Heterotrophic euglenid *Rhabdomonas costata* resembles its phototrophic relatives in many aspects of molecular and cell biology

Petr Soukal[1#], Štěpánka Hrdá[1#], Anna Karnkowska[2], Rafał Milanowski[2], Jana Szabová[1], Miluše Hradilová[3], Hynek Strnad[3], Čestmír Vlček[3], Ivan Čepička[4] and Vladimír Hampl[1]*

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

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

[3]Institute of Molecular Genetics, Academy of Sciences of the Czech Republic, Prague, Czech Republic.

[4]Department of Zoology, Charles University, Prague, Czech Republic.

[#]Authors contributed equally to the manuscript

*Author for Correspondence: Vladimír Hampl, Department of Parasitology, BIOCEV, Charles University, Vestec, Czech Republic, +420325873911, vlada@natur.cuni.cz

Abstract

Euglenids represent a group of protists with diverse modes of feeding. To date, only a partial genomic sequence of *Euglena gracilis* and transcriptomes of several phototrophic and secondarily osmotrophic species are available, while primarily heterotrophic euglenids are seriously undersampled. In this work, we begin to fill this gap by presenting genomic and transcriptomic drafts of a primary osmotroph, *Rhabdomonas costata*. The current genomic assembly length of 100 Mbp is 14× smaller than that of *E. gracilis*. Despite being too fragmented for comprehensive gene prediction, comparison of the transcriptomic and genomic data revealed features of its introns, including several candidates for nonconventional introns. 16 % of transcripts bear a recognizable partial splice leader sequence. A set of 39,585 putative *R. costata* proteins were predicted from the transcriptome. Annotation of the mitochondrial core metabolism provides the first data on the facultatively anaerobic mitochondrion of *R. costata*, which in most respects resembles the mitochondrion of *E. gracilis* with certain level of streamlining. *R. costata* synthesises heme by a mitochondrial-cytoplasmatic C4 pathway with enzymes orthologous to those found in *E. gracilis*. The low percentage of green algae-affiliated genes, supports the ancestrally osmotrophic status of this species.

Keywords: Mitochondrial metabolism, nonconventional introns, genomic draft, transcriptome, heme synthesis, phylogeny

Introduction

Euglenids are a species-rich group (>1,500 described species) of unicellular eukaryotes[1] classified into the phylum Euglenozoa and defined by both ultrastructural and molecular features. Generally, they possess one or two emergent flagella inserted in a paraflagellar pocket and reinforced by a paraflagellar rod. The surface of their cells is formed by a distinctive pellicle consisting of three layers – the cytoplasmic membrane, a proteinaceous belt supported by microtubules, and the vesicles of the endoplasmic reticulum. The pellicle enables some euglenids to move in a characteristic manner by undulated shifts in the shape of the cell, which is referred to as metaboly or euglenoid movement. Storage of carbohydrates as paramylon, a β-1,3-glucan, is unique to euglenids[1].

All major types of eukaryotic nutrition are present in Euglenida – phagotrophy (eukaryovory and bacteriovory), osmotrophy, and phototrophy. Phagotrophic euglenids form several clades in the phylogenetic tree of this group, including several deepest branches[2]. Phototrophic euglenids (*Euglenophyceae*) which arose from one of these clades are, by all means, the best-studied subgroup. Their cell contains a triple membrane-bound plastid derived from a secondary endosymbiosis with a green alga. Since *Euglenophyceae* certainly constitute a clade, it is assumed that this endosymbiosis occurred in the clade's exclusive common ancestor[3]; however, early acquisition of a plastid in the Euglenozoa lineage has also been proposed to account for plant-like traits in trypanosomatids[4]. Although all known members of *Euglenophyceae* contain plastids, six species have lost the ability to photosynthesize and have become secondarily osmotrophic, of which the best-known is *Euglena longa* (formerly known as *Astasia longa*)[5]. Primarily osmotrophic euglenids form a distinct clade Aphagea[6] branching within phagotrophic euglenids[2]. Members of the order *Rhabdomonadales* are distinguished from other members of the osmotrophic clade (i.e., *Distigma*, *Astasia*) by their lack of euglenoid movement.

Groups related to Euglenida are the poorly studied marine Symbiontida, Diplonemea, which despite being largely unknown have surprised scientists by their massive abundance in ocean waters[7,8], and Kinetoplastea, which comprise many infamous parasites, among which species of the genera

Trypanosoma and *Leishmania* are some of the most studied protists[9]. Currently, mitochondrial and nuclear genomes of 50 species and strains of kinetoplastids have been sequenced[10], but besides the kinetoplastids there are no complete and well-annotated nuclear genome sequences available for Euglenozoa. Partial genome sequences of marine diplonemids have been obtained by single-cell approaches[7]. In Euglenida, genomic studies have as yet covered relatively well the iconic model flagellate *Euglena gracilis*, these yielding the chloroplast and mitochondrial genomes[11,12] and a very fragmented and unannotated nuclear genome[13]. Genomic data set from the remaining euglenids comprise exclusively the chloroplast genomes, 30 of which have been published until present[11,14–16]. Recently, the proteomes of *E. gracilis* plastid[17] and mitochondrion[18] have been characterised by mass spectrometry proteomics.

Results

Light and electron microscopy and phylogenetic position

The cells were initially observed by light and electron microscopy (Fig. 1). They possess two flagella inserted in the flagellar pocket, but only one extends beyond. The surface of the cell is formed into 7–9 ridges supported by the pellicle and microtubules. Conspicuous paramylon grains were observed in the cytoplasm, as were many mitochondrial cross-sections with discoidal cristae. The nucleus contains a large nucleolus and multiple heterochromatin regions. Based on the microscopic observations and the phylogenetic analysis of the gene for 18S rRNA (Fig. 2), the organism was positively identified as *Rhabdomonas costata* (Korshikov) Pringsheim 1942.

General characteristics of the genome and transcriptome

Basic characteristics of the *R. costata* genome assembly are given in Table 1. The assembly is very fragmented: 36,105 contigs above 1 kb with the N50 being 1,194 and the average coverage being ~114×. The total length of the assembly is 106.9 Mbp. Another indication the genome is quite incomplete is that only 43.5 % of transcriptome reads mapped to the assembly. As we had obtained a highly fragmented and incomplete draft genome, we did not proceed with gene prediction.

Transcriptome sequencing resulted in 39,585 non-redundant protein predictions. Completeness of the transcriptome measured by BUSCO (complete BUSCOs: 76.0%, fragmented BUSCOs: 12.5%, missing BUSCOs: 11.5%, n: 303) was satisfactory and comparable to published transcriptomes of euglenophytes (missing BUSCOs in *E. gracilis*, *E. longa*, and *E. gymnastica* were 8.3 %, 7.3 %, and 39.3 %, respectively), suggesting that this data set is sufficient for describing the selected features of *R. costata*. Basic characteristics of the transcriptome are given in Table 2. The proteins were automatically annotated by BLAST against NCBI nr, and only 26,052 proteins have at least one good homologue (e-value < 10^{-5}). The taxonomic affiliations of *R. costata* proteins are summarized in Fig. 3. As expected, most proteins of the 3,340 for which a relationship could be robustly established (bootstrap 75 or higher) affiliated with taxa belonging to Excavata (55 %). 13 % of the proteins branched with prokaryotes and remaining 32 % affiliated with any of the other small bins. The proteins that affiliated with prokaryotes may represent contamination, but some may be *R. costata* proteins having strong affiliation to prokaryotic homologues due to increased divergence, absence of eukaryotic homologues in the nr database, or origin via horizontal gene transfer. KEGG analysis of the 26,052 putative proteins led to 13,130 KEGG annotations (Fig. 4).

All 93,852 transcripts were searched for the presence of the published sequence of the *R. costata* spliced leader sequence (ACATTACTGGGTGTAATCATTTTTTCG)[19]. Only 15.3 % of transcripts contained the 10-nucleotides partial SL (CATTTTTTCG) or a longer fraction of the full-length SL at one of the ends (in case the transcript was in reverse complement orientation). The longest part of the SL we were able to find was only 17 nucleotides long (out of the 27 published). More details and comparison with other euglenid transcriptomes are given in Supplementary Table S1 online.

Intron characteristics

Although it was not suitable for gene prediction, we used the genomic assembly of *R. costata* for an analysis of the types of introns. Introns were detected by mapping the transcripts to genomic contigs. First, this mapping was done manually in the case of genes for which the presence of introns has been reported in other euglenids[20]. We have identified 29 complete introns in these six genes (Table 3). Some gene regions were not covered by transcripts, and so the presence of additional introns in these genes

cannot be excluded. All complete detected introns have conventional GT/AG boundaries. For 14 introns, only one end was found in the data, and these were marked as incomplete. The type of these introns could not be determined with certainty, but all of them have at least one end with conventional boundaries.

The positions of the introns in the two best-studied genes in this respect (*tubA* and *tubB*) were compared with their homologues in other euglenids (Fig. 5). All *R. costata* introns were in the same positions as those described by Milanowski et al[20] in tubulin genes of the primary heterotroph *Menoidium bibacillatum*. The second conventional intron in *tubA* of the two heterotrophs is in a position identical to the intron in phototrophic euglenids. The positions of all other introns in the tubulin genes of heterotrophs and phototrophs are different. Only one of 15 introns in *hsp90* is in the same position as in *Euglena agilis hsp90*; on the other hand, the gene for *R. costata* fibrillarin shares 3 of 4 intron sites with the gene of *E. gracilis* (not shown).

We also predicted the putative introns by mapping the transcriptome to the genome assembly (Supplementary Tables S2 and S3 online), revealing 105 contigs containing putative introns with nonconventional boundaries (not GT(GC)/AG) (Supplementary Table S4 online). Of these, 26 represented genes with homologues in NCBI detected by blastX (Supplementary Table S5 online). These were manually inspected. The manual inspection revealed seven cases of putative nonconventional introns with boundaries confirmed by a transcript and reads mapping (Supplementary Table S6 online), of which four are very likely nonconventional introns as no alternative transcripts were observed. One example intron in the hypothetical protein encoded in the genomic contig NODE_718 and the transcript TR27401 is shown in Fig. 6.

Mitochondrial proteome

We used the set of proteins predicted from the transcriptome to *in silico* determine the mitochondrial proteome of this primarily osmotrophic euglenid. *R. costata* strain PANT2 grows in a polyxenic culture at the bottom of stationary 15 ml tubes, indicating that it thrives in a low-oxygen environment, in spite

of containing mitochondria with well-developed cristae (Fig. 1). In our experience, it was able to survive only short-term (one month) cultivation in complete anaerobiosis.

From the 1,539 proteins of the predicted mitochondrial proteome, 1,018 were assigned to functional categories adopted from KEGG (Supplementary Table S7 online, Fig. 7). 1,275 proteins have orthologs in the experimentally established and manually curated proteome of the *E. gracilis* mitochondrion[18]. Contigs containing sequences homologous to all seven protein-coding genes reported from the *E. gracilis* mitochondrial genome[12] were detected in the genome assembly (Supplementary Table S8 online). Given that in all cases the contigs contained only a single gene or its part, it could not be established whether they originated from the mitochondrial or nuclear genome. In addition, most of the coding sequences were interrupted by multiple stop codons indicating that they are pseudogenes. Although we did not detect any contigs that could be confidently considered part of the mitochondrial genome, the presence of this genome is expected as the predicted mitoproteome set contains a repertoire of over 140 proteins involved in DNA and RNA metabolism, ribosome biogenesis, and translation (Supplementary Table S7 online). These include two of the already published DNA polymerases I[21].

Predicted functions and metabolic pathways of the mitochondrion are summarised in Fig. 8. Pyruvate and malate are probable substrates for energy metabolism. A malic enzyme (RCo000811, and RCo012435) catalyses the oxidative decarboxylation of malate to pyruvate. Pyruvate:NADH oxidoreductase (PNO) is the only enzyme in the transcriptome with activity for the oxidative decarboxylation of pyruvate to acetyl-CoA. It is present in five transcripts that are not identical in sequence, representing at least three different versions of the protein (RCo008351, RCo008357, RCo021209, RCo032638, and RCo052779). The canonical mitochondrial pyruvate dehydrogenase complex is apparently absent, as only the E3 subunit (dihydrolipoamide dehydrogenase; RCo010764) was recovered in the dataset. Transcript RCo000646 for subunit E1 is likely contamination, as it has 95 % identity with the bacterium *Magnetospirillum aberrantis* according to the blastp. The subunit E3 is also a component of other mitochondrial enzyme complexes, and in *R. costata* it is probably involved in the glycine cleavage system. Pyruvate can be alternatively reduced to L-/D-lactate by lactate dehydrogenases (RCo006066, RCo011296, RCo018299, RCo022174, RCo038617, RCo038618). The

gluconeogenetic enzymes pyruvate carboxylase (RCo04329) and phosphoenolpyruvate carboxykinase (RCo024588, RCo025226) are present. There are also three similar copies of 1,3-β-D-glucan synthase (RCo003309, RCo003310, RCo022891). This enzyme is orthologous to the *E. gracilis* glucan synthase-like 2 protein, which is responsible for the synthesis of paramylon[22] and has been reported in the mitochondrial proteome of *E. gracilis*[18]. Unlike *E. gracilis* proteins, based on the prediction software the probability that glucan synthase-like 2 protein in *R. costata* has mitochondrial localisation is low, and so we consider it putatively cytosolic. This is consistent with the cytosolic localization of paramylon grains in the cytoplasm of *R. costata* (Fig. 1).The enzyme endo-1,3(4)-β-glucanase, which is potentially involved in the degradation of paramylon, is also present (RCo017523, RCo019147, RCo047076, RCo017521, RCo018319, and RCo023166) and *in silico* predicted to localise in the cytosol.

The TCA cycle seems complete. The α-ketoglutarate dehydrogenase complex is absent, but this step is bypassed by enzymes α-ketoglutarate decarboxylase (RCo009834) and succinate-semialdehyde dehydrogenase (RCo025294). Both subunits of succinyl-CoA synthetase are present (RCo022592, and RCo042191) and this enzyme probably works in the direction of succinyl-CoA formation, thus leading to fatty acid synthesis. The GABA shunt is present, but the glyoxylate cycle enzymes are absent. Succinate dehydrogenase (SDH, Complex II) is similar in subunit composition to kinetoplastid SDH[23]. In total, six subunits were found, including the conserved eukaryotic subunits SDH1-2 (SDHA-B) and Euglenozoa-specific subunits 6-9[23]. The FeS cluster-containing subunit SDH2 splits into two polypeptides (N- and C- terminus), similar to that in trypanosomes and *E. gracilis*[24]. The Euglenozoa-specific subunits SDH5, 10 and 11 are missing in the data, as well as in the *E. gracilis* transcriptome[13]. Other components of the respiratory chain detected in the transcriptome include complex I (21 subunits including 3 Euglenozoa-specific), complex III (5 subunits), complex IV (8 subunits), F_oF_1 ATPase (subunits of F_1 part α, β, γ, δ and OSCP and three Euglenozoa-specific), and electron-transferring-flavoprotein dehydrogenase (RCo000411, and RCo000414). The alternative oxidase that was reported in other Euglenozoa, is absent from the transcriptome and was not detected by PCR with specific primers. Three types of membrane-associated electron carriers are present – cytochrome c, ubiquinone (UQ; most enzymes involved in its synthesis are present in the transcriptome), and rhodoquinone (RQ), which is

formed from ubiquinone by rhodoquinone methyltransferase (RCo043299, RCo043301, RCo043302, and RCo043305). Soluble electron-transferring flavoprotein (ETF; RCo028350, RCo035495, and RCo011455) may serve as an electron donor for fatty acid metabolism.

The ability to synthesise RQ provides *R. costata* with the opportunity to transfer electrons from Complex I via Complex II to fumarate, but the same reaction can be performed by FAD-dependent fumarate reductase (Rco013263, Rco047402, Rco047403, and Rco048726), which uses ubiquinol for fumarate reduction. The succinate produced is the substrate of succinyl-CoA synthetase producing succinyl-CoA, which may enter the synthesis of wax esters described in *E. gracilis*. *R. costata* contains all of the enzymes needed for this process. Propionyl-CoA, the first committed substrate for wax ester synthesis, is produced by methylmalonyl-CoA mutase (Rco028046), methylmalonyl-CoA epimerase (Rco048567), and propionyl-CoA carboxylase α and β (Rco005823, and Rco014953). The condensation of propionyl-CoA and acetyl-CoA can be, in principle, catalysed by acetyl-CoA acyltransferase (Rco038684, Rco011994, Rco024188, Rco031732, and Rco035370) instead of the missing α-ketoacyl synthase. *R. costata* contains 3-hydroxyacyl-CoA dehydrogenase (Rco47669) and enoyl-CoA hydratase (Rco002785, RCo016455, and RCo030891), as well as enzymes needed for the reduction of trans-enoyl-CoA: trans-2-enoyl-CoA reductase (RCo048015, and RCo048574), acyl-CoA dehydrogenase (RCo029909) and ETF, which can provide electrons. The pathway further proceeds outside the mitochondrion, where carnitine O-palmitoyltransferases 1 and 2 (RCo004453, RCo008998, RCo016397, and RCo033121) export the acyl-CoA. Neither fatty acyl-CoA reductase (an ER enzyme) nor wax ester synthase (a cytosolic enzyme) were detected; however, 17 transcripts encoding a bifunctional enzyme ester synthases/diacylglycerol acyltransferases (WSD) robustly branching with their orthologues in *E. gracilis* were detected (Supplementary Fig. S1 online). This enzyme was firstly characterised in *Acinetobacter calcoaceticus*[25] and later demonstrated as the dominant enzymes for the wax ester synthesis in *Euglena gracilis*[26]. Notably, two of the *E. gracilis* proteins closely related to *R. costata* homologues (BAV82975.1 and BAV82978.1) seem to play pivotal role in this process[26].

The organelle may be able to import sulphate via a putative transporter (RCo008807), although the identity of this protein is uncertain. It also contains enzymes needed for sulphate activation, sulphate

adenylyltransferase (RCo049652) and adenylylsulphate kinase (CysC; RCo011997), to produce phosphoadenosyl-5'-phosphosulphate (PAPS). The enzymes that metabolise inorganic sulphur compounds, thiosulphate sulphur transferase and sulphite oxidase, are present (RCo000455, and RCo010220); however, the enzymes necessary for the production of sulphide were not detected. Still, transcripts of the sulphide-dependent enzyme, cysteine synthase (RCo009036, RCo016104, RCo016108, RCo013733, and RCo051020), are present, as is the L-serine producing serine O-acetyltransferase (RCo000499, RCo028429, and RCo042343). The mitochondrion also contains a rich set of enzymes for the early and late synthesis of FeS clusters, including the mitochondrial export system.

The predicted mitochondrial proteome contains enzymes involved in the synthesis and metabolism of 10 proteinogenic amino acids (S, C, T, G, A, V, L, I, Q, and P). A complete glycine cleavage system and serine/glycine hydroxymethyl transferase, which are involved in the folate one-carbon pool, are also present. The set contains 136 entries involved in metabolite and ion transport across membranes, of which 31 are ABC transporters (including Atm1, involved in FeS cluster export) and 75 are mitochondrial carrier family proteins (solute carrier family 25) that cluster into 71 distinct clades (Supplementary Fig. S2 online).

More than 30 transcripts of proteins putatively involved in protein transport and maturation were detected. These encode four outer membrane proteins – the translocation pore Tom40 (RCo005874), its insertase Tob55 (RCo035243, and RCo050551), and Euglenozoa specific proteins Atom69 (RCo012450, and RCo012448) and Atom46 (RCo030973) – as well as two distinct homologues of small Tim (RCo000369, and RCo022175) localised to the intermembrane space, and seven proteins localised in the inner membrane – inner membrane protease subunit 2 (RCo035651), Tim22 (RCo035157), Tim17 (RCo003131 and RCo018944), Tim44 (RCo035673), Tim16 (RCo040212), Pam16 (RCo035852), and Oxa1 (RCo002990, RCo023198, and RCo030475). Homologues of both subunits of mitochondrial processing peptidase were detected (RCo000876, and RCo039561); however, the β–subunit may also be part of Complex III. Soluble chaperones, Mge1 (RCo018360) and Hsp60 (RCo005843), are present. 17 different transcripts for the chaperon Hsp70 were detected of which two (RCo049920 and

RCo045932) robustly branch within the mitochondrial clade (Supplementary Fig. S3 online) likely representing the mtHSP70 orthologues.

Tetrapyrrole synthesis pathways

We found transcripts for the full set of heme biosynthesis enzymes in the transcriptome (Fig. 9). All enzymes formed monophyletic clades with *E. gracilis* mitochondrial-cytosolic C4 pathway enzymes with various statistical support, and were most likely present in their common ancestor. The mitochondrial 5-aminolevulinate synthase (ALAS; RCo053079) branches within the eukaryotic clade and appears closely related to α-proteobacteria, suggesting a mitochondrial origin (Supplementary Fig. S4 online). Although ALAS should localize in the mitochondrion, *in silico* prediction for this localisation give probability below the 0.5 threshold. The following four steps take place in the cytosol and the pathway ends in the mitochondrion. Porphobilinogen synthase (ALAD; RCo046560), and porphobilinogen deaminase (PBGD; RCo016092) homologues are closely related to homologues from another bacterivorous euglenid (*Distigma* sp.), and together with the cytosolic isoforms of *E. gracilis*, they branch within eukaryotic genes (Supplementary Figs. S5 and S6 online). *R. costata* uroporphyrinogen synthase (UROS; RCo031923) branches with photosynthetic euglenids, and the clade is weakly supported sister group to oomycetes (Supplementary Fig. S7 online). Although this protein was annotated as a plastidial form in *E. gracilis*, it is probably cytosolic or dual-localized in both the cytosol and plastid. Cytosolic localization is supported by the presence of a second, so far unnoticed and putatively plastidial UROS homolog in the transcriptome of *E. gracilis* (EG_transcript_17485), which robustly branches with green algae and cyanobacteria, though it lacks a clear plastidial targeting signal. The next enzyme, uroporphyrinogen decarboxylase (UROD), has three annotated isoforms in *E. gracilis*, and *E. gymnastica*. One of them is plastidial and does not originate from green algae but more likely from cryptophytes. The other two are probably originaly cytosolic and branch within the eukaryotic clade. These isoforms seem to originate from an ancient gene duplication, and we found only one isoform (RCo052619) in *R. costata* (Supplementary Fig. S8 online).

The most complicated situation is in the case of coproporphyrinogen oxidase (CPOX). *E. gracilis* has at least eight isoforms. Five of them are plastidial, oxygen-dependent CPOX (HemF), and no homologue

was found in *R. costata*. Three other isoforms of *E. gracilis* belong to phylogenetically distant oxygen-independent CPOX (HemN) clade, where they occupy different positions (Supplementary Fig. S9 online). Two *E. gracilis* isoforms group together with *R. costata* (RCo019415, and RCo045123) homologues and branch within the α-proteobacterial clade, suggesting their origin from HGT. These two isoforms have a clear mitochondrial targeting signal. The third one is more divergent, and the *R. costata* homologue (RCo029168) is not closely related to *E. gracilis*, but rather to α-proteobacteria. *R. costata* has several other homologues without specific mitochondrial targeting signal branching on various places in the tree. A putatively mitochondrial isoform of protoporphyrinogen oxidase (PPOX; RCo026828) branches within a well-supported eukaryotic clade together with *Distigma* sp. and *E. gracilis* (Supplementary Fig. S10 online), but the targeting prediction is weak. Plastid isoforms of *E. gracilis* and *E. gymnastica* PPOX are placed within plastid genes with "chromalveolate" origin This contrasts with *E. gracilis*, and *E. gymnastica* PPOX plastid isoforms, which are placed within plastid genes with "chromalveolate" origin. Lastly, the mitochondrial isoform of ferrochelatase (FECH) is not derived from eukaryotes; instead, euglenid sequences, including the *R. costata* homologue (RCo012206), branch within bacterial clades and were probably obtained by HGT (Supplementary Fig. S11 online). Consistent with the probable absence of a plastid, there are no traces of the plastidial C5 pathway in *R. costata*.

Discussion

In this work, we contribute to the understanding of nonphotosynthetic euglenids by presenting draft genome and transcriptome assemblies of *R. costata*. Although the genome assembly is very fragmented and incomplete, we demonstrate its usefulness by deducing new information about intron composition in *R. costata*. We believe that the fragmentation and incompleteness is not caused by insufficient sequencing depth, but rather is an artefact of uneven whole genome amplification (WGA) in the DNA sample preparation. Other reasons for extreme fragmentation (e.g. presence of repeats) cannot be excluded at this point. It should be mentioned that the level of fragmentation is comparable to the published draft genome of *E. gracilis*[27] and the total number of contigs in the assembly is lower in *R. costata* (143,763 vs. 2,066,288). Similar to *E. gracilis*, the *R. costata* transcriptome is richer in GC

content than the genome: 58 % vs. 51 %, respectively. In total, the assembly is 106 Mbp in length, which, considering its completeness estimated from transcriptome mapping (43.5 %), gives a haploid genome size estimate of approx. 250 Mbp. This is half of the value estimated for *E. gracilis*[13].

Conversely, the quality of the transcriptome dataset was sufficient for functional annotation. The number of non-redundant predicted proteins (39,585) is comparable to *E. gracilis* (36,526)[13]. Mature euglenid transcripts often contain splice leader (SL) sequences acquired by trans-splicing[19]. However, it is likely that not all transcripts require SL for successful translation[28] and in the transcriptomes of *E. gracilis*, and *E. longa*, only 54 % and 48.5 % of transcripts, respectively, have been reported to contain at least a fragment of SL[29,30] somewhere in their sequence. This could be to some extent caused by truncation of the N-termini. We applied a stricter rule for SL identification, in which the SL was only searched for by an exact match at either end of the transcripts (in case of reverse complement orientation). This search obviously revealed a lower fraction of SL-containing transcripts (Supplementary Table S1 online), but was within the range of other euglenid transcriptomes, which vary from 5.7 % in *E. gymnastica* to 28.6 % in *E. longa* (*R. costata* transcriptome contains 15.3 % transcripts longer than 9 nucleotides).

Almost all eukaryotic genomes, including *E. gracilis*, contain introns that are removed by the ribonucleoprotein complex spliceosome. These conventional introns have the consensus sequence GT(GC)/AG at their ends, and they are excised by two sequential transesterifications. They have been described in genes encoding 13 proteins[20,28,31–35]. However, the genes of euglenids also contain also nonconventional introns variable in length and with no clear pattern of the nucleotide sequence at the exon/intron junction. They form the stem-loop RNA structures, and their excision is probably independent of the spliceosome, taking place after the excision of spliceosomal introns[20,36]. Besides these main types of introns, so-called intermediate introns that combine the features of both types have been reported[20].

The genome of *R. costata* seems to be relatively intron rich. We have confirmed the presence of introns in all genes for which introns have been investigated in other euglenids[20,28,31,33], and our automatic search for introns in the fragmented draft genome revealed hundreds of putative introns. Seven of these introns

are likely nonconventional. The presence of nonconventional introns in *R. costata* is expected, as they have been reported from several euglenid species and from a marine diplonemid[7].

The phylogenetic affiliation of predicted *R. costata* proteins is similar to that assessed for *E. gracilis*[13], with the notable but expected difference in the fraction of genes affiliated with Viridiplantae. The fraction is much higher in *E. gracilis* than in *R. costata* (14 % as compared to 5 %) and reflects the symbiotic history of *E. gracilis*, the plastid of which originated from a green algal endosymbiont. The *in silico* predicted proteome of the *R. costata* mitochondrion is smaller in size than the experimental proteome of *E. gracilis* – 1,554 in *R. costata* vs. approx. 2,500 in *E. gracilis*[18]. This difference may be real but also it may reflect the data set completeness and/or the procedure used to generate the set of proteins in the mitochondrial proteome. Direct orthology comparison showed that, of the 1,782 experimentally verified proteins of the *E. gracilis* mitochondrion[18], 1,083 have at least one ortholog in *R. costata*, and in total, 1,606 *R. costata* proteins are orthologous to this set. From this comparison and from BUSCO estimation (11.5 % missing), we infer that the completeness of the predicted *R. costata* protein set is reasonably high.

The mitochondrion of *E. gracilis* bears a unique combination of metabolic features. It contains a set of enzymes for the facultatively anaerobic metabolism that, in the presence of oxygen, metabolises pyruvate or malate by the pyruvate dehydrogenase complex, followed by a slightly modified TCA cycle, and then the full set of respiratory complexes, including the alternative oxidase also described from kinetoplastids[37]. In the absence of oxygen, pyruvate is oxidatively decarboxylated by pyruvate:NADH oxidoreductase, and the mitochondrial NADH is recycled by respiratory complex I and rhodoquinone-dependent fumarate reductase, producing succinate and propionyl-CoA. The latter is condensed with acetyl-CoA into fatty acids and wax esters that are stored in the cytoplasm at high concentrations[38]. These are recycled under aerobic conditions for ATP production or for the synthesis of organic compounds through a functional glyoxylate cycle, uniquely localised in the mitochondrion[39].

The biochemistry of the *R. costata* mitochondrion resembles, in several aspects, that of the *E. gracilis* organelle, but it seems to be more streamlined. It contains a complete TCA cycle with the euglenid-

specific bypass of α-ketoglutarate decarboxylase. It also uses rhodoquinone to reverse the electron flow in the truncated electron transport chain under low oxygen conditions to fumarate as the final electron acceptor, and the succinate produced is then consumed during the synthesis of wax monoesters as in *E. gracilis*. Unlike *E. gracilis*, the *R. costata* mitochondrion uses only PNO for oxidative decarboxylation of pyruvate, and it does not contain an alternative oxidase in its electron transport chain, which is consistent with our inability to amplify these transcripts by PCR. Intriguingly, the *R. costata* mitochondrion does not contain enzymes of the glyoxylate cycle, a shortcut of the TCA cycle, which is used to generate four-carbon molecules for the anabolic reactions from the acetyl-CoA released after the degradation of lipids and wax esters. How this is solved in *R. costata* remains to be elucidated.

An interesting difference between *E. gracilis* and *R. costata* is in sulphate metabolism. *E. gracilis* is apparently capable of assimilating sulphate into cysteine in the mitochondrion[18,40] and sulphite reductase has been detected in its chloroplast fraction[17]. In contrast, *R. costata* can only activate sulphate to the form of PAPS, an important coenzyme in sulphotransferase reactions, but the enzymes phosphoadenosine phosphosulfate reductase and sulphite reductase, necessary for the internal production of sulphide from PAPS, were not detected. Still, the presence of sulphide-dependent cysteine synthase suggests that *R. costata* may be able to synthesise cysteine from sulphide putatively produced from PAPS by unclear mechanism or taken up from the anaerobic environment.

Heme ranks among the essential cofactors in cellular metabolism because it is involved in many key biochemical processes. In most heterotrophic eukaryotes, the heme synthesis C4 pathway involves eight steps that are localised partially in the mitochondrion and partially in the cytosol. The first step is the condensation of glycine and succinyl-CoA into 5-amino-levulinate by ALAS in the mitochondrial matrix. The next four (or five) steps take place in the cytosol. The pathway ends in the mitochondrion with two reactions that take place in the intermembrane space, and a final reaction occurring in the mitochondrial matrix [41]. In most eukaryotes with a plastid, an alternative C5 pathway is present, and all steps are localised in the plastid. In this case, 5-amino-levulinate is synthesised from glutamic acid by three consecutive enzymes: glutamyl-tRNA synthetase (GltX), glutamyl-tRNA reductase (GTR), and glutamate-1-semialdehyde 2, 1-aminomutase (GSA-AT). The pathway then follows the same steps as

the classical pathway, but it is localised in the plastid and catalysed by enzymes of mostly cyanobacterial origin[42,43]. Unexpectedly, *E. gracilis* and the chlorarachniophyte *Bigelowiella natans* (algae with complex plastids that originated from green algae) have both pathways, the mosaic evolutionary origin of their enzymes reflecting the complex evolutionary history of these eukaryotes[42,44–46].

The *R. costata* transcriptome contains a complete set of enzymes of the mitochondrial/cytosolic C4 pathway that are orthologous to *E. gracilis* enzymes. Comparing these two euglenids helped reveal that one UROS and two UROD isoforms of *E. gracilis* examined so far are probably cytosolic C4 pathway enzymes, as they all have *R. costata* orthologs. Interestingly, we recovered an unnoticed UROS homolog from the *E. gracilis* transcriptome related to green algae and putatively involved in the C5 pathway in its plastid. While several plastidial isoforms of the oxygen-dependent CPOX (HemF) in *E. gracilis* have two distinct origins in the eukaryotic kingdom, the newly discovered isoforms of the oxygen independent CPOX (HemN) that function in the mitochondrion have α-proteobacterial origin. The complete absence of the plastidial C5 pathway for tetrapyrrole synthesis together with the overall limited number of transcripts affiliated to green algae is consistent with the absence of a chloroplast in *R. costata*, and supports the osmotrophic lifestyle of this euglenid as its primary state.

Methods

Strain origin and cultivation

Rhabdomonas costata strain PANT2 was isolated from a freshwater sediment sample collected ca 40 km south of Poconé municipality, Mato Grosso, Brazil (16 37′’ S, 56°44′’ W) and grown in monoeukaryotic culture together with a non-characterised mixture of bacteria in a standard 802 Sonneborn's *Paramecium* medium (ATCC medium #802) at room temperature and subcultured approximately once every 3 or 4 weeks. For the purpose of this project, we prepared by serial dilution a clonal lineage that was used for DNA and RNA extraction.

Microscopy

The DIC microscopy was performed on living cells, using an Olympus BX51 microscope with an Olympus DP70 camera. For the scanning electron microscopy, the pelleted cells were dropped on filter

paper and fixed with 2.5% glutaraldehyde in 0.1 M cacodylate buffer for 24 hours. Further processing was done by a service lab. The samples were observed using a JEOL JSM-6380LV microscope (JEOL, Akishima, Japan). For transmission electron microscopy, the pelleted cells were fixed in 2.5% glutaraldehyde in 0.1 M cacodylate buffer, postfixed with OsO_4 and the ultrathin sections were contrasted by uranyl acetate and observed using a JEOL JEM-1011 microscope.

Isolation of DNA

We used two methods to obtain high-quality, non-contaminated DNA of *R. costata*. (1) The culture cells were sorted by FACS with the value of the drop diameter equal to 70 μm (length of the cell is around 25 μm). Around 2,000 positive drops, most of them containing a single cell of *R. costata* were collected and used for DNA extraction. (2) By a combination of FACS and laser microdissection in which 69 drops from FACS were subsequently used for microdissection of individual cells, from which DNA was extracted. Both samples were then subjected to whole genome amplification (WGA; Sigma-Aldrich WGA4-10Rxn) to increase the amount of DNA. After amplification, the samples contained 27.7 μg and 32.0 μg of DNA, respectively. PCR with general prokaryotic primers for 16S rRNA had produced a specific product when sample 1 was used as a template, but no product with sample 2 (not shown). This supported our expectation that while sample 1 is still contaminated by bacterial DNA, sample 2 is likely contamination-free.

Genome sequencing and assembly

Both samples were sequenced on the Illumina platform. Initially, two MiSeq runs of sample 2 (l = 250 bp) produced 3.9 Gbp of sequences. Unfortunately, we were not able to produce a reasonable assembly from these data. Therefore, we also sequenced sample 1 using HiSeq (l = 100 bp) and obtained 8.2 Gbp of data. The raw reads from the sequencing of both samples were assembled by SPAdes v3.7.0 into 143,763 contigs (36,105 contigs longer than 1,000 bp).

RNA isolation, transcriptome sequencing and assembly

RNA was isolated three times using three slightly different approaches from three different specimens of our *R. costata* clonal lineage: 1) Direct isolation of mRNA (Dynabeads mRNA Direct Kit, including

polyA-selection; Thermo Fisher Scientific, Waltham, MA, USA) from which the library was prepared, including a polyA-selection step; 2) Isolation of total RNA (GeneAll Hybrid-R RNA purification kit; GeneAll Biotechnology, Seoul, South Korea) followed by mRNA purification (Dynabeads mRNA Direct Kit) from which the library was prepared, including a polyA-selection step; and 3) Isolation of total RNA (GeneAll Hybrid-R RNA purification kit) and from 4 μg of the total RNA, the library was prepared according to the standard TruSeq Stranded mRNA Sample Preparation Guide including a polyA-selection step. Unlike the first two approaches, the latter approach included only one step of polyA selection. The total RNA was always quantified using a Quantus Fluorometer (Promega, Madison, WI, USA) and its quality was checked with an Agilent Bioanalyzer (Agilent Technologies, Santa Clara, CA). All three samples were sequenced on an Illumina MiSeq instrument (Illumina, San Diego, CA, USA) using 150 base-length read chemistry in the paired-end mode. As no principal differences in bacterial contaminations among the three libraries were observed, they were assembled with Trinity v2.0.6 using default parameters and 93,852 contigs were created. We used Transdecoder for basal protein prediction and obtained 55,783 putative proteins. The software package cd-hit v.6[47] was applied with the default threshold 90 % to remove redundancy. The set was further partially decontaminated by removing proteins with the highest similarity to bacteria *Tolumonas auensis*, strain DSM 9187. The final data set contained 39,585 non-redundant amino acid sequences.

Prediction and characterisation of introns

Introns were automatically predicted by mapping the assembled transcripts to genome contigs using Exonerate (version est2genome 2.2.0). 1164 introns with predicted nonconventional boundaries were selected. Transcript reads were mapped to these contigs by STAR, resulting in 105 contigs in which nonconventional boundaries were supported. Out of these, 26 contigs have good and well-annotated hits on NCBI (e-value < 10^{-5}) and for these contigs the intron boundaries were manually inspected. The RNA secondary structures were drawn using Varna 3.9[48] and logo by WebLogo[49].

For selected genes (*tubA*, *tubB*, *tubG*, *hsp90*, *gapC* and *nop1p*), the introns were checked by aligning transcripts to genome contigs. The position of these introns was compared to other euglenids.

Phylogenetic ancestry of R. costata *proteins*

Ancestry of proteins was assessed with the same methodology as in Ebenezer et al.[13]. Briefly, homologues with e-value < 10^{-2} were retrieved from a custom database containing 207 taxa (additional file 3 in Ebenezer et al.[13]), aligned by MAFFT 7.273 with default parameters[50] and trimmed in trimAl 1.2 with default parameters[51]. 13,696 alignments with more than 3 taxa and longer than 74 amino acid residues were used for tree reconstruction in RAxML v8.1.17 with 100 rapid bootstraps [52] in Metacentrum (The National Grid Infrastructure in the Czech Republic). Custom scripts (Python 3.7) were used to sort the trees into bins based on the taxonomic affiliation of the clan in which *R. costata* branched. In 3,445 cases, the tree was included in a taxonomically uniform bin because it contained a bipartition supported by bootstrap 75 or higher comprised of *R. costata* and members of only one other defined taxonomic group.

Prediction of the mitochondrial proteome

The proteome of the mitochondrion was predicted using the following procedure. (1) The complete set of proteins predicted from the transcriptome was BLAST-searched against the Mitominer database, and 7,985 proteins with e-value < 10^{-4} were selected. To lower the redundancy, only the best *R. costata* hit for each protein included in the Mitominer database was kept, yielding a set of 1,501 proteins. For every protein, the probability of mitochondrial targeting was predicted using TargetP[53] and MitoFates[54] tools. Only proteins with probability of targeting equal to or higher than 0.5 in one or both tools were kept, producing 265 candidates. (2) This initial set was enriched by *R. costata* orthologues of proteins enriched in the mitochondrial fraction of *E gracilis*[18]. Orthology was established using the OrthoMCL software package[55] as described in Soukal et al. (in preparation). In total, 1,275 proteins fulfilled this criterion, 122 of which were included in the previous step. (3) An additional 121 proteins were included in the list as their presence in mitochondria is very likely – those that are typically part of the respiration chain and protein import complexes, as well as mitochondrial carrier family proteins (Solute carrier family 25). For every gene, the presence of a partial spliced leader (SL) at the 5' terminus and affiliation to prokaryotes was established in the custom database. All proteins mentioned above contained either a partial SL or their best hit was a eukaryote; 179 putative candidates that did not fulfill these criteria were

removed during the process of candidate selection. For each entry, the KEGG ID was assigned using the single-directional best hit method in KAAS[56] or transferred from *E. gracilis* homologue annotation, and the best BLAST hit was identified using blastp against the NCBI nr database. The final predicted mitoproteome consists of 1,539 proteins.

The conspicuous absence of an alternative oxidase in the dataset was verified by gene-specific PCR using degenerate primers based on sequence information from *E. gracilis* (forward primer: GARGARGCNGARAAYGARAGRATGCA; reverse primer GCRAANGTRTGRTTNACRTSNCGRTG), with *E. gracilis* and *Eutreptiella gymnastica* gDNAs as positive controls.

Phylogenies

The partial sequence for the 18S rRNA gene was amplified using primers EPA-23 (5'-GTCATATGCTTYKTTCAAGGRCTAAGCC-3'), and EPA-2286 (5'-TCACCTACARCWACCTTGTTACGAC -3') according to Müllner et al.[57] and sequenced using internal primers.

Phylogenetic trees of the gene for 18S rRNA and proteins of interest were generated by the following procedure. *R. costata* 18S rRNA/protein(s) and their homologs were downloaded from databases. In the case of the tetrapyrrole biosynthesis pathway, the dataset of Lakey and Triemer[46] was used as the seed and enriched by *E. gracilis* and *R. costata* transcripts as well as their best NCBI blast hits and the best hits from the local database. All entries were aligned by online version of MAFFT[50], using the automated strategy, trimmed in BMGE version 1.12[58], and manually inspected for misaligned positions. The phylogeny, including 1,000 ultrafast bootstraps, was inferred in IQ-Tree 2.0[59] using the BIC best-selected model (specified in the legends).

References

1. Leander, B. S., Lax, G., Karnkowska, A. & Simpson, A. G. B. Euglenida. In Handbook of the Protists 1–42 (Springer International Publishing, 2017).
2. Lax, G., Lee, W. J., Eglit, Y. & Simpson, A. ploeotids represent much of the phylogenetic diversity of euglenids. *Protist* **170**, 233–257 (2019).
3. Leander, B. S. Did trypanosomatid parasites have photosynthetic ancestors? *Trends Microbiol.* **12**, 251–258 (2004).
4. Hannaert, V. et al. Plant-like traits associated with metabolism of *Trypanosoma* parasites. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 1067–1071 (2003).
5. Marin, B. Origin and fate of chloroplasts in the Euglenoida. *Protist* **155**, 13–14 (2004).
6. Busse, I. & Preisfeld, A. Systematics of primary osmotrophic euglenids: a molecular approach to the phylogeny of *Distigma* and *Astasia* (Euglenozoa). *Int. J. Syst. Evol. Microbiol.* **53**, 617–624 (2003).
7. Gawryluk, R. M. R. et al. Morphological identification and single-cell genomics of marine diplonemids. *Curr. Biol.* **26**, 3053–3059 (2016).
8. Flegontova, O. et al. Extreme diversity of diplonemid eukaryotes in the ocean. *Curr. Biol.* **26**, 3060–3065 (2016).
9. Gibson, W. Kinetoplastea. (2016). In Handbook of the Protists 1–42 (Springer International Publishing, 2017).
10. Aurrecoechea, C. et al. EuPathDB: the eukaryotic pathogen genomics database resource. *Nucleic Acids Res.* **45**, 581–591 (2017).
11. Hallick, R. B. et al. Complete sequence of *Euglena gracilis* chloroplast DNA. *Nucleic Acids Res.* **21**, 3537–44 (1993).
12. Dobáková, E., Flegontov, P., Skalický, T. & Lukeš, J. Unexpectedly streamlined mitochondrial genome of the euglenozoan Euglena gracilis. *Genome Biol. Evol.* **7**, 3358–3367 (2015).
13. Ebenezer, T. E. et al. Transcriptome, proteome and draft genome of *Euglena gracilis*. *BMC Biol.* **17**, 1–23 (2019).
14. Gockel, G. & Hachtel, W. Complete gene map of the plastid genome of the nonphotosynthetic euglenoid flagellate *Astasia longa*. *Protist* **151**, 347–51 (2000).
15. Kasiborski, B. A., Bennett, M. S. & Linton, E. W. The chloroplast genome of Phacus orbicularis (Euglenophyceae): an initial datum for the Phacaceae. *J. Phycol.* **411**, 404–411 (2016).
16. Karnkowska, A., Bennett, M. S. & Triemer, R. E. Dynamic evolution of inverted repeats in Euglenophyta plastid genomes. *Sci. Rep.* **8**, 16071 (2018).
17. Novák Vanclová, A. M. G. et al. Metabolic quirks and the colourful history of the *Euglena gracilis* secondary plastid. *New Phytol.* **225**, 1578-1592 (2019).
18. Hammond, M. J. et al. A uniquely complex mitochondrial proteome from *Euglena gracilis*. *Mol. Biol. Evol.* **37**, 2173-2191 (2020).
19. Frantz, C., Ebel, C., Paulus, F. & Imbault, P. Characterization of trans-splicing in euglenoids. *Curr. Genet.* **37**, 349–355 (2000).
20. Milanowski, R., Karnkowska, A., Ishikawa, T. & Zakrys, B. Distribution of conventional and nonconventional nitrons in tubulin (a and b) genes of euglenids. *Mol. Biol. Evol.* **31**, 584–593 (2014).
21. Harada, R. et al. Inventory and evolution of mitochondrion-localized family a DNA polymerases in Euglenozoa*. Pathogens* **9**, 257 (2020).
22. Tanaka, Y. et al. Glucan synthase-like 2 is indispensable for paramylon synthesis in *Euglena gracilis*. *FEBS Letters* **591**, 1360–1370 (2017).
23. Morales, J. et al. Novel mitochondrial complex II isolated from *Trypanosoma cruzi* is composed of 12 peptides including a heterodimeric ip subunit. *J. Biol. Chem.* **284**, 7255–7263 (2009).

24. Gawryluk, R. & Gray, M. W. A split and rearranged nuclear gene encoding the iron-sulfur subunit of mitochondrial succinate dehydrogenase in Euglenozoa. *BMC Res. Notes* **7**, 1–7 (2009).
25. Kalscheuer, R. & Steinbüchel, A. A novel bifunctional wax ester synthase/acyl-CoA:Diacylglycerol acyltransferase mediates wax ester and triacylglycerol biosynthesis in *Acinetobacter calcoaceticus* ADP1. *J. Biol. Chem.* **278**, 8075–8082 (2003).
26. Tomiyama, T. et al. Wax ester synthase/diacylglycerol acyltransferase isoenzymes play a pivotal role in wax ester biosynthesis in *Euglena gracilis*. *Sci. Rep.* **7**, 13504 (2017).
27. Ebenezer, T. E. et al. Transcriptome, proteome and draft genome of *Euglena gracilis*. BMC Biol. **17**, 11 (2019).
28. Russell, A. G., Watanabe, Y. I., Charette, J. M. & Gray, M. W. Unusual features of fibrillarin cDNA and gene structure in *Euglena gracilis*: Evolutionary conservation of core proteins and structural predictions for methylation-guide box C/D snoRNPs throughout the domain Eucarya. *Nucleic Acids Res.* **33**, 2781–2791 (2005).
29. Yoshida, Y. et al. De novo assembly and comparative transcriptome analysis of *Euglena gracilis* in response to anaerobic conditions. *BMC Genomics* **17**, 1–10 (2016).
30. Záhonová, K. et al. Peculiar features of the plastids of the colourless alga *Euglena longa* and photosynthetic euglenophytes unveiled by transcriptome analyses. *Sci. Rep.* **8**, 1–15 (2018).
31. Breckenridge, D. G., Watanabe, Y., Greenwood, S. J., Gray, M. W. & Schnare, M. N. U1 small nuclear RNA and spliceosomal introns in *Euglena gracilis*. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 852–6 (1999).
32. Ebel, C., Frantz, C., Paulus, F. & Imbault, P. Trans-splicing and cis-splicing in the colourless euglenoid, *Entosiphon sulcatum*. *Curr. Genet.* **35**, 542–50 (1999).
33. Canaday, J., Tessier, L. H., Imbault, P. & Paulus, F. Analysis of *Euglena gracilis* alpha-, beta- and gamma-tubulin genes: introns and pre-mRNA maturation. *Mol. Genet. Genomics* **265**, 153–60 (2001).
34. Vesteg, M. et al. A possible role for short introns in the acquisition of stroma-targeting peptides in the flagellate *Euglena gracilis*. *DNA Res.* **17**, 223–231 (2010).
35. Milanowski, R., Gumińska, N., Karnkowska, A., Ishikawa, T. & Zakryś, B. Intermediate introns in nuclear genes of euglenids - are they a distinct type? *BMC Evol. Biol.* **16**, 49 (2016).
36. Gumińska, N., Płecha, M., Zakryś, B. & Milanowski, R. Order of removal of conventional and nonconventional introns from nuclear transcripts of *Euglena gracilis*. *PLoS Genet.* **14**, e1007761 (2018).
37. Hellemond, J. J. Van, Bakker, B. M. & Tielens, A. G. M. Energy metabolism and its compartmentation in *Trypanosoma brucei*. *Adv. Microb. Physiol.* **50**, 199–226 (2005).
38. Tucci, S., Vacula, R., Krajcovic, J., Proksch, P. & Martin, W. Variability of wax ester fermentation in natural and bleached *Euglena gracilis* strains in response to oxygen and the elongase inhibitor flufenacet. *J. Eukaryot. Microbiol.* **57**, 63–69 (2010).
39. Müller, M. et al. Biochemistry and evolution of anaerobic energy metabolism in eukaryotes. *Microbiol. Mol. Biol. Rev.* **76**, 444–95 (2012).
40. Saidha, T., Stern, A. I., Lee, D. H. & Schiff, J. A. Localization of a sulphate-activating system within Euglena mitochondria. *Biochem. J.* **232**, 357–365 (1985).
41. Hamza, I. & Dailey, H. A. One ring to rule them all: Trafficking of heme and heme synthesis intermediates in the metazoans. *Biochimica et Biophysica Acta* **1823**, 1617–1632 (2012).
42. Kořený, L. & Oborník, M. Sequence evidence for the presence of two tetrapyrrole pathways in *Euglena gracilis*. *Genome Biol. Evol.* **3**, 359–364 (2011).
43. Cenci, U. et al. Heme pathway evolution in kinetoplastid protists. *BMC Evol. Biol.* **16**, 109 (2016).
44. Woo, Y. H. et al. Chromerid genomes reveal the evolutionary path from photosynthetic algae to obligate intracellular parasites. *Elife* **4**, 1–41 (2015).

45. Cihlář, J., Füssy, Z., Horák, A. & Oborník, M. Evolution of the tetrapyrrole biosynthetic pathway in secondary algae: Conservation, redundancy and replacement. *PLoS One* **11**, e0166338 (2016).
46. Lakey, B. & Triemer, R. The tetrapyrrole synthesis pathway as a model of horizontal gene transfer in euglenoids. *J. Phycol.* **53**,198-217 (2017).
47. Li, W. & Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
48. Crooks, G. E. et al. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinfomatics* **25**, 1974–1975 (2009).
49. Crooks, G. E., Hon, G., Chandonia, J. & Brenner, S. E. WebLogo: A sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).
50. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–80 (2013).
51. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. TrimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
52. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–3 (2014).
53. Emanuelsson, O., Nielsen, H., Brunak, S. & Von Heijne, G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* **300**, 1005–1016 (2000).
54. Fukasawa, Y. et al. MitoFates: Improved prediction of mitochondrial targeting sequences and their cleavage sites. *Mol. Cell. Proteomics* **14**, 1113–1126 (2015).
55. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
56. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C. & Kanehisa, M. KAAS: An automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* **35**, W182-5 (2007).
57. Müllner, A. N., Angeler, D. G., Samuel, R., Linton, E. W. & Triemer, R. E. Phylogenetic analysis of phagotrophic, phototrophic and osmotrophic euglenoids by using the nuclear 18S rDNA sequence. *Int. J. Syst. Evol. Microbiol.* **51**, 783–791 (2001).
58. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* **10**, 210 (2010).
59. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

Author contributions

VH and PS conceived the study, IČ provided the culture, JS prepared clonal lineage and sequenced the 18S rRNA gene, PS prepared the amplified gDNA, MH prepared the gDNA libraries and sequenced them, HS and ČV assembled the genome, ŠH prepared the cDNA libraries, PS assembled the transcriptome, annotated the splice leaders and analysed the phylogenetic origin of genes, ŠH and VH annotated mitochondrial metabolism and heme synthesis, VH computed the trees, AK and RM searched for and annotated the introns, VH and ŠH wrote the manuscript, PS, IČ, AK and RM edited the manuscript All authors approved the final version.

Acknowledgement

All sequencing and the salaries of VH and PS were financially supported by the Czech Science Foundation project nr. 13-24983S and by a project of the Ministry of Education, Youth and Sports of CR within the National Sustainability Program II (Project BIOCEV-FAR) LQ1604 as well as by the project "BIOCEV" (CZ.1.05/1.1.00/02.0109). AK was supported by an EMBO Installation Grant. RM work was supported by grant 2015/19/B/NZ8/ 00166 from the National Science Centre, Poland, and IČ by the Czech Science Foundation grant no. 19-19297S. Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum provided under the programme "Projects of Large Research, Development, and Innovations Infrastructures" (CESNET LM2015042), is greatly appreciated. The authors would like to thank Martina Johnson Pokorná for help with FACS sorting and laser microdissection, František Šťáhlavský for providing the sample of sediment from which the culture of *R. costata* was derived and Ivan Hrdý and Zoltán Fűssy for reading and commenting the manuscript.

Additional information

The authors declare no competing interests.

Data availability statement

The transcriptomic and genomic reads are available in GenBank under the BioProject ID PRJNA550357, the assemblies and predicted proteins are available in the Zenodo repository (https://zenodo.org/record/4249289#.X6UlDWhKhPY). The sequence of the gene for 18S rRNA is deposited under GenBank accession nr. MW113742.

Figure legends

Fig. 1: Microscopic investigation of *Rhabdomonas costata.* Cells in DIC contrast (A) with visible pellicular stripes and paramylon grains. SEM microscopy (B) of a cell showing surface invaginations and a flagellum inserted in the flagellar pocket. Longitudinal (C) and transverse (D) TEM sections and details of the pellicle and mitochondria (E), nucleus (F), and flagellar pocket with two flagella (G). Ax – axoneme, Gb – Golgi body, Fl – flagella, Fp – flagellar pocket, Mt – mitochondrial cross-sections, Mtb – subpellicular microtubules, Nu – nucleus, Ncl – nucleolus, Par – holes after paramylon grain, Pe – pellicle.

Fig. 2: Phylogenetic tree of euglenids based on the 18S rRNA gene. The tree was constructed in IQ-Tree using the TIM2e+G4 model selected by Bayesian information criterion from a trimmed alignment containing 1,569 nucleotide positions. The values at the nodes represent ultrafast bootstraps from 1,000 repetitions, where above 50. The strain analysed in this study is shown in blue. The tree was rooted by the genus *Distigma*.

Fig. 3: Graph summarising taxonomical affinities of the predicted proteins of *Rhabdomonas costata*. 3,445 protein phylogenies, in which *R. costata* was robustly (BS>=75) placed into a taxonomically homogeneous clan, were sorted accordingly into taxonomic bins. Discoba and Kinetoplastea represent the subgroups of Excavata.

Fig. 4: KEGG Functional categories of predicted proteins of *R. costata*. 13,130 proteins (33 % of 39,585) were ascribed to functional categories according to KEGG. All categories with less than 10 members were merged into the category “other”.

Fig. 5: Comparison of the positions of introns in tubulin α and β genes in four euglenids. Heterotrophic euglenids *Rhabdomonas costata* and *Menoidium bibacillatum* (orange lines) and phototrophic euglenids *Euglena gracilis* and *Euglena agilis* (green lines). Introns shared between heterotrophs – orange ellipses, phototrophs – green ellipses, all four euglenids – brown ellipse.

Fig. 6: Example of the nonconventional intron of *R. costata*. The secondary structure of a nonconventional intron from NODE_718 (A), and the sequence logo of the boundaries of seven putative nonconventional introns detected in *R. costata* (B).

Fig. 7: KEGG functional categories of proteins predicted to the mitochondrial proteome of *R. costata*. 1,551 proteins of the mitochondrial proteome were ascribed to functional categories according to KEGG with some modifications. All categories with less than 10 members were merged into the category "other".

Fig. 8: Metabolic map of the *R. costata* mitochondrion. The map is based on *in silico* prediction of the mitochondrial proteome from the transcriptomic dataset. Blue circles represent enzymes present in *R. costata* with homologues in *E. gracilis*, light green circles represent enzymes absent in *R. costata* but present in *E. gracilis*, and brown circles represent typical eukaryotic enzymes missing in both euglenids. These colours allow for comparison, inspired by Ebenezer et al. 2019.

Fig. 9: Origins and subcellular localisation of tetrapyrrole synthesis enzymes in *R. costata*. Predicted localisation of the enzyme is indicated by its position in the diagram. Phylogenetic origin is indicated by colour (orange – eukaryotic origin, blue – eubacterial origin). Presence of mitochondrial targeting peptide (TP) is indicated by frame (red frame – high TP value in *R. costata*, green frame – high TP value in euglenophytes).

Tables

Table 1: Parameters of the genomic assembly.

Number of reads Miseq	11,624,864
Number of reads Hiseq	82,258,718
Number of contigs	143,763
Number of contigs (≥ 500 bp)	82,982
Number of contigs (≥ 1000 bp)	36,105
Number of contigs (≥ 5000 bp)	153
Median contig length (bp)	661
Median contig length (≥ 500 bp)	924
Max contig length (bp)	8,093
Total length (bp)	106,888,161
Total length (≥ 500 bp)	94,209,074
Total length (≥ 1000 bp)	58,674,546
N50 (bp)	1,194
L50	25,493
Fraction of GC (%)	51.63
Number of N's per 100 kbp	31.22

741 **Table 2:** Parameters of the transcriptomic assembly.

742

Number of reads	12,103,119
Number of contigs	93,852
Number of contigs (≥ 1000 bp)	19,335
Median contig length (bp)	430
Max contig length (bp)	17,913
Total length (bp)	66,880,466
Fraction of GC (%)	58.25
Putative proteins	55,783
Unique proteins	39,585
Any homologue (e-value ≤ 10^{-5})	26,052

743

744 **Table 3:** Introns identified in selected genes of *Rhabdomonas costata.* * Only one intron boundary
745 was found in the data, thus the intron type could not be determined with certainty. ** Part of ORF
746 length mapped to gDNA. The number of introns may not be definitive in low-percentage coverage.
747
748
749

Coded protein	Gene abb.	Complete conventional introns	Incomplete introns *	ORF length (nt)	gDNA coverage (%)**
α-tubulin	*tubA*	2	0	1356	100
β-tubulin	*tubB*	4	0	1335	100
γ-tubulin	*tubG*	6	7	1680	76
Heat shock protein 90	*hsp90*	10	5	2112	68
GAPDH	*gapC*	3	1	1062	70
Fibrillarin	*nop1p*	4	1	903	85

750
751
752