

Heterotrophic Euglenid Rhabdomonas Costata Resembles Its Phototrophic Relatives in Many Aspects of Molecular and Cell Biology

Petr Soukal

Charles University

Štěpánka Hrdá

Charles University

Anna Karnkowska

University of Warsaw

Rafał Milanowski

University of Warsaw

Jana Szabová

Charles University

Miluše Hradilová

Institute of Molecular Genetics

Hynek Strnad

Institute of Molecular Genetics

Čestmír Vlček

Institute of Molecular Genetics

Ivan Čepička

Charles University

Vladimír Hampl (✉ vlada@natur.cuni.cz)

Charles University

Research Article

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

Posted Date: November 30th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-112847/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

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

4

5

6 Petr Soukal^{1#}, Štěpánka Hrdá^{1#}, Anna Karnkowska², Rafał Milanowski², Jana Szabová¹, Miluše
7 Hradilová³, Hynek Strnad³, Čestmír Vlček³, Ivan Čepička⁴ and Vladimír Hampl^{1*}

8

9

10

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

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

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

15 ⁴Department of Zoology, Charles University, Prague, Czech Republic.

16

17

18 [#]Authors contributed equally to the manuscript

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

21

22 Abstract

23

24 Euglenids represent a group of protists with diverse modes of feeding. To date, only a partial genomic
25 sequence of *Euglena gracilis* and transcriptomes of several phototrophic and secondarily osmotrophic
26 species are available, while primarily heterotrophic euglenids are seriously undersampled. In this work,
27 we begin to fill this gap by presenting genomic and transcriptomic drafts of a primary osmotroph,
28 *Rhabdomonas costata*. The current genomic assembly length of 100 Mbp is 14× smaller than that of
29 *E. gracilis*. Despite being too fragmented for comprehensive gene prediction, comparison of the
30 transcriptomic and genomic data revealed features of its introns, including several candidates for
31 nonconventional introns. 16 % of transcripts bear a recognizable partial splice leader sequence. A set of
32 39,585 putative *R. costata* proteins were predicted from the transcriptome. Annotation of the
33 mitochondrial core metabolism provides the first data on the facultatively anaerobic mitochondrion of
34 *R. costata*, which in most respects resembles the mitochondrion of *E. gracilis* with certain level of
35 streamlining. *R. costata* synthesises heme by a mitochondrial-cytoplasmatic C4 pathway with enzymes
36 orthologous to those found in *E. gracilis*. The low percentage of green algae-affiliated genes, supports
37 the ancestrally osmotrophic status of this species.

38

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

41

42 Introduction

43 Euglenids are a species-rich group (>1,500 described species) of unicellular eukaryotes¹ classified into
44 the phylum Euglenozoa and defined by both ultrastructural and molecular features. Generally, they
45 possess one or two emergent flagella inserted in a paraflagellar pocket and reinforced by a paraflagellar
46 rod. The surface of their cells is formed by a distinctive pellicle consisting of three layers – the
47 cytoplasmic membrane, a proteinaceous belt supported by microtubules, and the vesicles of the
48 endoplasmic reticulum. The pellicle enables some euglenids to move in a characteristic manner
49 by undulated shifts in the shape of the cell, which is referred to as metaboly or euglenoid movement.
50 Storage of carbohydrates as paramylon, a β -1,3-glucan, is unique to euglenids¹.

51 All major types of eukaryotic nutrition are present in Euglenida – phagotrophy (eukaryovory and
52 bacteriovory), osmotrophy, and phototrophy. Phagotrophic euglenids form several clades in the
53 phylogenetic tree of this group, including several deepest branches². Phototrophic euglenids
54 (*Euglenophyceae*) which arose from one of these clades are, by all means, the best-studied subgroup.
55 Their cell contains a triple membrane-bound plastid derived from a secondary endosymbiosis with a
56 green alga. Since *Euglenophyceae* certainly constitute a clade, it is assumed that this endosymbiosis
57 occurred in the clade's exclusive common ancestor³; however, early acquisition of a plastid in the
58 Euglenozoa lineage has also been proposed to account for plant-like traits in trypanosomatids⁴. Although
59 all known members of *Euglenophyceae* contain plastids, six species have lost the ability to
60 photosynthesize and have become secondarily osmotrophic, of which the best-known is *Euglena longa*
61 (formerly known as *Astasia longa*)⁵. Primarily osmotrophic euglenids form a distinct clade Aphagea⁶
62 branching within phagotrophic euglenids². Members of the order *Rhabdomonadales* are distinguished
63 from other members of the osmotrophic clade (i.e., *Distigma*, *Astasia*) by their lack of euglenoid
64 movement.

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

68 *Trypanosoma* and *Leishmania* are some of the most studied protists⁹. Currently, mitochondrial and
69 nuclear genomes of 50 species and strains of kinetoplastids have been sequenced¹⁰, but besides the
70 kinetoplastids there are no complete and well-annotated nuclear genome sequences available for
71 Euglenozoa. Partial genome sequences of marine diplomonads have been obtained by single-cell
72 approaches⁷. In Euglenida, genomic studies have as yet covered relatively well the iconic model
73 flagellate *Euglena gracilis*, these yielding the chloroplast and mitochondrial genomes^{11,12} and a very
74 fragmented and unannotated nuclear genome¹³. Genomic data set from the remaining euglenids
75 comprise exclusively the chloroplast genomes, 30 of which have been published until present^{11,14–16}.
76 Recently, the proteomes of *E. gracilis* plastid¹⁷ and mitochondrion¹⁸ have been characterised by mass
77 spectrometry proteomics.

78 Results

79 *Light and electron microscopy and phylogenetic position*

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

87 *General characteristics of the genome and transcriptome*

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

93 Transcriptome sequencing resulted in 39,585 non-redundant protein predictions. Completeness of the
94 transcriptome measured by BUSCO (complete BUSCOs: 76.0%, fragmented BUSCOs: 12.5%, missing
95 BUSCOs: 11.5%, n: 303) was satisfactory and comparable to published transcriptomes of euglenophytes
96 (missing BUSCOs in *E. gracilis*, *E. longa*, and *E. gymnastica* were 8.3 %, 7.3 %, and 39.3 %,
97 respectively), suggesting that this data set is sufficient for describing the selected features of *R. costata*.
98 Basic characteristics of the transcriptome are given in Table 2. The proteins were automatically
99 annotated by BLAST against NCBI nr, and only 26,052 proteins have at least one good homologue (e-
100 value < 10⁻⁵). The taxonomic affiliations of *R. costata* proteins are summarized in Fig. 3. As expected,
101 most proteins of the 3,340 for which a relationship could be robustly established (bootstrap 75 or higher)
102 affiliated with taxa belonging to Excavata (55 %). 13 % of the proteins branched with prokaryotes and
103 remaining 32 % affiliated with any of the other small bins. The proteins that affiliated with prokaryotes
104 may represent contamination, but some may be *R. costata* proteins having strong affiliation to
105 prokaryotic homologues due to increased divergence, absence of eukaryotic homologues in the nr
106 database, or origin via horizontal gene transfer. KEGG analysis of the 26,052 putative proteins led to
107 13,130 KEGG annotations (Fig. 4).

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

114 *Intron characteristics*

115 Although it was not suitable for gene prediction, we used the genomic assembly of *R. costata* for an
116 analysis of the types of introns. Introns were detected by mapping the transcripts to genomic contigs.
117 First, this mapping was done manually in the case of genes for which the presence of introns has been
118 reported in other euglenids²⁰. We have identified 29 complete introns in these six genes (Table 3). Some
119 gene regions were not covered by transcripts, and so the presence of additional introns in these genes

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

124 The positions of the introns in the two best-studied genes in this respect (*tubA* and *tubB*) were compared
125 with their homologues in other euglenids (Fig. 5). All *R. costata* introns were in the same positions as
126 those described by Milanowski et al²⁰ in tubulin genes of the primary heterotroph *Menoidium*
127 *bibacillatum*. The second conventional intron in *tubA* of the two heterotrophs is in a position identical
128 to the intron in phototrophic euglenids. The positions of all other introns in the tubulin genes of
129 heterotrophs and phototrophs are different. Only one of 15 introns in *hsp90* is in the same position as in
130 *Euglena agilis hsp90*; on the other hand, the gene for *R. costata* fibrillarin shares 3 of 4 intron sites with
131 the gene of *E. gracilis* (not shown).

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

141 *Mitochondrial proteome*

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

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

147 From the 1,539 proteins of the predicted mitochondrial proteome, 1,018 were assigned to functional
148 categories adopted from KEGG (Supplementary Table S7 online, Fig. 7). 1,275 proteins have orthologs
149 in the experimentally established and manually curated proteome of the *E. gracilis* mitochondrion¹⁸.
150 Contigs containing sequences homologous to all seven protein-coding genes reported from the
151 *E. gracilis* mitochondrial genome¹² were detected in the genome assembly (Supplementary Table S8
152 online). Given that in all cases the contigs contained only a single gene or its part, it could not be
153 established whether they originated from the mitochondrial or nuclear genome. In addition, most of the
154 coding sequences were interrupted by multiple stop codons indicating that they are pseudogenes.
155 Although we did not detect any contigs that could be confidently considered part of the mitochondrial
156 genome, the presence of this genome is expected as the predicted mitoproteome set contains a repertoire
157 of over 140 proteins involved in DNA and RNA metabolism, ribosome biogenesis, and translation
158 (Supplementary Table S7 online). These include two of the already published DNA polymerases I²¹.

159 Predicted functions and metabolic pathways of the mitochondrion are summarised in Fig. 8. Pyruvate
160 and malate are probable substrates for energy metabolism. A malic enzyme (RCo000811, and
161 RCo012435) catalyses the oxidative decarboxylation of malate to pyruvate. Pyruvate:NADH
162 oxidoreductase (PNO) is the only enzyme in the transcriptome with activity for the oxidative
163 decarboxylation of pyruvate to acetyl-CoA. It is present in five transcripts that are not identical
164 in sequence, representing at least three different versions of the protein (RCo008351, RCo008357,
165 RCo021209, RCo032638, and RCo052779). The canonical mitochondrial pyruvate dehydrogenase
166 complex is apparently absent, as only the E3 subunit (dihydrolipoamide dehydrogenase; RCo010764)
167 was recovered in the dataset. Transcript RCo000646 for subunit E1 is likely contamination, as it has
168 95 % identity with the bacterium *Magnetospirillum aberrantis* according to the blastp. The subunit E3
169 is also a component of other mitochondrial enzyme complexes, and in *R. costata* it is probably involved
170 in the glycine cleavage system. Pyruvate can be alternatively reduced to L-/D-lactate by lactate
171 dehydrogenases (RCo006066, RCo011296, RCo018299, RCo022174, RCo038617, RCo038618). The

172 gluconeogenic enzymes pyruvate carboxylase (RCo04329) and phosphoenolpyruvate carboxykinase
173 (RCo024588, RCo025226) are present. There are also three similar copies of 1,3- β -D-glucan synthase
174 (RCo003309, RCo003310, RCo022891). This enzyme is orthologous to the *E. gracilis* glucan synthase-
175 like 2 protein, which is responsible for the synthesis of paramylon²² and has been reported in the
176 mitochondrial proteome of *E. gracilis*¹⁸. Unlike *E. gracilis* proteins, based on the prediction software
177 the probability that glucan synthase-like 2 protein in *R. costata* has mitochondrial localisation is low,
178 and so we consider it putatively cytosolic. This is consistent with the cytosolic localization of paramylon
179 grains in the cytoplasm of *R. costata* (Fig. 1). The enzyme endo-1,3(4)- β -glucanase, which is potentially
180 involved in the degradation of paramylon, is also present (RCo017523, RCo019147, RCo047076,
181 RCo017521, RCo018319, and RCo023166) and *in silico* predicted to localise in the cytosol.

182 The TCA cycle seems complete. The α -ketoglutarate dehydrogenase complex is absent, but this step is
183 bypassed by enzymes α -ketoglutarate decarboxylase (RCo009834) and succinate-semialdehyde
184 dehydrogenase (RCo025294). Both subunits of succinyl-CoA synthetase are present (RCo022592, and
185 RCo042191) and this enzyme probably works in the direction of succinyl-CoA formation, thus leading
186 to fatty acid synthesis. The GABA shunt is present, but the glyoxylate cycle enzymes are absent.
187 Succinate dehydrogenase (SDH, Complex II) is similar in subunit composition to kinetoplastid SDH²³.
188 In total, six subunits were found, including the conserved eukaryotic subunits SDH1-2 (SDHA-B) and
189 Euglenozoa-specific subunits 6-9²³. The FeS cluster-containing subunit SDH2 splits into two
190 polypeptides (N- and C- terminus), similar to that in trypanosomes and *E. gracilis*²⁴. The Euglenozoa-
191 specific subunits SDH5, 10 and 11 are missing in the data, as well as in the *E. gracilis* transcriptome¹³.
192 Other components of the respiratory chain detected in the transcriptome include complex I (21 subunits
193 including 3 Euglenozoa-specific), complex III (5 subunits), complex IV (8 subunits), F₀F₁ ATPase
194 (subunits of F₁ part α , β , γ , δ and OSCP and three Euglenozoa-specific), and electron-transferring-
195 flavoprotein dehydrogenase (RCo000411, and RCo000414). The alternative oxidase that was reported
196 in other Euglenozoa, is absent from the transcriptome and was not detected by PCR with specific primers.
197 Three types of membrane-associated electron carriers are present – cytochrome c, ubiquinone (UQ; most
198 enzymes involved in its synthesis are present in the transcriptome), and rhodoquinone (RQ), which is

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

202 The ability to synthesise RQ provides *R. costata* with the opportunity to transfer electrons from Complex
203 I via Complex II to fumarate, but the same reaction can be performed by FAD-dependent fumarate
204 reductase (Rco013263, Rco047402, Rco047403, and Rco048726), which uses ubiquinol for fumarate
205 reduction. The succinate produced is the substrate of succinyl-CoA synthetase producing succinyl-CoA,
206 which may enter the synthesis of wax esters described in *E. gracilis*. *R. costata* contains all of the
207 enzymes needed for this process. Propionyl-CoA, the first committed substrate for wax ester synthesis,
208 is produced by methylmalonyl-CoA mutase (Rco028046), methylmalonyl-CoA epimerase (Rco048567),
209 and propionyl-CoA carboxylase α and β (Rco005823, and Rco014953). The condensation of propionyl-
210 CoA and acetyl-CoA can be, in principle, catalysed by acetyl-CoA acyltransferase (Rco038684,
211 Rco011994, Rco024188, Rco031732, and Rco035370) instead of the missing α -ketoacyl synthase.
212 *R. costata* contains 3-hydroxyacyl-CoA dehydrogenase (Rco47669) and enoyl-CoA hydratase
213 (Rco002785, RCo016455, and RCo030891), as well as enzymes needed for the reduction of trans-enoyl-
214 CoA: trans-2-enoyl-CoA reductase (RCo048015, and RCo048574), acyl-CoA dehydrogenase
215 (RCo029909) and ETF, which can provide electrons. The pathway further proceeds outside the
216 mitochondrion, where carnitine O-palmitoyltransferases 1 and 2 (RCo004453, RCo008998, RCo016397,
217 and RCo033121) export the acyl-CoA. Neither fatty acyl-CoA reductase (an ER enzyme) nor wax ester
218 synthase (a cytosolic enzyme) were detected; however, 17 transcripts encoding a bifunctional enzyme
219 ester synthases/diacylglycerol acyltransferases (WSD) robustly branching with their orthologues in *E.*
220 *gracilis* were detected (Supplementary Fig. S1 online). This enzyme was firstly characterised in
221 *Acinetobacter calcoaceticus*²⁵ and later demonstrated as the dominant enzymes for the wax ester
222 synthesis in *Euglena gracilis*²⁶. Notably, two of the *E. gracilis* proteins closely related to *R. costata*
223 homologues (BAV82975.1 and BAV82978.1) seem to play pivotal role in this process²⁶.

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

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

234 The predicted mitochondrial proteome contains enzymes involved in the synthesis and metabolism of
235 10 proteinogenic amino acids (S, C, T, G, A, V, L, I, Q, and P). A complete glycine cleavage system
236 and serine/glycine hydroxymethyl transferase, which are involved in the folate one-carbon pool, are also
237 present. The set contains 136 entries involved in metabolite and ion transport across membranes, of
238 which 31 are ABC transporters (including Atm1, involved in FeS cluster export) and 75 are
239 mitochondrial carrier family proteins (solute carrier family 25) that cluster into 71 distinct clades
240 (Supplementary Fig. S2 online).

241 More than 30 transcripts of proteins putatively involved in protein transport and maturation were
242 detected. These encode four outer membrane proteins – the translocation pore Tom40 (RCo005874), its
243 insertase Tob55 (RCo035243, and RCo050551), and Euglenozoa specific proteins Atom69 (RCo012450,
244 and RCo012448) and Atom46 (RCo030973) – as well as two distinct homologues of small Tim
245 (RCo000369, and RCo022175) localised to the intermembrane space, and seven proteins localised in
246 the inner membrane – inner membrane protease subunit 2 (RCo035651), Tim22 (RCo035157), Tim17
247 (RCo003131 and RCo018944), Tim44 (RCo035673), Tim16 (RCo040212), Pam16 (RCo035852), and
248 Oxa1 (RCo002990, RCo023198, and RCo030475). Homologues of both subunits of mitochondrial
249 processing peptidase were detected (RCo000876, and RCo039561); however, the β -subunit may also
250 be part of Complex III. Soluble chaperones, Mge1 (RCo018360) and Hsp60 (RCo005843), are present.
251 17 different transcripts for the chaperon Hsp70 were detected of which two (RCo049920 and

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

254 *Tetrapyrrole synthesis pathways*

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

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

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

296 Discussion

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

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

309 Conversely, the quality of the transcriptome dataset was sufficient for functional annotation. The number
310 of non-redundant predicted proteins (39,585) is comparable to *E. gracilis* (36,526)¹³. Mature euglenid
311 transcripts often contain splice leader (SL) sequences acquired by trans-splicing¹⁹. However, it is likely
312 that not all transcripts require SL for successful translation²⁸ and in the transcriptomes of *E. gracilis*,
313 and *E. longa*, only 54 % and 48.5 % of transcripts, respectively, have been reported to contain at least a
314 fragment of SL^{29,30} somewhere in their sequence. This could be to some extent caused by truncation of
315 the N-termini. We applied a stricter rule for SL identification, in which the SL was only searched for by
316 an exact match at either end of the transcripts (in case of reverse complement orientation). This search
317 obviously revealed a lower fraction of SL-containing transcripts (Supplementary Table S1 online), but
318 was within the range of other euglenid transcriptomes, which vary from 5.7 % in *E. gymnastica* to
319 28.6 % in *E. longa* (*R. costata* transcriptome contains 15.3 % transcripts longer than 9 nucleotides).

320 Almost all eukaryotic genomes, including *E. gracilis*, contain introns that are removed by the
321 ribonucleoprotein complex spliceosome. These conventional introns have the consensus sequence
322 GT(GC)/AG at their ends, and they are excised by two sequential transesterifications. They have been
323 described in genes encoding 13 proteins^{20,28,31–35}. However, the genes of euglenids also contain also
324 nonconventional introns variable in length and with no clear pattern of the nucleotide sequence at the
325 exon/intron junction. They form the stem-loop RNA structures, and their excision is probably
326 independent of the spliceosome, taking place after the excision of spliceosomal introns^{20,36}. Besides
327 these main types of introns, so-called intermediate introns that combine the features of both types have
328 been reported²⁰.

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

332 are likely nonconventional. The presence of nonconventional introns in *R. costata* is expected, as they
333 have been reported from several euglenid species and from a marine diplomonid⁷.

334 The phylogenetic affiliation of predicted *R. costata* proteins is similar to that assessed for *E. gracilis*¹³,
335 with the notable but expected difference in the fraction of genes affiliated with Viridiplantae. The
336 fraction is much higher in *E. gracilis* than in *R. costata* (14 % as compared to 5 %) and reflects the
337 symbiotic history of *E. gracilis*, the plastid of which originated from a green algal endosymbiont. The
338 *in silico* predicted proteome of the *R. costata* mitochondrion is smaller in size than the experimental
339 proteome of *E. gracilis* – 1,554 in *R. costata* vs. approx. 2,500 in *E. gracilis*¹⁸. This difference may be
340 real but also it may reflect the data set completeness and/or the procedure used to generate the set of
341 proteins in the mitochondrial proteome. Direct orthology comparison showed that, of the 1,782
342 experimentally verified proteins of the *E. gracilis* mitochondrion¹⁸, 1,083 have at least one ortholog in
343 *R. costata*, and in total, 1,606 *R. costata* proteins are orthologous to this set. From this comparison and
344 from BUSCO estimation (11.5 % missing), we infer that the completeness of the predicted *R. costata*
345 protein set is reasonably high.

346 The mitochondrion of *E. gracilis* bears a unique combination of metabolic features. It contains a set of
347 enzymes for the facultatively anaerobic metabolism that, in the presence of oxygen, metabolises
348 pyruvate or malate by the pyruvate dehydrogenase complex, followed by a slightly modified TCA cycle,
349 and then the full set of respiratory complexes, including the alternative oxidase also described from
350 kinetoplastids³⁷. In the absence of oxygen, pyruvate is oxidatively decarboxylated by pyruvate:NADH
351 oxidoreductase, and the mitochondrial NADH is recycled by respiratory complex I and rhodoquinone-
352 dependent fumarate reductase, producing succinate and propionyl-CoA. The latter is condensed with
353 acetyl-CoA into fatty acids and wax esters that are stored in the cytoplasm at high concentrations³⁸.
354 These are recycled under aerobic conditions for ATP production or for the synthesis of organic
355 compounds through a functional glyoxylate cycle, uniquely localised in the mitochondrion³⁹.

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

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

367 An interesting difference between *E. gracilis* and *R. costata* is in sulphate metabolism. *E. gracilis* is
368 apparently capable of assimilating sulphate into cysteine in the mitochondrion^{18,40} and sulphite reductase
369 has been detected in its chloroplast fraction¹⁷. In contrast, *R. costata* can only activate sulphate to the
370 form of PAPS, an important coenzyme in sulphotransferase reactions, but the enzymes
371 phosphoadenosine phosphosulfate reductase and sulphite reductase, necessary for the internal
372 production of sulphide from PAPS, were not detected. Still, the presence of sulphide-dependent cysteine
373 synthase suggests that *R. costata* may be able to synthesise cysteine from sulphide putatively produced
374 from PAPS by unclear mechanism or taken up from the anaerobic environment.

375 Heme ranks among the essential cofactors in cellular metabolism because it is involved in many key
376 biochemical processes. In most heterotrophic eukaryotes, the heme synthesis C4 pathway involves eight
377 steps that are localised partially in the mitochondrion and partially in the cytosol. The first step is the
378 condensation of glycine and succinyl-CoA into 5-amino-levulinate by ALAS in the mitochondrial
379 matrix. The next four (or five) steps take place in the cytosol. The pathway ends in the mitochondrion
380 with two reactions that take place in the intermembrane space, and a final reaction occurring in the
381 mitochondrial matrix⁴¹. In most eukaryotes with a plastid, an alternative C5 pathway is present, and all
382 steps are localised in the plastid. In this case, 5-amino-levulinate is synthesised from glutamic acid by
383 three consecutive enzymes: glutamyl-tRNA synthetase (GltX), glutamyl-tRNA reductase (GTR), and
384 glutamate-1-semialdehyde 2, 1-aminomutase (GSA-AT). The pathway then follows the same steps as

385 the classical pathway, but it is localised in the plastid and catalysed by enzymes of mostly cyanobacterial
386 origin^{42,43}. Unexpectedly, *E. gracilis* and the chlorarachniophyte *Bigeloviella natans* (algae with
387 complex plastids that originated from green algae) have both pathways, the mosaic evolutionary origin
388 of their enzymes reflecting the complex evolutionary history of these eukaryotes^{42,44-46}.

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

400 Methods

401 *Strain origin and cultivation*

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

408 *Microscopy*

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

411 paper and fixed with 2.5% glutaraldehyde in 0.1 M cacodylate buffer for 24 hours. Further processing
412 was done by a service lab. The samples were observed using a JEOL JSM-6380LV microscope (JEOL,
413 Akishima, Japan). For transmission electron microscopy, the pelleted cells were fixed in 2.5%
414 glutaraldehyde in 0.1 M cacodylate buffer, postfixed with OsO₄ and the ultrathin sections were
415 contrasted by uranyl acetate and observed using a JEOL JEM-1011 microscope.

416 *Isolation of DNA*

417 We used two methods to obtain high-quality, non-contaminated DNA of *R. costata*. (1) The culture cells
418 were sorted by FACS with the value of the drop diameter equal to 70 µm (length of the cell is around
419 25 µm). Around 2,000 positive drops, most of them containing a single cell of *R. costata* were collected
420 and used for DNA extraction. (2) By a combination of FACS and laser microdissection in which 69
421 drops from FACS were subsequently used for microdissection of individual cells, from which DNA was
422 extracted. Both samples were then subjected to whole genome amplification (WGA; Sigma-Aldrich
423 WGA4-10Rxn) to increase the amount of DNA. After amplification, the samples contained 27.7 µg and
424 32.0 µg of DNA, respectively. PCR with general prokaryotic primers for 16S rRNA had produced
425 a specific product when sample 1 was used as a template, but no product with sample 2 (not shown).
426 This supported our expectation that while sample 1 is still contaminated by bacterial DNA, sample 2 is
427 likely contamination-free.

428 *Genome sequencing and assembly*

429 Both samples were sequenced on the Illumina platform. Initially, two MiSeq runs of sample 2
430 (1 = 250 bp) produced 3.9 Gbp of sequences. Unfortunately, we were not able to produce a reasonable
431 assembly from these data. Therefore, we also sequenced sample 1 using HiSeq (1 = 100 bp) and obtained
432 8.2 Gbp of data. The raw reads from the sequencing of both samples were assembled by SPAdes v3.7.0
433 into 143,763 contigs (36,105 contigs longer than 1,000 bp).

434 *RNA isolation, transcriptome sequencing and assembly*

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

437 polyA-selection; Thermo Fisher Scientific, Waltham, MA, USA) from which the library was prepared,
438 including a polyA-selection step; 2) Isolation of total RNA (GeneAll Hybrid-R RNA purification kit;
439 GeneAll Biotechnology, Seoul, South Korea) followed by mRNA purification (Dynabeads mRNA
440 Direct Kit) from which the library was prepared, including a polyA-selection step; and 3) Isolation of
441 total RNA (GeneAll Hybrid-R RNA purification kit) and from 4 µg of the total RNA, the library was
442 prepared according to the standard TruSeq Stranded mRNA Sample Preparation Guide including a
443 polyA-selection step. Unlike the first two approaches, the latter approach included only one step of
444 polyA selection. The total RNA was always quantified using a Quantus Fluorometer (Promega, Madison,
445 WI, USA) and its quality was checked with an Agilent Bioanalyzer (Agilent Technologies, Santa Clara,
446 CA). All three samples were sequenced on an Illumina MiSeq instrument (Illumina, San Diego, CA,
447 USA) using 150 base-length read chemistry in the paired-end mode. As no principal differences in
448 bacterial contaminations among the three libraries were observed, they were assembled with Trinity
449 v2.0.6 using default parameters and 93,852 contigs were created. We used Transdecoder for basal
450 protein prediction and obtained 55,783 putative proteins. The software package cd-hit v.6⁴⁷ was applied
451 with the default threshold 90 % to remove redundancy. The set was further partially decontaminated by
452 removing proteins with the highest similarity to bacteria *Tolomonas auensis*, strain DSM 9187. The final
453 data set contained 39,585 non-redundant amino acid sequences.

454 *Prediction and characterisation of introns*

455 Introns were automatically predicted by mapping the assembled transcripts to genome contigs using
456 Exonerate (version est2genome 2.2.0). 1164 introns with predicted nonconventional boundaries were
457 selected. Transcript reads were mapped to these contigs by STAR, resulting in 105 contigs in which
458 nonconventional boundaries were supported. Out of these, 26 contigs have good and well-annotated
459 hits on NCBI (e-value < 10⁻⁵) and for these contigs the intron boundaries were manually inspected.
460 The RNA secondary structures were drawn using Varna 3.9⁴⁸ and logo by WebLogo⁴⁹.
461 For selected genes (*tubA*, *tubB*, *tubG*, *hsp90*, *gapC* and *nop1p*), the introns were checked by aligning
462 transcripts to genome contigs. The position of these introns was compared to other euglenids.

463 *Phylogenetic ancestry of R. costata proteins*

464 Ancestry of proteins was assessed with the same methodology as in Ebenezer et al.¹³. Briefly,
465 homologues with e-value < 10⁻² were retrieved from a custom database containing 207 taxa (additional
466 file 3 in Ebenezer et al.¹³), aligned by MAFFT 7.273 with default parameters⁵⁰ and trimmed in trimAl
467 1.2 with default parameters⁵¹. 13,696 alignments with more than 3 taxa and longer than 74 amino acid
468 residues were used for tree reconstruction in RAxML v8.1.17 with 100 rapid bootstraps⁵² in
469 Metacentrum (The National Grid Infrastructure in the Czech Republic). Custom scripts (Python 3.7)
470 were used to sort the trees into bins based on the taxonomic affiliation of the clan in which *R. costata*
471 branched. In 3,445 cases, the tree was included in a taxonomically uniform bin because it contained a
472 bipartition supported by bootstrap 75 or higher comprised of *R. costata* and members of only one other
473 defined taxonomic group.

474 *Prediction of the mitochondrial proteome*

475 The proteome of the mitochondrion was predicted using the following procedure. (1) The complete set
476 of proteins predicted from the transcriptome was BLAST-searched against the Mitominer database, and
477 7,985 proteins with e-value < 10⁻⁴ were selected. To lower the redundancy, only the best *R. costata* hit
478 for each protein included in the Mitominer database was kept, yielding a set of 1,501 proteins. For every
479 protein, the probability of mitochondrial targeting was predicted using TargetP⁵³ and MitoFates⁵⁴ tools.
480 Only proteins with probability of targeting equal to or higher than 0.5 in one or both tools were kept,
481 producing 265 candidates. (2) This initial set was enriched by *R. costata* orthologues of proteins
482 enriched in the mitochondrial fraction of *E. gracilis*¹⁸. Orthology was established using the OrthoMCL
483 software package⁵⁵ as described in Soukal et al. (in preparation). In total, 1,275 proteins fulfilled this
484 criterion, 122 of which were included in the previous step. (3) An additional 121 proteins were included
485 in the list as their presence in mitochondria is very likely – those that are typically part of the respiration
486 chain and protein import complexes, as well as mitochondrial carrier family proteins (Solute carrier
487 family 25). For every gene, the presence of a partial spliced leader (SL) at the 5' terminus and affiliation
488 to prokaryotes was established in the custom database. All proteins mentioned above contained either a
489 partial SL or their best hit was a eukaryote; 179 putative candidates that did not fulfill these criteria were

490 removed during the process of candidate selection. For each entry, the KEGG ID was assigned using
491 the single-directional best hit method in KAAS⁵⁶ or transferred from *E. gracilis* homologue annotation,
492 and the best BLAST hit was identified using blastp against the NCBI nr database. The final predicted
493 mitoproteome consists of 1,539 proteins.

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

499 *Phylogenies*

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

504 Phylogenetic trees of the gene for 18S rRNA and proteins of interest were generated by the following
505 procedure. *R. costata* 18S rRNA/protein(s) and their homologs were downloaded from databases. In the
506 case of the tetrapyrrole biosynthesis pathway, the dataset of Lakey and Triemer⁴⁶ was used as the seed
507 and enriched by *E. gracilis* and *R. costata* transcripts as well as their best NCBI blast hits and the best
508 hits from the local database. All entries were aligned by online version of MAFFT⁵⁰, using the automated
509 strategy, trimmed in BMGE version 1.12⁵⁸, and manually inspected for misaligned positions. The
510 phylogeny, including 1,000 ultrafast bootstraps, was inferred in IQ-Tree 2.0⁵⁹ using the BIC best-
511 selected model (specified in the legends).

512

513

514 References

- 515
- 516 1. Leander, B. S., Lax, G., Karnkowska, A. & Simpson, A. G. B. Euglenida. In Handbook of the
517 Protists 1–42 (Springer International Publishing, 2017).
- 518 2. Lax, G., Lee, W. J., Eglit, Y. & Simpson, A. ploetids represent much of the phylogenetic
519 diversity of euglenids. *Protist* **170**, 233–257 (2019).
- 520 3. Leander, B. S. Did trypanosomatid parasites have photosynthetic ancestors? *Trends Microbiol.*
521 **12**, 251–258 (2004).
- 522 4. Hannaert, V. et al. Plant-like traits associated with metabolism of *Trypanosoma* parasites. *Proc.*
523 *Natl. Acad. Sci. U. S. A.* **100**, 1067–1071 (2003).
- 524 5. Marin, B. Origin and fate of chloroplasts in the Euglenoida. *Protist* **155**, 13–14 (2004).
- 525 6. Busse, I. & Preisfeld, A. Systematics of primary osmotrophic euglenids: a molecular approach
526 to the phylogeny of *Distigma* and *Astasia* (Euglenozoa). *Int. J. Syst. Evol. Microbiol.* **53**, 617–
527 624 (2003).
- 528 7. Gawryluk, R. M. R. et al. Morphological identification and single-cell genomics of marine
529 diplomemids. *Curr. Biol.* **26**, 3053–3059 (2016).
- 530 8. Flegontova, O. et al. Extreme diversity of diplomemid eukaryotes in the ocean. *Curr. Biol.* **26**,
531 3060–3065 (2016).
- 532 9. Gibson, W. Kinetoplastea. (2016). In Handbook of the Protists 1–42 (Springer International
533 Publishing, 2017).
- 534 10. Aurrecochea, C. et al. EuPathDB: the eukaryotic pathogen genomics database resource.
535 *Nucleic Acids Res.* **45**, 581–591 (2017).
- 536 11. Hallick, R. B. et al. Complete sequence of *Euglena gracilis* chloroplast DNA. *Nucleic Acids Res.*
537 **21**, 3537–44 (1993).
- 538 12. Dobáková, E., Flegontov, P., Skalický, T. & Lukeš, J. Unexpectedly streamlined mitochondrial
539 genome of the euglenozoan *Euglena gracilis*. *Genome Biol. Evol.* **7**, 3358–3367 (2015).
- 540 13. Ebenezer, T. E. et al. Transcriptome, proteome and draft genome of *Euglena gracilis*. *BMC Biol.*
541 **17**, 1–23 (2019).
- 542 14. Gockel, G. & Hachtel, W. Complete gene map of the plastid genome of the nonphotosynthetic
543 euglenoid flagellate *Astasia longa*. *Protist* **151**, 347–51 (2000).
- 544 15. Kasiborski, B. A., Bennett, M. S. & Linton, E. W. The chloroplast genome of *Phacus orbicularis*
545 (Euglenophyceae): an initial datum for the Phacaceae. *J. Phycol.* **411**, 404–411 (2016).
- 546 16. Karnkowska, A., Bennett, M. S. & Triemer, R. E. Dynamic evolution of inverted repeats in
547 Euglenophyta plastid genomes. *Sci. Rep.* **8**, 16071 (2018).
- 548 17. Novák Vanclová, A. M. G. et al. Metabolic quirks and the colourful history of the *Euglena*
549 *gracilis* secondary plastid. *New Phytol.* **225**, 1578–1592 (2019).
- 550 18. Hammond, M. J. et al. A uniquely complex mitochondrial proteome from *Euglena gracilis*. *Mol.*
551 *Biol. Evol.* **37**, 2173–2191 (2020).
- 552 19. Frantz, C., Ebel, C., Paulus, F. & Imbault, P. Characterization of trans-splicing in euglenoids.
553 *Curr. Genet.* **37**, 349–355 (2000).
- 554 20. Milanowski, R., Karnkowska, A., Ishikawa, T. & Zakrys, B. Distribution of conventional and
555 nonconventional nitrons in tubulin (a and b) genes of euglenids. *Mol. Biol. Evol.* **31**, 584–593
556 (2014).
- 557 21. Harada, R. et al. Inventory and evolution of mitochondrion-localized family a DNA polymerases
558 in Euglenozoa. *Pathogens* **9**, 257 (2020).
- 559 22. Tanaka, Y. et al. Glucan synthase-like 2 is indispensable for paramylon synthesis in *Euglena*
560 *gracilis*. *FEBS Letters* **591**, 1360–1370 (2017).
- 561 23. Morales, J. et al. Novel mitochondrial complex II isolated from *Trypanosoma cruzi* is composed
562 of 12 peptides including a heterodimeric ip subunit. *J. Biol. Chem.* **284**, 7255–7263 (2009).

- 563 24. Gawryluk, R. & Gray, M. W. A split and rearranged nuclear gene encoding the iron-sulfur
564 subunit of mitochondrial succinate dehydrogenase in Euglenozoa. *BMC Res. Notes* **7**, 1–7
565 (2009).
- 566 25. Kalscheuer, R. & Steinbüchel, A. A novel bifunctional wax ester synthase/acyl-
567 CoA:Diacylglycerol acyltransferase mediates wax ester and triacylglycerol biosynthesis in
568 *Acinetobacter calcoaceticus* ADP1. *J. Biol. Chem.* **278**, 8075–8082 (2003).
- 569 26. Tomiyama, T. et al. Wax ester synthase/diacylglycerol acyltransferase isoenzymes play a
570 pivotal role in wax ester biosynthesis in *Euglena gracilis*. *Sci. Rep.* **7**, 13504 (2017).
- 571 27. Ebenezer, T. E. et al. Transcriptome, proteome and draft genome of *Euglena gracilis*. *BMC Biol.*
572 **17**, 11 (2019).
- 573 28. Russell, A. G., Watanabe, Y. I., Charette, J. M. & Gray, M. W. Unusual features of fibrillar
574 cDNA and gene structure in *Euglena gracilis*: Evolutionary conservation of core proteins and
575 structural predictions for methylation-guide box C/D snoRNPs throughout the domain Eucarya.
576 *Nucleic Acids Res.* **33**, 2781–2791 (2005).
- 577 29. Yoshida, Y. et al. De novo assembly and comparative transcriptome analysis of *Euglena gracilis*
578 in response to anaerobic conditions. *BMC Genomics* **17**, 1–10 (2016).
- 579 30. Záhonová, K. et al. Peculiar features of the plastids of the colourless alga *Euglena longa* and
580 photosynthetic euglenophytes unveiled by transcriptome analyses. *Sci. Rep.* **8**, 1–15 (2018).
- 581 31. Breckenridge, D. G., Watanabe, Y., Greenwood, S. J., Gray, M. W. & Schnare, M. N. U1 small
582 nuclear RNA and spliceosomal introns in *Euglena gracilis*. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 852–
583 6 (1999).
- 584 32. Ebel, C., Frantz, C., Paulus, F. & Imbault, P. Trans-splicing and cis-splicing in the colourless
585 euglenoid, *Entosiphon sulcatum*. *Curr. Genet.* **35**, 542–50 (1999).
- 586 33. Canaday, J., Tessier, L. H., Imbault, P. & Paulus, F. Analysis of *Euglena gracilis* alpha-, beta- and
587 gamma-tubulin genes: introns and pre-mRNA maturation. *Mol. Genet. Genomics* **265**, 153–60
588 (2001).
- 589 34. Vesteg, M. et al. A possible role for short introns in the acquisition of stroma-targeting
590 peptides in the flagellate *Euglena gracilis*. *DNA Res.* **17**, 223–231 (2010).
- 591 35. Milanowski, R., Gumińska, N., Karnkowska, A., Ishikawa, T. & Zakryś, B. Intermediate introns in
592 nuclear genes of euglenids - are they a distinct type? *BMC Evol. Biol.* **16**, 49 (2016).
- 593 36. Gumińska, N., Płecha, M., Zakryś, B. & Milanowski, R. Order of removal of conventional and
594 nonconventional introns from nuclear transcripts of *Euglena gracilis*. *PLoS Genet.* **14**,
595 e1007761 (2018).
- 596 37. Hellemond, J. J. Van, Bakker, B. M. & Tielens, A. G. M. Energy metabolism and its
597 compartmentation in *Trypanosoma brucei*. *Adv. Microb. Physiol.* **50**, 199–226 (2005).
- 598 38. Tucci, S., Vacula, R., Krajcovic, J., Proksch, P. & Martin, W. Variability of wax ester fermentation
599 in natural and bleached *Euglena gracilis* strains in response to oxygen and the elongase
600 inhibitor flufenacet. *J. Eukaryot. Microbiol.* **57**, 63–69 (2010).
- 601 39. Müller, M. et al. Biochemistry and evolution of anaerobic energy metabolism in eukaryotes.
602 *Microbiol. Mol. Biol. Rev.* **76**, 444–95 (2012).
- 603 40. Saidha, T., Stern, A. I., Lee, D. H. & Schiff, J. A. Localization of a sulphate-activating system
604 within *Euglena* mitochondria. *Biochem. J.* **232**, 357–365 (1985).
- 605 41. Hamza, I. & Dailey, H. A. One ring to rule them all: Trafficking of heme and heme synthesis
606 intermediates in the metazoans. *Biochimica et Biophysica Acta* **1823**, 1617–1632 (2012).
- 607 42. Kořený, L. & Oborník, M. Sequence evidence for the presence of two tetrapyrrole pathways in
608 *Euglena gracilis*. *Genome Biol. Evol.* **3**, 359–364 (2011).
- 609 43. Cenci, U. et al. Heme pathway evolution in kinetoplastid protists. *BMC Evol. Biol.* **16**, 109
610 (2016).
- 611 44. Woo, Y. H. et al. Chromerid genomes reveal the evolutionary path from photosynthetic algae
612 to obligate intracellular parasites. *Elife* **4**, 1–41 (2015).

- 613 45. Cihlář, J., Füßy, Z., Horák, A. & Oborník, M. Evolution of the tetrapyrrole biosynthetic pathway
614 in secondary algae: Conservation, redundancy and replacement. *PLoS One* **11**, e0166338
615 (2016).
- 616 46. Lakey, B. & Triemer, R. The tetrapyrrole synthesis pathway as a model of horizontal gene
617 transfer in euglenoids. *J. Phycol.* **53**,198-217 (2017).
- 618 47. Li, W. & Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or
619 nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
- 620 48. Crooks, G. E. et al. VARNA: Interactive drawing and editing of the RNA secondary structure.
621 *Bioinformatics* **25**, 1974–1975 (2009).
- 622 49. Crooks, G. E., Hon, G., Chandonia, J. & Brenner, S. E. WebLogo: A sequence logo generator.
623 *Genome Res.* **14**, 1188–1190 (2004).
- 624 50. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7:
625 improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–80 (2013).
- 626 51. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. TrimAl: a tool for automated
627 alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
- 628 52. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
629 phylogenies. *Bioinformatics* **30**, 1312–3 (2014).
- 630 53. Emanuelsson, O., Nielsen, H., Brunak, S. & Von Heijne, G. Predicting subcellular localization of
631 proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* **300**, 1005–1016 (2000).
- 632 54. Fukasawa, Y. et al. MitoFates: Improved prediction of mitochondrial targeting sequences and
633 their cleavage sites. *Mol. Cell. Proteomics* **14**, 1113–1126 (2015).
- 634 55. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: Identification of ortholog groups for eukaryotic
635 genomes. *Genome Res.* **13**, 2178–2189 (2003).
- 636 56. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C. & Kanehisa, M. KAAS: An automatic genome
637 annotation and pathway reconstruction server. *Nucleic Acids Res.* **35**, W182-5 (2007).
- 638 57. Müllner, A. N., Angeler, D. G., Samuel, R., Linton, E. W. & Triemer, R. E. Phylogenetic analysis of
639 phagotrophic, phototrophic and osmotrophic euglenoids by using the nuclear 18S rDNA
640 sequence. *Int. J. Syst. Evol. Microbiol.* **51**, 783–791 (2001).
- 641 58. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new
642 software for selection of phylogenetic informative regions from multiple sequence alignments.
643 *BMC Evol. Biol.* **10**, 210 (2010).
- 644 59. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective
645 stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–
646 274 (2015).
- 647

648 Author contributions

649 VH and PS conceived the study, IČ provided the culture, JS prepared clonal lineage and sequenced the
650 18S rRNA gene, PS prepared the amplified gDNA, MH prepared the gDNA libraries and sequenced
651 them, HS and ČV assembled the genome, ŠH prepared the cDNA libraries, PS assembled the
652 transcriptome, annotated the splice leaders and analysed the phylogenetic origin of genes, ŠH and VH
653 annotated mitochondrial metabolism and heme synthesis, VH computed the trees, AK and RM
654 searched for and annotated the introns, VH and ŠH wrote the manuscript, PS, IČ, AK and RM edited
655 the manuscript All authors approved the final version.

656 Acknowledgement

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

669 Additional information

670 The authors declare no competing interests.
671

672 Data availability statement

673 The transcriptomic and genomic reads are available in GenBank under the BioProject ID PRJNA550357,
674 the assemblies and predicted proteins are available in the Zenodo repository
675 (<https://zenodo.org/record/4249289#.X6UIDWhKhPY>). The sequence of the gene for 18S rRNA is
676 deposited under GenBank accession nr. MW113742.

677

678

679

680

681

682 Figure legends

683

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

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

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

700 **Fig. 4: KEGG Functional categories of predicted proteins of *R. costata*.** 13,130 proteins (33 % of
701 39,585) were ascribed to functional categories according to KEGG. All categories with less than 10
702 members were merged into the category “other”.

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

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

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

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

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

724

725

726

727

728

729

730

731

732

733

734

735

736

737 **Tables**

738 **Table 1:** Parameters of the genomic assembly.

739

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

740

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 $\leq 10^{-5}$)	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	<i>tubA</i>	2	0	1356	100
β-tubulin	<i>tubB</i>	4	0	1335	100
γ-tubulin	<i>tubG</i>	6	7	1680	76
Heat shock protein 90	<i>hsp90</i>	10	5	2112	68
GAPDH	<i>gapC</i>	3	1	1062	70
Fibrillarin	<i>nop1p</i>	4	1	903	85

750

751

752

Figures

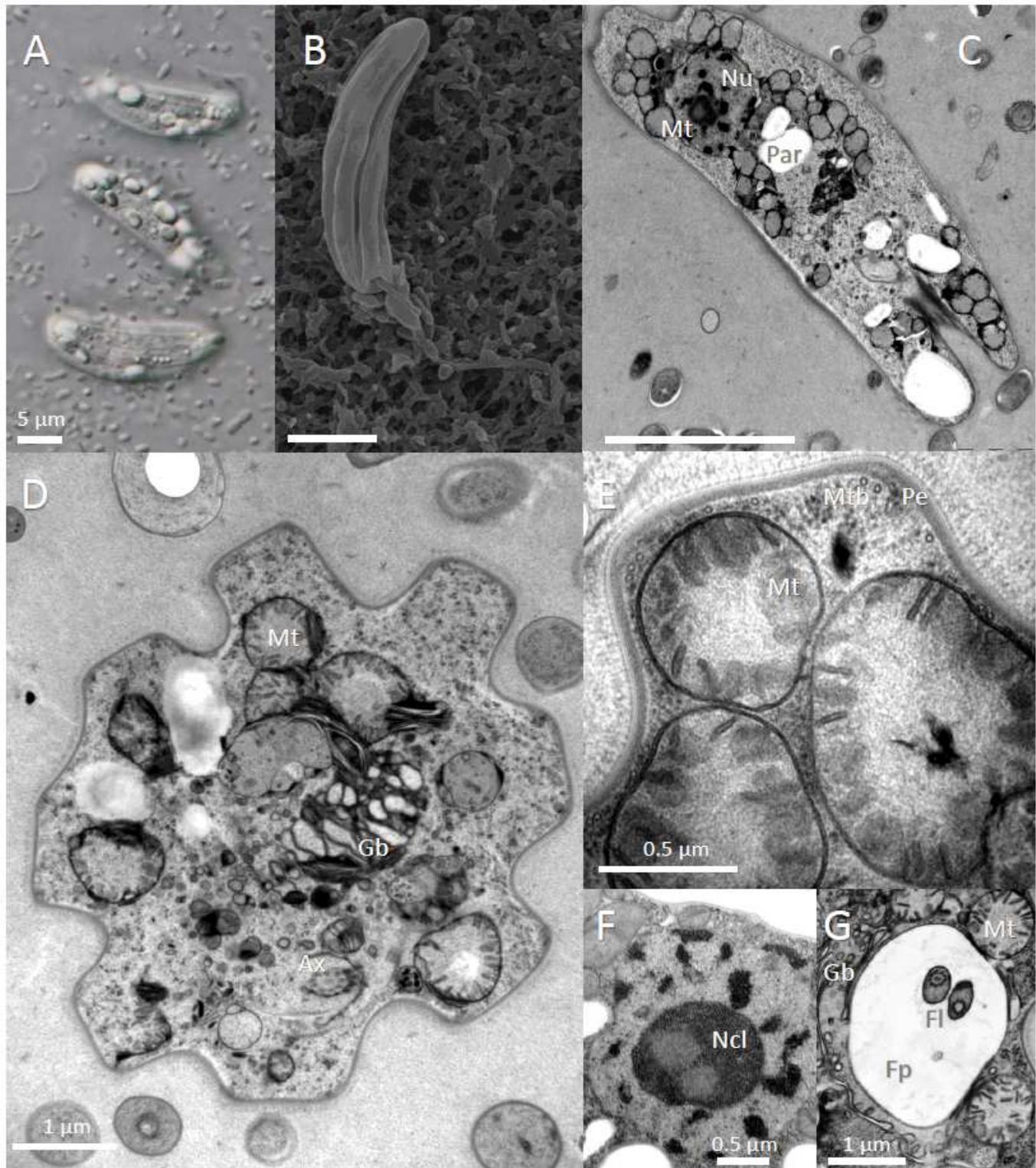


Figure 1

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.

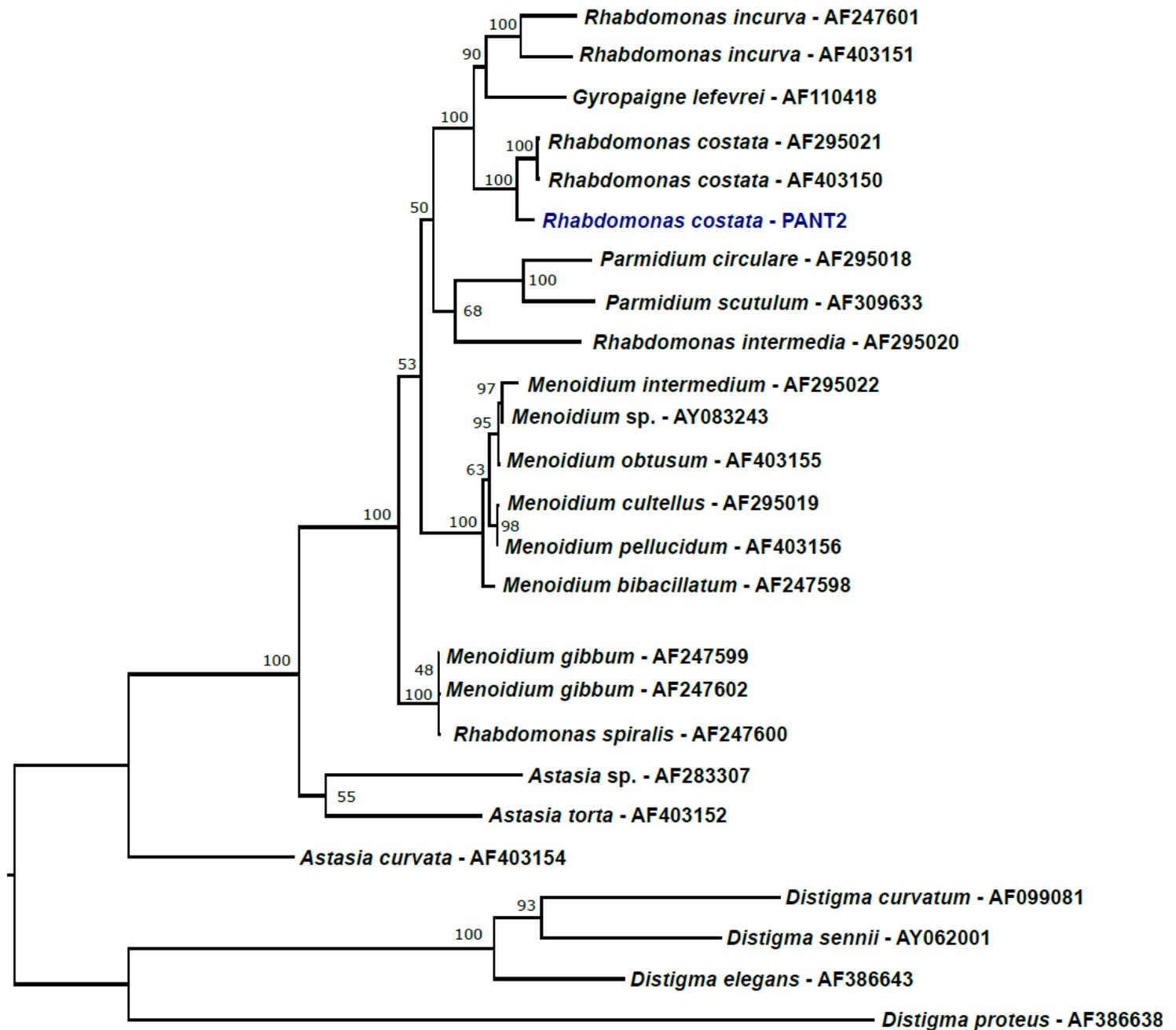


Figure 2

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*.

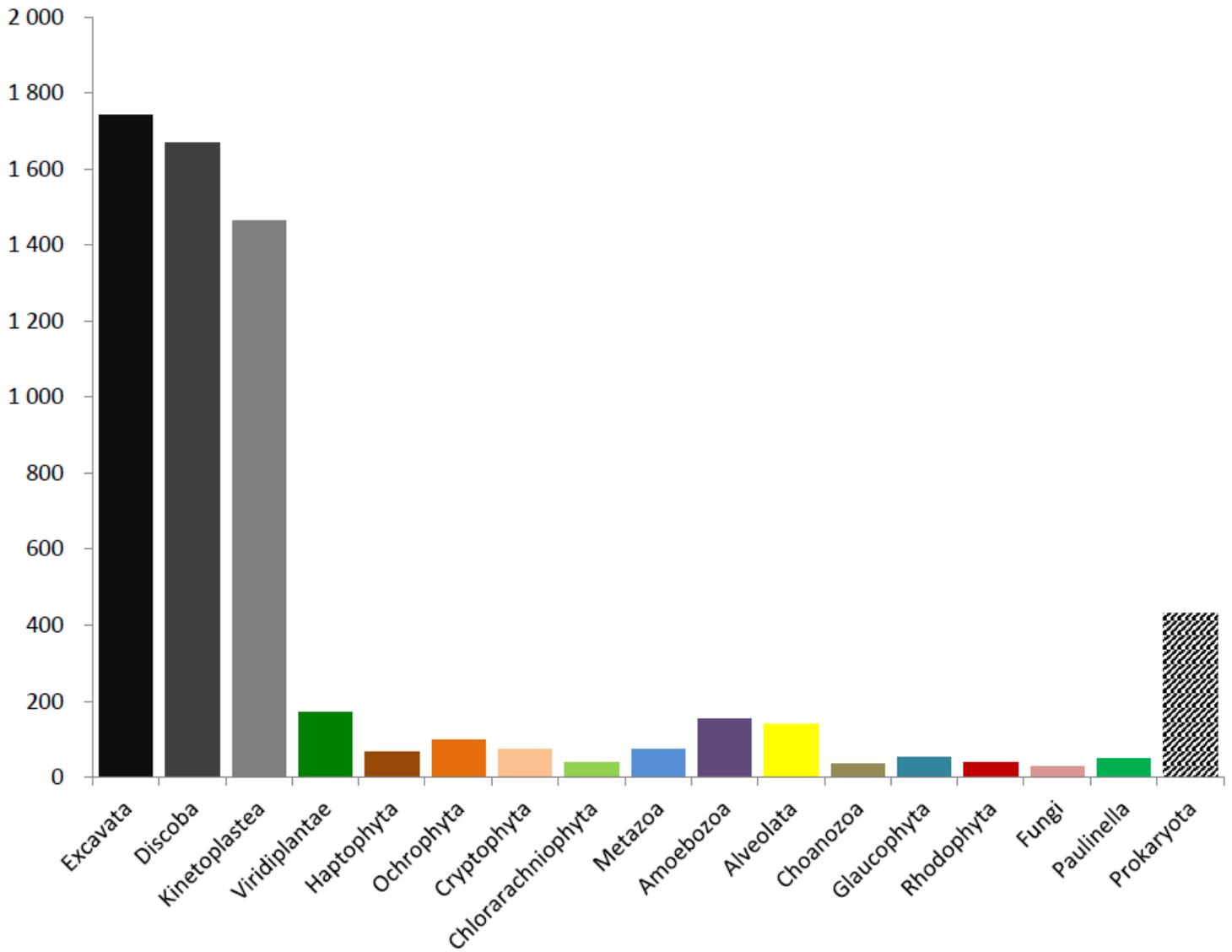


Figure 3

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

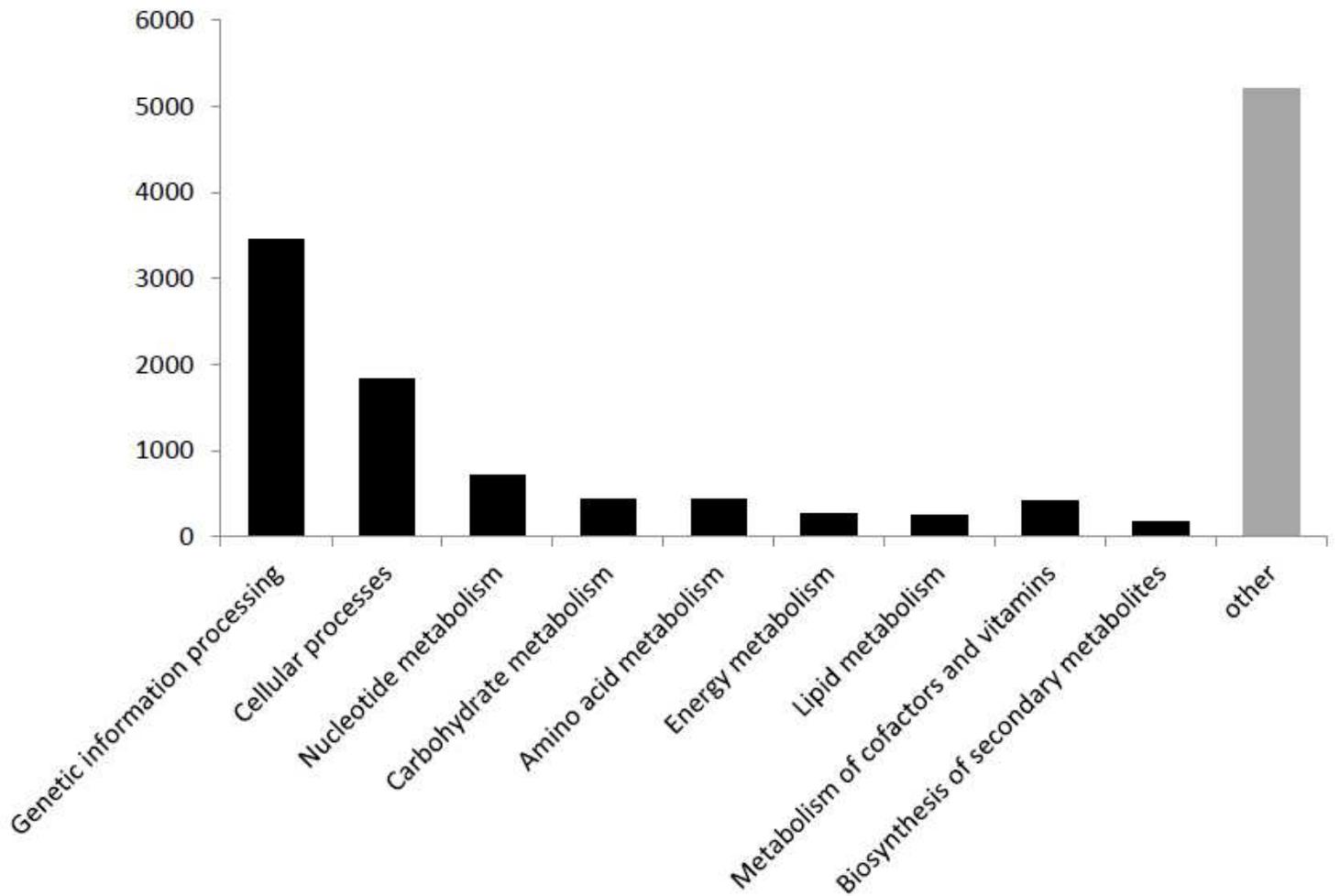
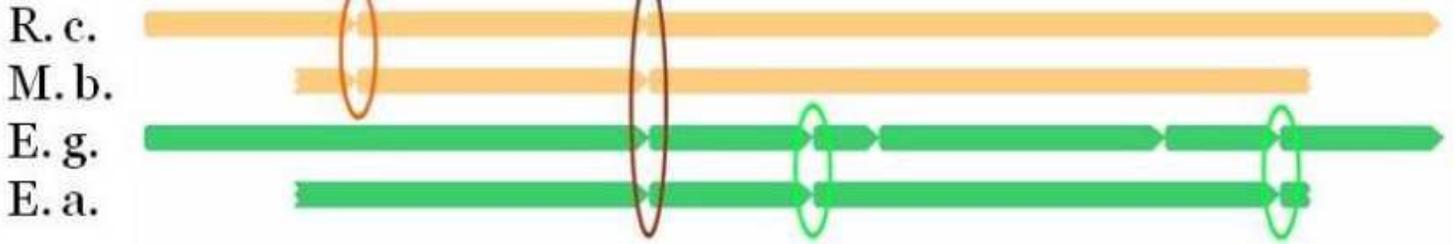


Figure 4

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".

Tubulin α



Tubulin β

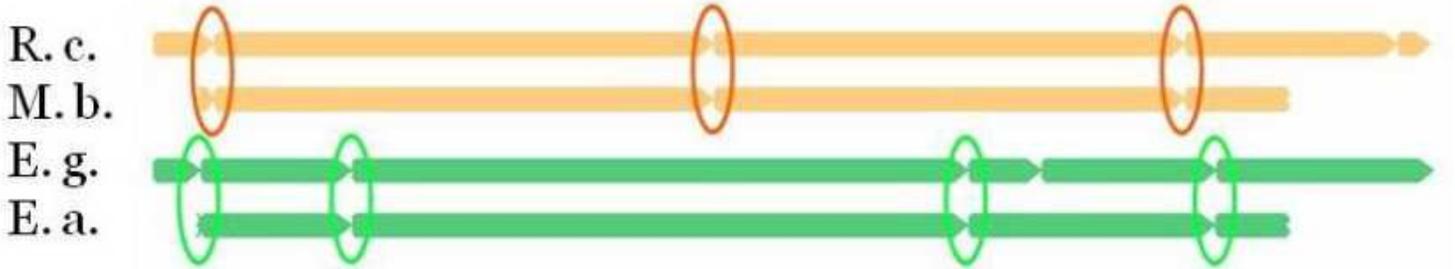
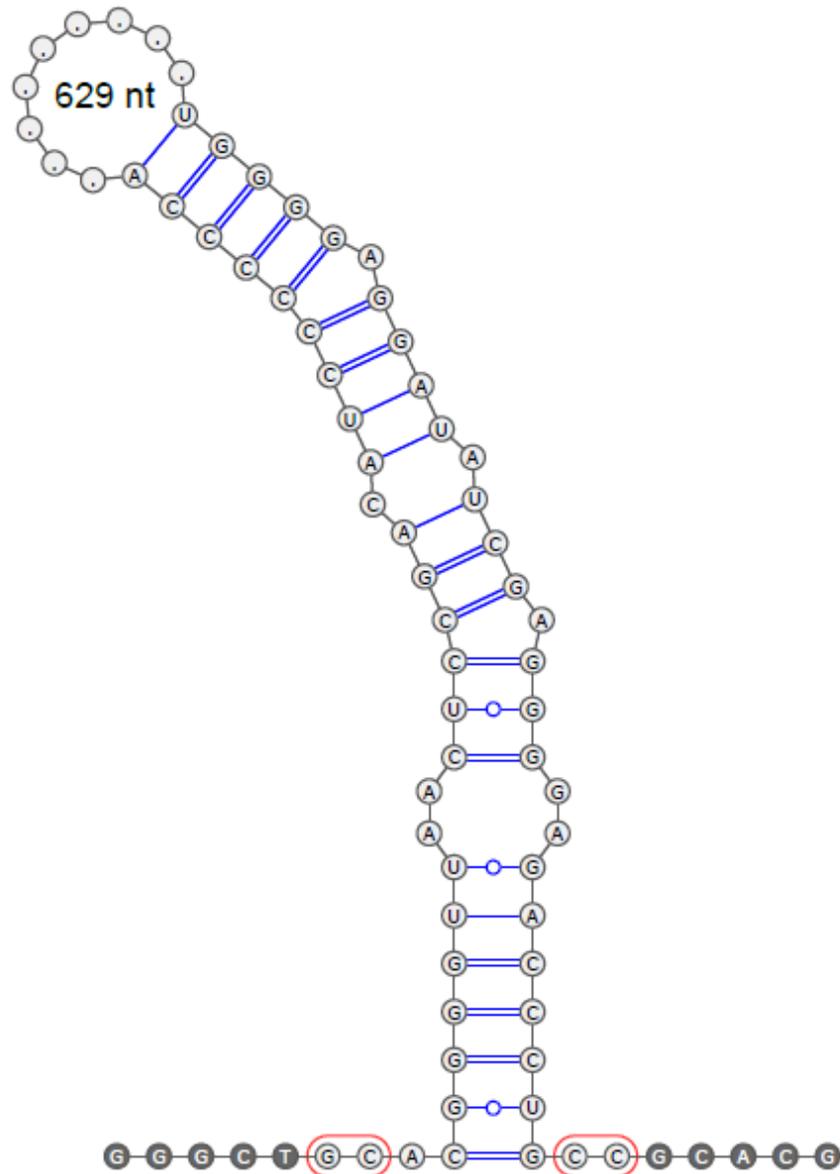
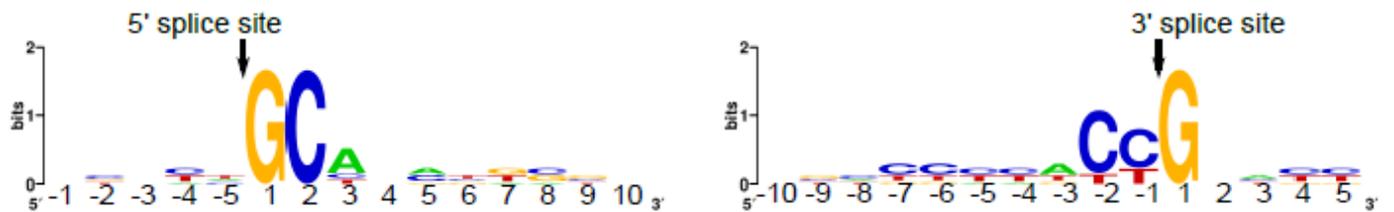


Figure 5

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.

A**B****Figure 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).

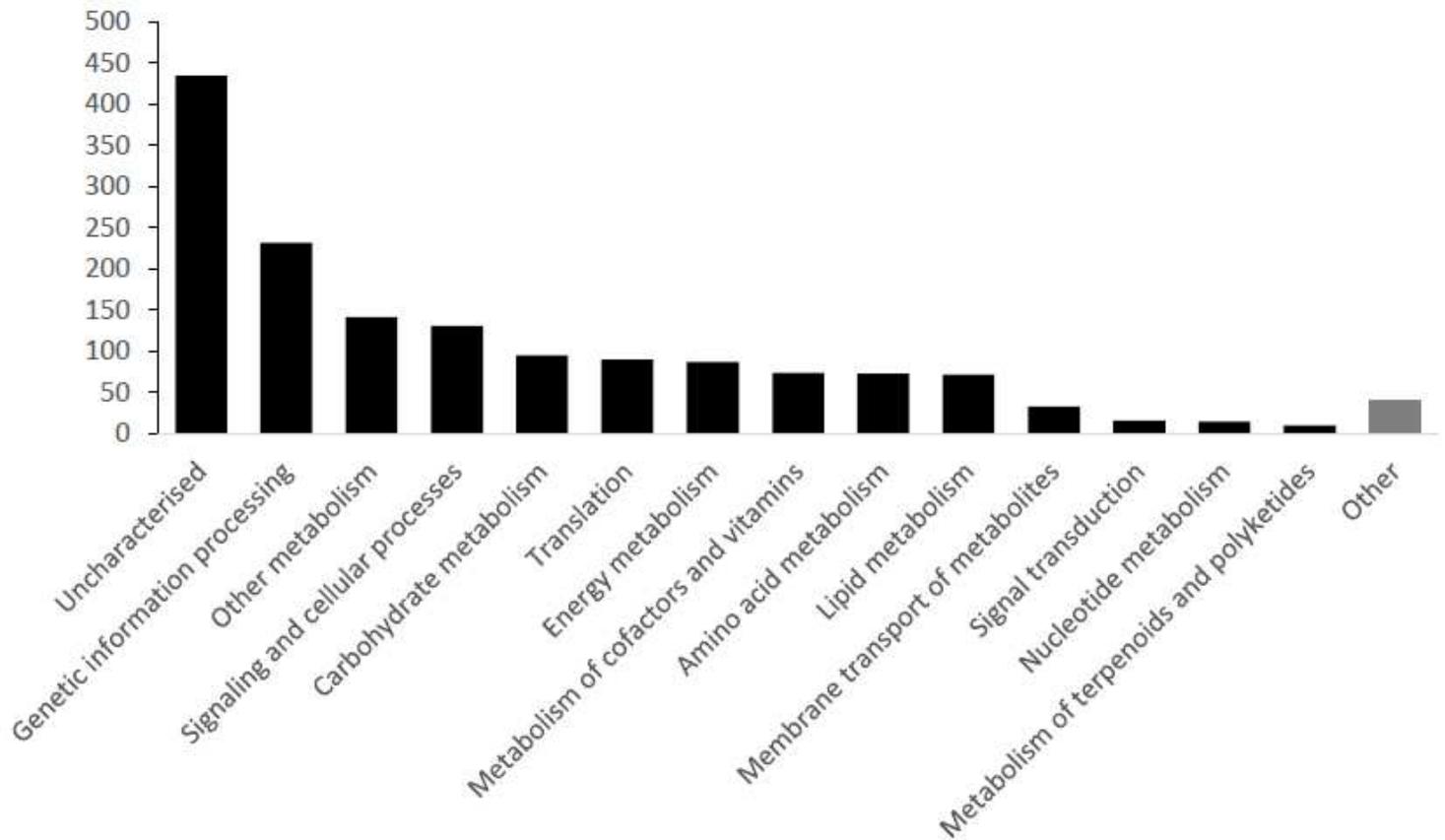


Figure 7

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”.

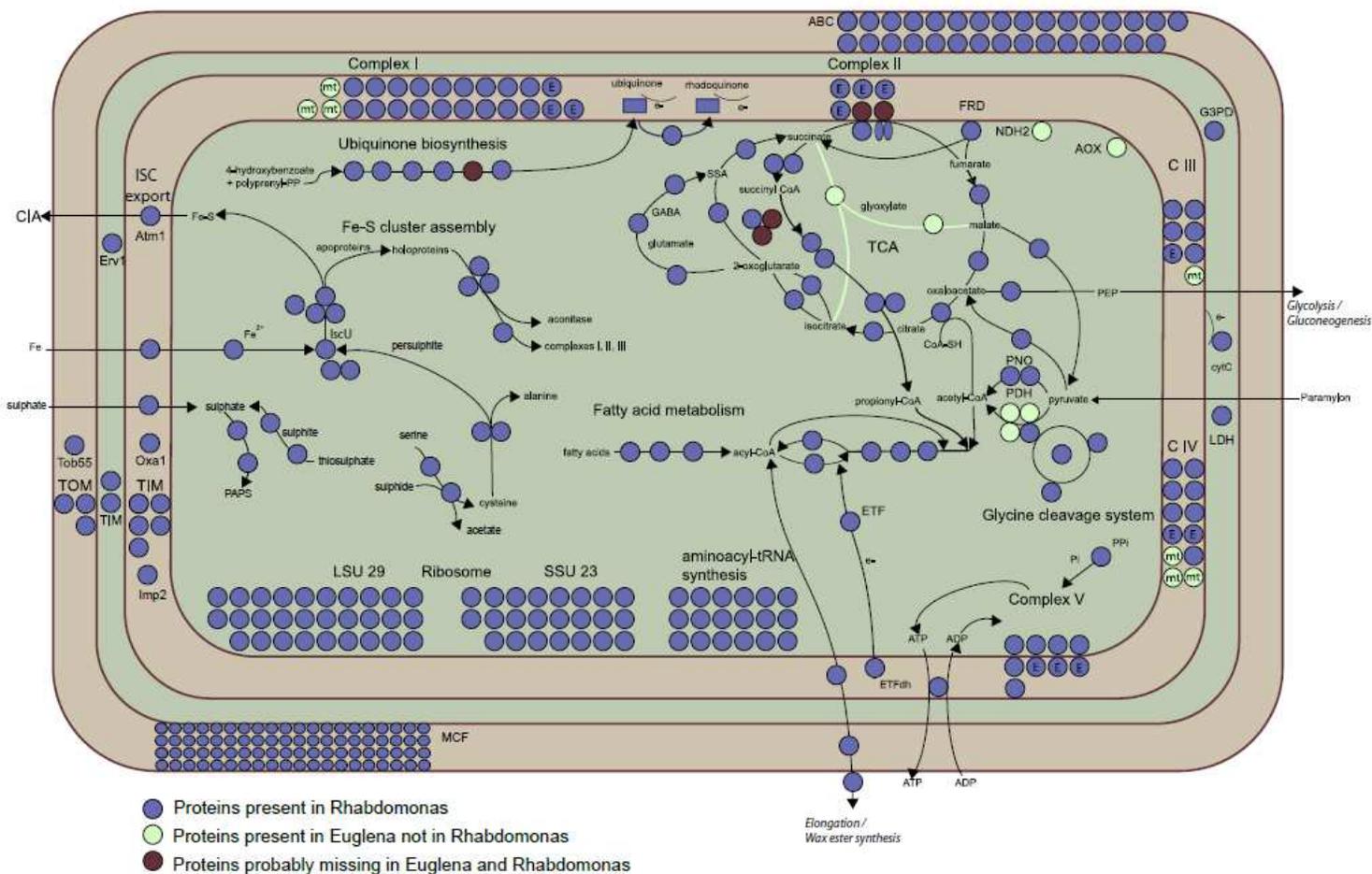


Figure 8

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.

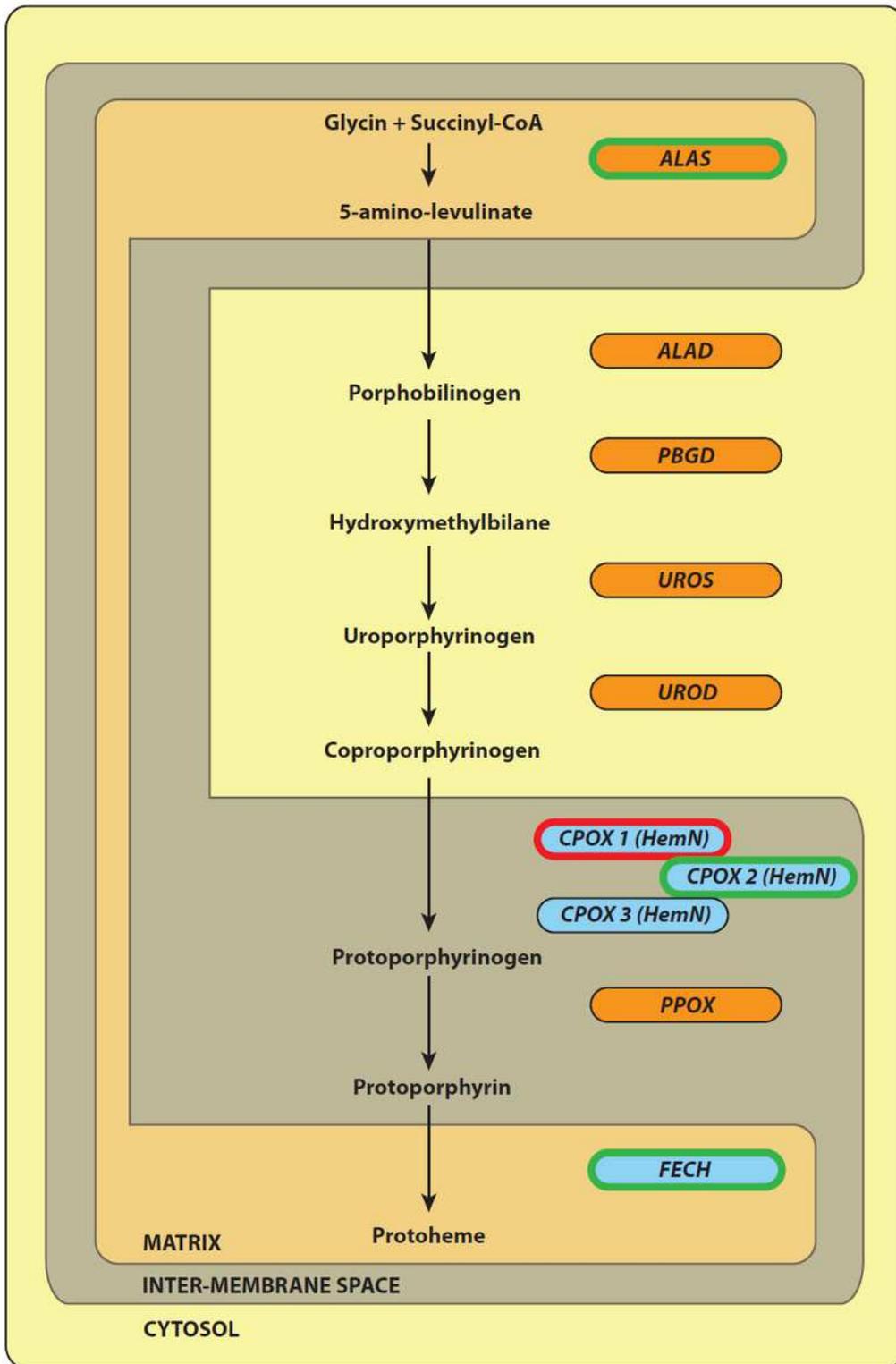


Figure 9

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).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [TableS1S8.xlsx](#)
- [FiguresS1S11.pdf](#)