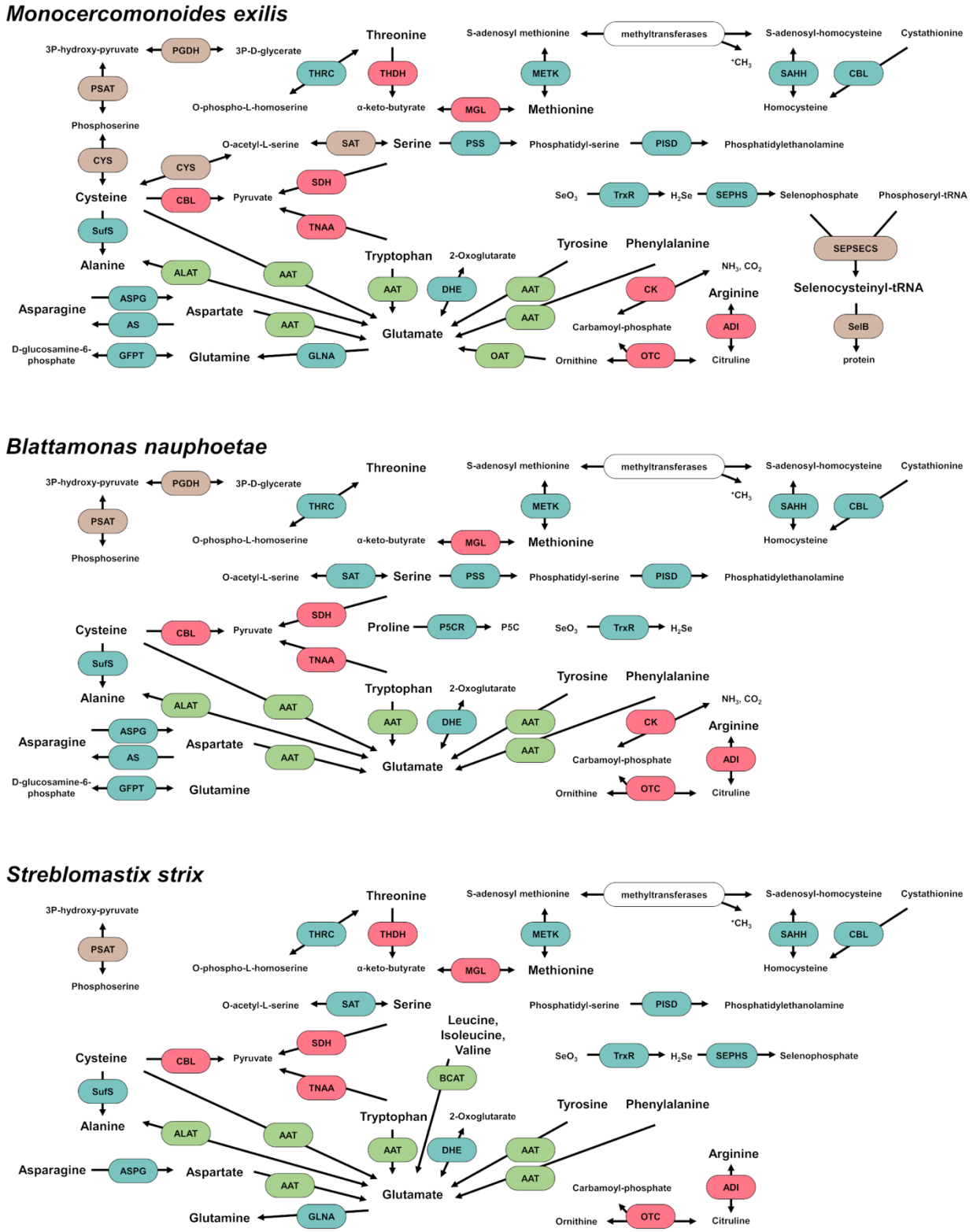
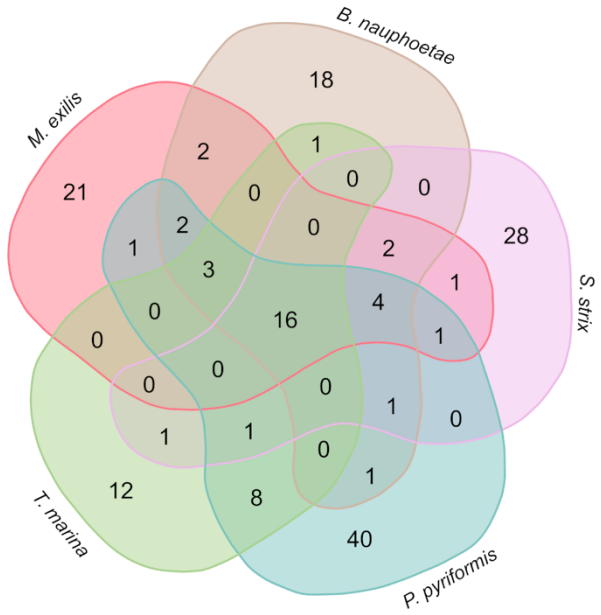

Monocercomonoides exilis
3P-hydroxy-pyruvate
PGDH
3P-D-glycerate
PSAT
Phosphoserine
CYS
Cysteine
SufS
Alanine
ALAT
Asparagine
ASPG
AS
Aspartate
D-glucosamine-6-phosphate
GFPT
Glutamine
GLNA
Threonine
THRC
THDH
O-phospho-L-homoserine
α-keto-butyrate
MGL
Methionine
S-adenosyl methionine
methyltransferases
S-adenosyl-homocysteine
Cystathionine
METK
SAHH
CBL
Homocysteine
O-acetyl-L-serine
SAT
Serine
PSS
Phosphatidyl-serine
PISD
Phosphatidylethanolamine
CYS
CBL
Pyruvate
SDH
TNAA
AAT
Tryptophan
2-Oxoglutarate
DHE
Glutamate
Tyrosine
Phenylalanine
CK
Carbamoyl-phosphate
OAT
Ornithine
OTC
Citruline
Arginine
ADI
SeO3
TrxR
H2Se
SEPHS
Selenophosphate
Phosphoseryl-tRNA
SEPSECS
Selenocysteinyl-tRNA
SelB
protein
Blattamonas nauphoetae
Proline
P5CR
P5C
Leucine, Isoleucine, Valine
BCAT

Figure S4


B. nauphoetae
M. exilis
S. strix
T. marina
P. pyriformis
18
21
28
12
40
2
1
0
0
0
2
1
3
0
2
1
0
16
4
1
0
0
0
0
1
0
1
1
0
1
8
40