Elsevier Editorial System(tm) for Mitochondrion
Manuscript Draft

Manuscript Number: MITOCH-D-13-00193R1

Title: The mitochondrial respiratory chain of the secondary green alga Euglena gracilis shares many additional subunits with parasitic Trypanosomatidae.

Article Type: Plant Mitochondrial Biolo

Keywords: Euglenozoa, Trypanosomatidae, OXPHOS, Proteomics, Sequence database mining, Large-scale phylogenetics

Corresponding Author: Dr. Pierre Cardol, PhD

Corresponding Author's Institution: University of Liege

First Author: Emilie Perez

Order of Authors: Emilie Perez; Marie Lapaille, PhD; Hervé Degand; Laura Cilibrasi; Alexa Villavicencio-Queijeiro; Pierre Morsomme; Diego González-Halphen; Mark C Field; Claire Remacle; Denis Baurain; Pierre Cardol, PhD

Abstract: The mitochondrion is an essential organelle for the production of cellular ATP in most eukaryotic cells. It is extensively studied, including in parasitic organisms such as trypanosomes, as a potential therapeutic target. Recently, numerous additional subunits of the respiratory-chain complexes have been described in Trypanosoma brucei and Trypanosoma cruzi. Since these subunits had apparently no counterparts in other organisms, they were interpreted as potentially associated with the parasitic trypanosome lifestyle. Here we used two complementary approaches to characterise the subunit composition of respiratory complexes in Euglena gracilis, a non-parasitic secondary green alga related to trypanosomes. First, we developed a phylogenetic pipeline aimed at mining sequence databases for identifying homologs to known respiratory-complex subunits with high confidence. Second, we used MS/MS proteomics after two-dimensional separation of the respiratory complexes by Blue Native- and SDS-PAGE to both confirm in silico predictions and to identify further additional subunits. Altogether, we identified 41 subunits that are restricted to E. gracilis, T. brucei and T. cruzi, along with 48 classical subunits described in other eukaryotes (i.e. plants, mammals and fungi). This moreover demonstrates that at least half of the subunits recently reported in T. brucei and T. cruzi are actually not specific to Trypanosomatidae, but extend at least to other Euglenozoa, and that their origin and function are thus not specifically associated with the parasitic lifestyle. Furthermore, preliminary biochemical analyses suggest that some of these additional subunits underlie the peculiarities of the respiratory chain observed in Euglenozoa.

Reviewer #1: The authors have carried out a comprehensive (indeed exhaustive) analysis of the protein components of the mitochondrial electron transport chain (ETC) of the euglenozoan, Euglena gracilis. They report that the E. gracilis ETC contains (besides "classical" components broadly distributed throughout eukaryotes) a number of additional proteins that had previously been described in (and were thought to be specific to) another euglenozoan group, the kinetoplastids. This conclusion is a novel and important one because it implies that these additional proteins arose at an early stage in the evolution of Euglenozoa, before the separation of the three main euglenozoan lineages.

That being said, and recognizing that the authors have used a combination of complementary approaches in their analysis, I have a number of concerns (mostly concerning the way in which the data and results are presented) that should be addressed prior to publication.

1. No primary MS data are presented (i.e., a list of identified peptides with ion scores) that would allow a reader to assess the robustness of the identification of the various euglenid proteins. Inclusion of a supplemental table summarizing such data would be helpful.

R: Primary MS data are now given in Supplemental Table 3 for each polypeptide for which the score was > 60.

2. The most important conclusion of this paper is that novel ETC components previously found only in kinetoplastids have counterparts in E. gracilis. However, no objective data (BLAST alignments, E-values, etc.) are actually presented in the paper to support the inference that a particular E. gracilis protein is a true homolog of a given kinetoplastid protein.

R: We edited the text in several places to better describe our general approach, which is akin to the strategy used to build large-scale datasets for phylogenomic studies.

The main issue when collecting homologous sequences is to ensure a good representation of the organism diversity without being overwhelmed by the thousands sequences stored in contemporary public databases. This is especially true when dealing with multigenic families, in which some paralogous genes can actually be more similar to a given reference sequence than orthologous genes that have extensively diverged. This is why it is unwise to carry out simple BLAST searches on public databases and then to « choose » an E-value threshold « where to cut the BLAST report » for separating orthologues from paralogues. Even if common, this approach is the perfect recipe for missing orthologues and including paralogues, which is very problematic in the context of a study aimed at assessing the taxonomic distribution of one hundred (sometimes poorly conserved) genes.

Instead, we took a series of measures to ensure a reasonably sensitive yet specific survey of respiratory complex subunits across the eukaryotic diversity. These methodological choices yielded conservative identifications, i.e., we may have missed some genuinely orthologous sequences, but have not included paralogous sequences, especially from homologous genes functioning in other cell compartments than the mitochondrion, such as the plastid.

1. We compiled our own database of eukaryotic and bacterial complete proteomes.
2. We compared each of the 422 235 proteins in our database to all the others using BLASTP on a computer cluster. This allowed us to build a graph relating the proteins based on their sequence similarity. The arcs of this graph were re-scaled to weaken the links between out-paralogues and to strengthen the links between orthologues (including in-paralogues). This part was done using OrthoMCL algorithms. Finally, we used Markov clustering (MCL) to split the graph into a few thousands individual groups, each one corresponding to a set of

orthologous proteins, possibly supplemented by in-paralogues (e.g., splicing variants or genuine recent gene duplications).

3. Using an inventory of mitochondrial reference proteins, we identified the orthologous groups containing proteins similar to respiratory complex subunits by BLASTing each sequence from each group against a database built from our inventory. Careful graphical analysis of the BLAST reports allowed us to eventually select one orthologous group for each subunit. All the other groups were discarded, as they consisted of proteins that have nothing to do with respiratory complexes or that are only distantly related to them (out-paralogues or proteins sharing one or more domains with respiratory complex subunits).
4. When including bacteria in a collection of eukaryotic sequences of endosymbiotic origin, all are technically orthologous. Indeed, eukaryotic sequences can be seen as « bacterial » lineages, either related to cyanobacteria (plastidial sequences) or to alpha-proteobacteria (mitochondrial sequences). That is why they often end up in a single orthologous group, even if from the point of view of a eukaryotic cell, the plastidial and mitochondrial genes are paralogues obtained in the course of two different endosymbiotic events. Since we are only interested in mitochondrial proteins, we don't want to count plastidial proteins in our inventory. Thus, for each orthologous group, we built a phylogenetic tree (as exemplified in Supplemental Figure 1) in order to identify and prune out the subtree corresponding to the mitochondrial gene (the « mitochondrial paralogue »). In some cases, the tree was quite complicated with much more than two subtrees (including distantly related cytosolic paralogous genes) and we had to proceed in several steps to reliably identify the mitochondrial paralogues (Supplemental Figure 2 and Figure 3).
5. To search for the corresponding subunits in Euglenozoa, we built HMM profiles from an alignment of the sequences included in each mitochondrial subtree and used these profiles to mine the complete proteomes of our selection of Euglenozoa. To ensure identifying only genuine mitochondrial subunits (if any) in euglenozoan genomes, we used the HaMStR pipeline, which applies a best reciprocal hit (BRH) orthology criterion on each HMM hit to filter out paralogues. Briefly, if BLASTing a candidate HMM hit to a set of complete reference genomes does not yield a sequence that is part of the original HMM profile as the first hit, the candidate is tagged as a paralogous gene and discarded.
6. When some subunits appeared to be specific to Euglenozoa, we used the same HMM profiles to mine a much larger database of broadly sampled eukaryotes and prokaryotes to ensure that we had not missed homologues in non-model organisms. The eukaryotic part of this database was also used to study in detail the evolution of QCR-like subunits.

We hope that this (admittedly lengthy) explanation clarifies our approach and will convince the reviewers that we worked and reported our results adequately in this revision of our manuscript.

At the very least, a table listing the novel E. gracilis components along with their supposed kinetoplastid homologs, together with a statement summarizing the basis for the assignment in each case, should be included in the paper. Whether the component was detected in more than one analytical condition (e.g., in silico vs. MS) should also be indicated. Although Fig. 2 is an effective way of displaying the distribution of various ETC subunits, it doesn't readily distinguish between "classical" components and the novel ones that are the subject of this paper. All the more reason to have a separate table that explicitly identifies the novel components for the reader.

R: According to the reviewer, Figure 2 is an effective way of displaying the various ETC subunits. Because we really feel that a table encompassing a list of names is neither appropriated nor useful, and in order to easily distinguish between classical components and the novel ones, we changed the fonts (now in bold and italic) of names for novel components. Thus Figure 2 has been slightly improved: the novel conserved components between E. gracilis and trypanosomes that are specific

to Euglenozoa are now highlighted. A sentence has also been added at the end of the Results to draw attention on this aspect.

3. There is inconsistency between the names used throughout the paper and in Fig. 2 and the annotation in the accompanying supplemental files. This makes it difficult to retrieve a given sequence from the fasta files. For examples, a simple text search for "NDTB1", "NDTB28", "NDTB31", "NDTB34" and "COB", among many others, fails to retrieve any sequences from any of the euglenid fasta files (these 5 entries are, however, present in the Tbru_ID-MCL_bioinfo_genes.fasta filel). Few of the CV "ATP" proteins shown in Fig. 2 are annotated as such. The fasta files should be re-annotated to ensure that there is consistency with the corresponding gene names used in the text and figures.

R: We apologize for the discrepancies between the nomenclature given in the text and in the original Supplemental Files. All entries have been carefully checked to ensure that a unique name is used across every file and figure.

4. There are other problems with Fig. 2. For example, underlining is supposed to indicate "proteins found in E. gracilis by proteomic analysis". However, NDUFA8, which is underlined, is shown outside the euglenid cluster, as is ATPA, the sequence of which (designated "alpha") is, in fact, present in the Egra_ID-MCL_bioinfo_genes.fasta file.

R: NDFUA8 is now placed inside the euglenid cluster and we added a comment in the figure legend. We thank the reviewer for drawing our attention on this mistake.
In contrast, ATPA subunit corresponds to ATP6 /subunit A (not alpha) from the Fo part (see Supplemental Table 1). This subunit does not correspond to subunit alpha from F1 catalytic fraction. We are sorry but the nomenclature of genes / proteins in respiratory complexes is hell.

5. According to Fig. 2, none of the core CI subunits that trace their ancestry to the alpha-proteobacterial ancestor of the mitochondrion (i.e., ND1-ND11 and ND4L) was detected by MS, and only 3 (ND1, ND4 and ND5) are shown as being identified computationally. Do the authors have any explanation for what seems to me a rather puzzling absence, given that they identify many "accessory" CI proteins? In the case of ND1, ND4 and ND5, is there any indication that these are encoded in the nuclear genome; or, do the authors think they are derived from mtDNA as part of their 454 sequencing of the E. gracilis genome? How certain are the authors of their identification of ND1, ND4 and ND5 as bona fide mitochondrial components?

R: There is a misunderstanding here. ND1–11+ND4L are the 12 subunits encoded by mitochondrial genes in Reclinomonas: NAD1,2,3,4,5,6,4L + NAD8=TYKY=NDUFS8, NDUFS3=30kD=NAD9, NDUFS7=NAD10=PSTT, NDUFS1=75kD=NAD11, NDUFS7=NAD7=49kD. This nomenclature has been added to Supplemental Table 1. These 12 subunits along with 24kD=NDUFV2 and 51kD=NDUFV1 subunits are the 14 subunits that compose complex I in most bacteria. Seven of these subunits are nucleus-encoded in mammals and yeasts and are hydrophilic subunits. 6 of them subunits have been identified as bona-fide components of Euglena complex I in our proteomic analyses (see Figure 2) while the last one (NDUFS8/NAD8) is identified by *in silico* analysis. Among the seven hydrophobic components (ND1-6 + ND4L), only three, ND1, ND4, ND5 have been identified by the *in silico* analysis based in the presence of the 454 reads, probably corresponding to mtDNA sequences. Usually, highly hydrophobic subunits are difficult to resolve on SDS gels. To clarify this point, we slightly modified paragraph 1.1 in the Results and we added a short paragraph in the first part of Discussion about the limitation of our proteomic analysis. To support the homology for ND1,4,5 sequences with mitochondrial components, we provided an additional example of phylogenetic trees that show mitochondrial and plastidial paralogues (Figure 3). In each of these trees, we identified and analysed

the subtree corresponding to the mitochondrial paralogue (by opposition to the plastidial paralogue) by its proximity to the two α-proteobacteria. The approach is explained at the end of the first section of the Results.

6. In 2.1 (Complex I) on pg. 19, the authors mention a number of proteins (e.g., G3PD, DNAJ) identified in the proteomic analysis of CI, and they imply that these proteins might be specifically associated with CI, albeit not actual components of the complex. However, there is no assessment in the manuscript of the purity of their mitochondrial fraction, or consideration of possible contamination of their mitochondrial fraction with non-mitochondrial proteins. What other proteins did the authors find during their proteomic analysis of isolated ETC complexes? On what basis did they conclude that some 'non-ETC' proteins might be specifically associated with these complexes rather than artifactual, co-migrating contaminants?

R: To assess the purity of the mitochondrial fraction, we first tested their chlorophyll content. This is now indicated in the Methods section 3.
In addition, based on the identification we performed on 2D gels, only few annotated proteins that perform other functions have been identified (see Supplemental Table S3). Thus, although we cannot rule out the possibility that DNAJ/HSP40 and G3PD might be non-mitochondrial contaminations of the mitochondrial fraction, we propose that they could be part of complex I. As already mentioned in the Discussion, DNAJ was also found associated to complex I in trypanosomes. This point is now further discussed in section 2.1 of the Discussion.
At last, to test the purity of the mitochondrial fraction, we performed a preliminary analysis of the total protein content of the mitochondrial fraction by LC-MS/MS analysis. Among the ~400 proteins that matched known proteins in public databases (E-value < 10^{-10}), > 80% corresponded to known mitochondrial components, and only ~5% to cytoplasmic components. This is now indicated in the Methods section 3.

Minor points:

1. pg. 5: ref. to Lang et al., 1997. A more recent paper describing genome organization and gene content throughout jakobids is: http://www.ncbi.nlm.nih.gov/pubmed/23335123

R: The new reference has been added.

2. pg. 7: two citations (Harris, 1989; Mego, 1974) are not included in the list of references.

R: The reference list has been updated

3. pg. 21: "Balabaskaran Nina et al." should be "Nina et al." (the first author's name is "Praveen Balabaskaran Nina").

R: According to PubMed database and published papers, the first name is Praveen and Surname is Balabaskaran Nina. This is maybe a mistake in their own papers.
(1**: Balabaskaran Nina P**, Dudkina NV, Kane LA, van Eyk JE, Boekema EJ, Mather MW,Vaidya AB. Highly divergent mitochondrial ATP synthase complexes in Tetrahymenathermophila. PLoS Biol. 2010 Jul 13;8(7):e1000418. doi:10.1371/journal.pbio.1000418. PubMed PMID: 20644710; PubMed Central PMCID:PMC2903591.)
(2: **Balabaskaran Nina** P, Morrisey JM, Ganesan SM, Ke H, Pershing AM, Mather MW,Vaidya AB. ATP synthase complex of Plasmodium falciparum: dimeric assembly inmitochondrial membranes and resistance to genetic disruption. J Biol Chem. 2011Dec 2;286(48):41312-22. doi: 10.1074/jbc.M111.290973. Epub 2011 Oct 7. PubMedPMID: 21984828; PubMed Central PMCID: PMC3308843.)

Reviewer #2:

The manuscript by Perez et al. was aimed to characterize the subunit composition of the essential mitochondrial respiratory chain and ATP synthase complexes (CI to CV) in free living protist Euglena gracilis. The results indicated that many of the unusual subunits previously found in trypanosomes were also present in E. gracilis, ruling out the proposed association of such subunits with the parasitic lifestyle of trypanosomes. The results presented in this manuscript seem relevant for the field of molecular taxonomy and phylogeny.

1. *The long introductory paragraphs (p. 1-3) on phylogeny and mtDNA are not directly related to the main goal of this study and should be condensed.*

R: We agree with the reviewer that the description of mtDNAs was slightly off topic. This paragraph has thus been deleted. However, as underlined by the referee, our paper is relevant for the taxonomic and phylogenetic fields. Moreover, our study was carried out and interpreted with the reported evolutionary framework in mind, which is needed for a full understanding of our arguments. That is why the remaining introductory part about euglenozoans was left unchanged.

2. *"Enzyme activity analyses were performed on membrane fractions from cells grown…"*

The procedures used for isolation of the mitochondrial membranes were omitted.

R: There is a misunderstanding here since we used crude total membrane fractions and not mitochondrial membrane fractions. The procedure was described for Chlamydomonas cells in Remacle et al. 2001 (see section 3). The section 4 of the Methods was updated to make it clearer.

There is also no indication on the fraction marker assays used as purity criteria for mitochondria after the Percoll centrifugation step; the authors should realize that plastids and other subcellular fractions can co-purify with mitochondria.

R: See our positive response to a similar comment of reviewer 1 (point 6).

The pH of the culture media used is missing.

R: This is now mentioned in section 1 of the Methods

Were the enzyme activities assayed (p. 17) fully blocked by their specific inhibitors? This information was also omitted.

R: By definition, the specific activities were measured as the activity inhibited by maximal concentration of inhibitors used *in vivo*. This paragraph has been modified.

The number of cells (or cellular protein) used to get the protein extracts was not indicated, nor the amount of protein used for enzyme activity and electrophoresis.

R: The amount of cells used was formerly given in the last sentence of section 1 of the Methods. The amount of protein used is now indicated in the legends of Figures 4 and 5 about electrophoresis experiments, and in section 4 of the Methods.

3. Inhibition constants (Ki) for several respiratory chain inhibitors were calculated. However, [inhibitor]-activity plots were not shown and the inhibition mechanism was not specified for each inhibitor. Most likely, the Ki values shown in fact represent IC50 values. These IC50 values (p. 15-16) are too high and they contradict reported Ki values. These discrepancies should be experimentally analyzed or, at least, they should be described and discussed.

R: The reviewer is right. Reported values are IC50 and not Ki. This has been changed in the manuscript and we apologize for this stupid mistake. This being said, it is very difficult to make comments on IC50 versus Ki values since the reported values are obtained *in vivo*, meaning that inhibitory effects can be masked by alternative pathways of respiration. This is illustrated for example in the case of complex I, which can be bypassed by alternative type-II NADH dehydrogenases. This might explain for sure the apparent high IC50 values. Since this was not at all the aim of this paper to study the enzymology of the respiratory complexes in Euglena, we are not willing to further discuss these results. Raw data are now given in supplemental figure 3.

4. One important weakness of the present study is the lack of experimental assessment of the degree of purity of the protein samples collected from the two-dimension electrophoresis, which in turn seem to proceed from impure mitochondrial preparations. Some of the presumed new subunits might be contaminants. Complex I should be purified and reconstituted for determining activity and then, only after verifying that the complex is active, the identification of new subunits would be undisputable.

R: Concerning the purity of the preparations, I hope that our answer to comment 6 from reviewer 1 is sufficient. About the suggestion of the referee to use an alternative purification method, we would like to remind that the preparative procedure that was used here for complex I–V separation (BN-PAGE) is a broadly acknowledged method for multiproteic complex separation/purification. Accordingly, complex I was resolved as a single band of ~1500 kDa, and this band showed the expected NADH dehydrogenase activity. This being said, we agree that some proteins identified could be contaminants. Thus, we added a sentence in the Discussion about the possibility that some proteins might be contaminants.

Minor observations:

5. "Pellets were thawed and DNA extracted using a Qiagen DNAeasy extraction kit for total RNA". Is this sentence correct?

R: The spelling mistake was corrected (RNA to DNA)

6. Subunits of CII were identified by LC-MS but they were not apparently analyzed by MS/MS. Does this mean that LC-MS is a better method for analysis of this complex?

R: Unfortunately, we were not able to resolve complex II from BN-PAGE. This is mentioned in the Results and in the Discussion. Thus, one should not to draw any conclusion about the relative merits of the two approaches.

7. English editing by an English-speaking scientist may be beneficial to pick up and correct the numerous subtle grammatical mistakes.

R: The MS has been carefully checked for English mistakes. We apologize but this is not our native language. We hope that the MS is sufficiently clear now and hope the Mitochondrion journal editing service will correct the last mistakes.

Reviewer #3: The manuscript describes a combined bioinformatic and experimental approach to obtain the respiratory chain subunits of Euglena gracilis, and compares them with those of related species.

In principle the authors have done a thorough job. And I do think these kind of contributions are invaluable if we ever want to understand what all these extra subunits do in mitochondria (even though this manuscript in itself does not resolve that). I therefore very much support publication, with only minor revisions. I have tried to filter out some mistakes in the usage of the English language, but I would strongly advise that the manuscript is read & corrected by a native speaker (I guess Mark Field could do that, specifically as he is one of the authors).

R: See our answer to comment 7 from referee 2.

The 46 conventional subunits of complex I have recently gone down. NDUFA4 is a complex IV subunit.

R: This has been corrected. Supplemental Table 1 has also been modified to take into account the referee's remark.

It would be nice if one could see in Figure 2 on which analysis the results were based (or are they all based on both computational and experimental results?)

R: This is now explained in the legend of Figure 2. See also our answer to comment 2 from reviewer 1.

Based on the impression I get from Figure 2, I would argue that the statement that "most" of the subunits that appeared unique for the kinetoplastids have also been found in E.gracilis, is a bit of a stretch. More subtle, quantitative use of language that actually mentions the numbers or fractions would do more justice to the manuscript.

R: We agree. The corresponding sentence in the abstract has been modified.

It might be nice in the Discussion to comment upon the apparent contrast between human mitochondria that appear to have gained most of their supernumerary subunits shortly after the origin of the mitochondria, and the ones from euglenozoa that appear to have relatively many "unique" proteins. Even if these Euglenozoa specific subunits are not specific to kinetoplastids, it is interesting to note that they are specific to Euglenozoa, as human has much less taxon specific subunits.

R: We thank the reviewer for his interesting suggestion but we feel that we have already answered to this point in the last paragraph of section 1 in the discussion.

Editorial:
Abstract. First sentence: Mitochondria produce ATP and convert energy. They do not produce energy!

R: The sentence has been modified.

Introduction:
Archaea is plural

R: Corrected to archaeon.

Eukaryotic relationships -Eukaryotic phylogenetic relationships (?)

R: Changed to phylogenetic relationships among Eukaryotes, for clarity.

"Jakobids harbour mitochondrion" ?

R: Corrected to mitochondria.

"energetic metabolism"?

R: Edited.

"life state"

R: Replaced by “life stage”

"conservative database mining" (i guess something along the lines of "homology detection with conservative parameter settings" is meant.

R: The sentence was reformulated and now reads: (…) by combining targeted proteomics, complete proteome database mining and thorough phylogenetic analyses.

page 11:
"too partial sequences" -should be "to partial sequences" but even then it does not sound gramatically correct

R: Actually, the intended meaning was different and the sentence has been reformulated. It now reads: the sequences that were too incomplete to be accurately positioned in the trees.

Page 13:
"To confirm that QCRTB2 was indeed homologous to QCR1/QCR2" I guess the authors mean orthologous rather than homologous here? (they also mention homology in the abstract, which is puzzling as they talk about a phylogenetic pipeline that would not be required to establish homology).

R: Please see our lengthy answer to comment 2 of reviewer 1 for the exact meaning of homology, orthology and paralogy in the context of this manuscript. We have also edited the text in several parts to improve the clarity with respect to these concepts.

Page 15
"On a cell basis" do the authors mean that they quantified respiration per cell?

R: Edited as suggested.

Page 18
"but with a low E-value (0.053)." do the authors mean a high E valie?

R: Indeed, changed to high.

Page 20
"elaboration" ? I do not understand what that word means in this context.

R: Replaced by “recruitment of taxon-specific subunits”.

Page 21
"The dimeric nature of ATP synthase is now largely acknowledged in many organisms, and in mammals and yeasts, the dimer has a molecular mass of ~1.2 MDa and is thought also to be responsible for shaping mitochondrial cristae" I guess the comma after yeasts has to go

R: Edited: the comma was deleted.

I do not understand the term "mitochondrial paralogs" in Figure 1. How do we know that the identified paralogs are mitochondrial?

R: See our answers to comments 2 and 5 of reviewer 1.

Page 22
"Euglena QCR9 protein is not homolog to the canonical QCR9" do you mean homologous? In any case, in the absence of structure data that case is hard to make/

R: The legend of Figure 2 has been corrected. The case of QRC9 is discussed in the end of the Results.

***Highlights (for review)**

>Canonical and lineage-specific composition of respiratory complexes in Euglena

>Unusually large mitochondrial F_1F_o ATP synthase and Complex I in Euglena

>The subunit composition of respiratory complexes is conserved among Euglenozoa

The mitochondrial respiratory chain of the secondary green alga *Euglena gracilis* shares many additional subunits with parasitic Trypanosomatidae.

Emilie Perez[1,2], Marie Lapaille[1], Hervé Degand[4], Laura Cilibrasi[1], Alexa Villavicencio-Queijeiro[5], Pierre Morsomme[4], Diego González-Halphen[5], Mark C. Field[6], Claire Remacle[1,3], Denis Baurain[2,3, §], Pierre Cardol[1,3, §]

[1] Genetics and Physiology of microalgae, Department of Life Sciences, University of Liège, B-4000 Liège, Belgium

[2] Eukaryotic Phylogenomics, Department of Life Sciences, University of Liège, B-4000 Liège, Belgium

[3] PhytoSYSTEMS, University of Liège, B-4000 Liège, Belgium

[4] Institut des Sciences de la Vie, Université Catholique de Louvain, Louvain-la-Neuve, Belgium

[5] Departamento de Genética Molecular, Instituto de Fisiología Celular, Universidad Nacional Autónoma de México, Mexico

[6] Division of Biological Chemistry and Drug Discovery, University of Dundee, Dundee, Scotland, DD1 5EH.

[§]To whom correspondence should be addressed:

Pierre Cardol. Bvd du Rectorat, 27, B22, Institute of Botany, Dept of Life Sciences, University of Liège, 4000 Liège, Belgium. Pierre.cardol@ulg.ac.be. Tel 324-3663840

Denis Baurain. Bvd du Rectorat, 27, B22, Institute of Botany, Dept of Life Sciences, University of Liège, 4000 Liège, Belgium. denis.baurain@ulg.ac.be. Tel 324-3663864

Running Title: Euglena respiratory-chain complexes

Keywords: Euglenozoa, Trypanosomatidae, OXPHOS, Proteomics, Sequence database mining, Large-scale phylogenetics

Abstract

The mitochondrion is an essential organelle for the production of cellular ATP in most eukaryotic cells. It is extensively studied, including in parasitic organisms such as trypanosomes, as a potential therapeutic target. Recently, numerous additional subunits of the respiratory-chain complexes have been described in *Trypanosoma brucei* and *Trypanosoma cruzi*. Since these subunits had apparently no counterparts in other organisms, they were interpreted as potentially associated with the parasitic trypanosome lifestyle. Here we used two complementary approaches to characterise the subunit composition of respiratory complexes in *Euglena gracilis,* a non-parasitic secondary green alga related to trypanosomes. First, we developed a phylogenetic pipeline aimed at mining sequence databases for identifying homologs to known respiratory-complex subunits with high confidence. Second, we used MS/MS proteomics after two-dimensional separation of the respiratory complexes by Blue Native- and SDS-PAGE to both confirm *in silico* predictions and to identify further additional subunits. Altogether, we identified 41 subunits that are restricted to *E. gracilis, T. brucei* and *T. cruzi*, along with 48 classical subunits described in other eukaryotes (*i.e.* plants, mammals and fungi). This moreover demonstrates that at least half of the subunits recently reported in *T. brucei* and *T. cruzi* are actually not specific to Trypanosomatidae, but extend at least to other Euglenozoa, and that their origin and function are thus not specifically associated with the parasitic lifestyle. Furthermore, preliminary biochemical analyses suggest that some of these additional subunits underlie the peculiarities of the respiratory chain observed in Euglenozoa.

Introduction

The mitochondrion is a eukaryotic organelle acquired through endosymbiosis of an α-proteobacterium by either an ancestral eukaryote (Dyall et al., 2004; Gray et al., 1999) or an archaeon (Martin and Muller, 1998). Subsequently the vast majority of genes encoding mitochondrial proteins were transferred from the mitochondrial genome to the nuclear genome, and the residual mitochondrial genome mainly encodes products involved in the synthesis of mitochondrial proteins (tRNA and rRNA) and some subunits of the oxidative phosphorylation complexes (Gabaldon and Huynen, 2005; Saccone et al., 2006). The mitochondrion plays a central role in energy production as the site of the latter stages of respiration, *i.e.*, the Krebs cycle in the matrix and oxidative phosphorylation in the inner membrane. The respiratory chain consists of four classical multi-protein complexes: complex I (CI or NADH:ubiquinone oxidoreductase, EC 1.5.6.3), complex II (CII or succinate:ubiquinone oxidoreductase, EC 1.3.5.1), complex III (CIII or ubiquinol:cytochrome *c* oxidoreductase, EC 1.10.2.2) and complex IV (CIV or cytochrome *c* oxidase, EC 1.9.3.1). Except for CII,

these complexes couple electron transfer with translocation of protons from the matrix to the intermembrane space, generating a proton gradient used by a fifth complex (CV or F_1F_o-ATP synthase, EC 3.6.3.14) for ATP synthesis. Many organisms however have alternative pathways that transfer electrons without concomitant proton translocation, such as type-II NAD(P)H dehydrogenases (from NAD(P)H to ubiquinone) and alternative oxidases (AOX; from ubiquinol to molecular oxygen) (Michalecka et al., 2003; Van Aken et al., 2009).

Phylogenetic relationships among eukaryotes are notoriously difficult to resolve and thus highly disputed. In two recent syntheses of the phylogenetic literature, eukaryotes are considered as composed of about 6–9 major lineages, of which well-studied green plants, fungi and animals only correspond to a very small fraction (Adl et al., 2012; Walker et al., 2011). Among these major lineages excavates are a putative assemblage of (often heterotrophic) flagellates, proposed on the basis of shared morphological characters (*e.g.*, the ventral feeding groove and associated cytoskeletal structures) and molecular data. Excavates are subdivided into two subgroups, Metamonada and Discoba, the latter consistently recovered in phylogenomic analyses (Hampl et al., 2009; Zhao et al., 2012). Within Metamonada, most lineages lack typical mitochondria and instead possess hydrogenosomes (*e.g.*, *Trichomonas* in parabasalids) or mitosomes (*e.g.*, *Giardia* in diplomonads) (Adl et al., 2012; Walker et al., 2011). In contrast, the three lineages comprising Discoba are mostly mitochondriate: Heterolobosea have a regular mitochondrion with a very gene-rich genome (*e.g.*, *Naegleria gruberi*) (Gray et al., 2004), Jakobids harbour mitochondria retaining one of the highest gene complements known to date (*e.g., Reclinomonas* and *Andalucia*) (Burger et al., 2013; Lang et al., 1997), and, finally, Euglenozoa are a very diverse collection of protists, some of which are parasitic (*e.g.,* Trypanosomatidae in kinetoplastids), others photosynthetic (*e.g.*, euglenophytes in euglenids) or free-living heterotrophs (*e.g.,* diplonemids). Yet, all of them present very complex mitochondria with discoidal cristae (Simpson, 1997).

Amongst the studied trypanosomatidae are many pathogens, including *Trypanosoma brucei*, *T. cruzi* and *Leishmania major*, respectively causing sleeping sickness, Chagas' disease and leishmaniasis in humans. Extensively studied, these parasites are characterised by a complex life cycle and an energetic metabolism possessing many unique features. On the one hand, an alternative electron transfer chain, composed of a glyceraldehyde-3-phosphate dehydrogenase (G3PD) and an alternative oxidase (TAO or AOX), replaces the classical complexes I–IV during the bloodstream (or mammalian-infective) stage (Chaudhuri et al., 2006). On the other hand, respiratory complexes I–IV and the ATP synthase (complex V) found in the procyclic (or insect-infective) stage of trypanosomes possess many apparently-specific additional subunits (Acestor et al., 2011; Morales et al., 2009; Panigrahi et al., 2008; Zikova et al., 2008; Zikova et al., 2009).

E. gracilis is a secondary green alga, stemming from a secondary endosymbiosis between a prasinophyte green alga and a eukaryotic phagotroph belonging to the euglenids (Gibbs, 1981) (Turmel et al., 2009). The *E. gracilis* mitochondrion harbours all of the classical respiratory-chain complexes. However, besides CIII, , that is sensitive to antimycin A but insensitive to myxothiazol (Moreno-Sánchez et al., 2000), an antimycin A-resistant alternative pathway with CIII-like activity (including the proton translocation) carries a minor fraction of the electron flux in the presence of antimycin A (Sharpless and Butow, 1970b). *E. gracilis* also has an AOX that is active at pH 6.5 and at temperatures below 20°C (Castro-Guerrero et al., 2004), as well as a cytochrome *c* oxidase activity partially insensitive to cyanide in the presence of L-lactate (Moreno-Sánchez et al., 2000). Its CIV, meanwhile, has a very unusual subunit composition (Bronstrup and Hachtel, 1989).

In this work, we analysed the subunit composition of the classical respiratory chain complexes (CI-V) in *Euglena gracilis* by combining targeted proteomics, complete proteome database mining and thorough phylogenetic analyses. We show that beyond the subunits found in all mitochondriate eukaryotes studied so far, *E. gracilis* mitochondria also contain a minimum of 40 subunits previously identified only in parasitic Trypanosomatidae (kinetoplastids). Their presence in this photosynthetic alga belonging to a group of non-parasitic euglenids and, for some also retained in the colourless *E. longa* and/or in the free-living *Diplonema papillatum* (diplonemids), extends their distribution to at least Euglenozoa as a whole. These data also strongly suggest that these additional subunits are not associated with a parasitic lifestyle, but rather have a fundamental role in the Euglenozoa.

Materials and Methods

1. Strain and growth conditions

The *E. gracilis* strain used in this work (SAG 1224-5/25) was obtained from the University of Göttingen (Sammlung von Algenkulturen, Germany). Cells were grown in the dark (DK) or under low (50 µmol photons. m^{-2}. s^{-1}; LL) or medium (200 µmol photons . m^{-2} . s^{-1}; ML) light at 25°C, in liquid mineral Tris-minimum-phosphate (TMP) at pH 7.0 supplemented with a mix of vitamins (biotin 10^{-7}%, B12 vitamin 10^{-7}% and B1 vitamin $2x10^{-5}$% (w/v)) and eventually with an organic carbon source (acetate 17 mM or 60 mM, ethanol 20 mM or 200 mM) (Harris, 1989; Mego and Farb, 1974; Sharpless and Butow, 1970a). In all experiments, cells in exponential phase (1–$2x10^{6}$ cells/mL) were harvested by centrifugation for 10 min at 500–1,500 x g.

2. Oxygen consumption

Dark respiration rates were measured using a Clark Electrode (Hansatech Instruments, King's Lynn, England) as described in Duby and Matagne (1999). Rotenone (CI, 1–200 µM), antimycin A (CIII, 0.5–20 µM), SHAM (AOX, 0.2–2 mM), myxothiazol (CIII, 1–20 µM), cyanide (CIV, 0.02–2 mM) and oligomycin (CV, 0.5–40 µM) were added to cells grown in the presence of 60 mM acetate under low light.

3. Protein extracts

Crude total membrane fractions were prepared according to Remacle et al. (2001). The crude mitochondrial fraction was obtained as described by Moreno-Sánchez and Raya (1987). At the end of this protocol, the supernatant was centrifuged for 10 min at 10,000 g at 4°C and the pellet containing all organelles suspended in 250 µL of buffer (Remacle et al., 2001) and loaded onto a mannitol/Percoll gradient (Percoll 250 mM mannitol 13%, 21% and 45% (v/v)). The gradient and sample were centrifuged 50 min at 40,000 g at 4°C. Mitochondria were collected at the interface between the 21% and 45% layers and then washed twice by centrifugation for 10 min at 11,000 g at 4°C. The mitochondrial pellet was finally recovered in 300 µL of buffer. Protein concentrations were determined by the method of Bradford (1976).

To assess the purity of mitochondrial fractions, we first determined that chlorophyll content was below the detection threshold (< 10^{-3} µg chlorophyll per µg of protein). We also performed a preliminary analysis of the total protein content of mitochondrial fractions by LC-MS/MS analysis. Among the ~400 proteins that matched known proteins in GenBank (nr) (Benson et al., 2013) (E-value < 10^{-10}), > 80% corresponded to known mitochondrial components (data not shown).

4. Enzyme activities

Enzyme activity analyses were performed on crude total membrane fractions (50–150 µg of protein) from cells grown in the presence of 17 mM acetate under low light. CI (rotenone-sensitive NADH:duroquinone oxidoreductase), CII+III (succinate:cytochrome *c* oxidoreductase) and CIV (cyanide-sensitive cytochrome *c* oxidase) activities were measured as described in Cardol et al. (2002) and Remacle et al. (2001). CV (oligomycin-sensitive ATP synthase) activity was measured as described in Villavicencio-Queijeiro et al. (2009) based on the method from Pullman et al. (1960). The specific activities were reported as the activity inhibited by maximal concentration of inhibitors used *in vivo* (see Results).

5. Protein electrophoresis

To conduct Blue native polyacrylamide gel electrophoresis (BN-PAGE) analyses (Schagger and von Jagow, 1991), protein complexes were first solubilised in the presence of either N-dodecyl-β-D-maltoside (0.25–1.5%), digitonin (0.5–3.5%) (w/v), 750 mM epsilon-amino-n-caproic acid, 0.5 mM EDTA and 50 mM Bis–Tris pH 7.0, and centrifuged for 15 min at 15,000 g at 4°C to remove insoluble matter. 0.5% (w/v) sodium taurodeoxycholate was then added to the supernatant prior to separation by electrophoresis on a 4–10% or 4–12% (w/v) polyacrylamide gradient BN gel.

Complex I activity was detected by incubating the gel in 100 mM HEPES-KOH pH 8.0 or Tris-HCl pH 7.5, containing 0.2 mM NADH and 0.1% (w/v) Nitrotetrazolium Blue chloride (NBT). ATP synthase activity was detected by incubating the gel in 50 mM HEPES pH 8.0, containing 10 mM ATP and 30 mM $CaCl_2$.

BN-gels were also electroblotted according to standard protocols onto polyvinylidene fluoride membranes (Amersham GE Healthcare). Detection was performed using a BM Chemiluminescence Western blotting kit (Roche, Basel, Switzerland) with anti-rabbit peroxidase-conjugated antibodies. We used rabbit sera obtained against *Polytomella* sp. Pringsheim 198.80 β subunit (1:200,000).

Coomassie-Blue staining and the second dimensional Tricine-sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) procedure were performed as previously described (Cardol et al., 2004).

6. Mass spectrometry (MS) analyses

Stained proteins associated with spots or bands of interest were manually excised and analysed by mass spectrometry as described in Lapaille et al. (2010). MS and MS/MS queries were performed either against the GenBank protein database (nr) (Benson et al., 2013) or against a database of conceptually translated *E. gracilis* sequences assembled from contigs of 454 genomic reads and from contigs of public ESTs (see below). Precursor tolerance of 150 ppm for MS spectra and 0.1 Da fragment tolerance for MS/MS spectra were allowed. A charge state of +1 was selected. A single trypsin miscleavage and variable modifications consisting of methionine oxidation and acrylamide-modified cysteine were allowed. For protein direct identification with MASCOT, protein scores greater than 60 were considered as significant ($P < 0.05$).

7. 454 sequencing of E. gracilis genome

E. gracilis DNA sequence data were obtained from strain Z1 (kind gift of William Martin, University of Düsseldorf). Cells were maintained in Huttner's medium at ~20˚C in Erlenmeyer flasks, and with ambient illumination. Cells were harvested by centrifugation for ten minutes at 1,000 x g, washed twice with ice cold PBS and snap frozen on dry ice. Pellets were thawed and DNA extracted using a Qiagen DNAeasy extraction kit for total DNA (Qiagen UK, Manchester), and quality controlled using a NanoDrop spectrophotometer and by 1% agarose gel electrophoresis to determine that DNA was of high molecular weight. DNA was sequenced using a Titanium Roche 454 instrument and standard protocols. Approximately 1.7 million reads, corresponding to ~500 million bases that passed quality control were obtained and used in the subsequent analysis. Assembly of reads was performed using Newbler (http://www.454.com/products/analysis-software/).

8. In silico analyses

The initial dataset was composed of the complete protein databases of 21 organisms downloaded from various sources: *Escherichia coli* K-12 MG1655, *Mesorhibozium loti* MAFF303099, *Rickettsia prowazekii* Rp22, *Nostoc* sp. PCC 7120, *Prochlorococcus marinus* MIT9303, *Microcoleus chthonoplastes* sp. PCC 7420, *Komagataella pastoris* (NCBI RefSeq; ftp://ftp.ncbi.nih.gov), *Homo sapiens*, *Bos taurus*, *Saccharomyces cerevisiae* (Ensembl 68; ftp://ftp.ensembl.org), *Chlamydomonas reinhardtii*, *Volvox carteri*, *Micromonas* sp. RCC299, *Naegleria gruberi*, *Phaeodactylum tricornutum, Dictyostelium purpureum* (JGI Genome Portal; http://genome.jgi.doe.gov), *T. brucei*, *L. major* (TriTrypDB; http://tritrypdb.org), *Arabidopsis thaliana* (TAIR; ftp://ftp.arabidopsis.org), *Tetrahymena thermophila* (TGD; http://www.ciliate.org) and *Cyanidioschyzon merolae* (University of Tokyo; http://merolae.biol.s.u-tokyo.ac.jp).

To these complete proteomes, we added our database of translated *E. gracilis* sequences used in the MS analyses (see above). Briefly, the 23,372 ESTs of *E. gracilis* available at the NCBI (as of May 24, 2012) were contiged with CAP3 (Huang and Madan, 1999) and then translated in the reading frame yielding the longest protein fragment (using BioPerl *longorf.pl* script) (Stajich et al., 2002), whereas ~250-Mbp of 454 reads were assembled with the Velvet *de novo* assembler [version 1.2.03] (Zerbino and Birney, 2008) and then translated in the same way as EST contigs.

OrthoMCL [version 2.0.3] (Li et al., 2003) was used according to the protocol of Fischer et al. (2011) with modified settings: percent-match cutoff = 20, E-value-exponent cutoff = 1. Five values were tested for the inflation parameter (I = 1.1–1.5), but except for the tree of Figure 3A (see below), the analyses reported in this article are all based on the MCL-1.2 run. Orthologous groups and singletons were annotated by BLASTP [version 2.2.25+] using a list of handpicked reference protein sequences

(classical subunits of the complexes of the respiratory chain and those described in trypanosomes), which allowed us to identify the groups corresponding to the respiratory complex subunits. The sequences in each annotated orthologous group were aligned with Clustal Omega [version 1.1.0] (Sievers et al., 2011) and most alignments automatically cleared of the ambiguously aligned positions and of the sequences that were too incomplete to be accurately positioned in the trees (see below for details). The final alignments were then submitted to phylogenetic analysis with PhyML [version 3.0] (Guindon et al., 2010) using the LG+F+Γ_4 model (Le and Gascuel, 2008; Yang, 1993). The starting tree for the heuristic search was computed by parsimony and the search included both NNI (nearest-neighbour interchange) and SPR (subtree pruning and regrafting) topological moves. The resulting trees were used to select the monophyletic subgroup of interest whenever there were several paralogues included in the same OrthoMCL orthologous group (*i.e.*, plastidial or cytosolic in addition to mitochondrial sequences). Each selected group or subgroup was then used to build a HMM profile with HMMER [version 3.0] (http://hmmer.janelia.org). Using a customized version of HaMStR [version 9] (Ebersberger et al., 2009), these HMM profiles allowed us to search for additional orthologous sequences in our untranslated *E. gracilis* database and in the (largely incomplete) EST databases of *E. longa* and *D. papillatum* available at the NCBI.

In order to analyse the evolution of QCR1, QCR2 and QCRTB2, we considered an additional batch of 47 broadly sampled organisms, for which we downloaded the complete protein databases: *Azospirillum brasilense*, *Batrachochytrium dendrobatidis*, *Rhizopus delemar*, *Acanthamoeba castelanii*, *Polysphondylium pallidum*, *Leishmania infantum*, *Leishmania donovani*, *Leishmania mexicana*, *Leishmania braziliensis, T. cruzi*, *Angomonas deanei*, *Strigomonas culicis*, *Trichomonas vaginalis*, *Galdieria sulphuraria*, *Ectocarpus siliculosus*, *Nannochloropsis gaditana*, *Cryptosporidium muris*, *Perkinsus marinus*, *Oxytricha trifallax* (NCBI RefSeq); *Monosiga brevicollis*, *Lottia gigantea*, *Guillardia theta*, *Emiliania huxleyi*, *Aurantiochytrium limacinum*, *Aureococcus anophagefferens*, *Bigelowiella natans*, *Chlorella variabilis*, *Ostreococcus tauri* (JGI Genome Portal); *Puccinia graminis*, *Ustilago maydis*, *Amphimedon queenslandica*, *Nematostella vectensis, Dictyostelium discoideum*, *Entamoeba histolytica*, *Giardia lamblia*, *Hyaloperonospora arabidopsidis*, *Pythium ultimum*, *Phytophthora infestans*, *Thalassiosira pseudonana*, *Plasmodium falciparum*, *Toxoplasma gondii* (Ensembl Genomes 17; http://ensemblgenomes.org), *Selaginella moellendorffii*, *Physcomitrella patens*, *Oryza sativa* (Phytozome 9.0; http://www.phytozome.net); *Chondrus crispus* (ENA; http://www.ebi.ac.uk/ena); *Pyropia yezoensis* (NRIFS; http://nrifs.fra.affrc.go.jp/ResearchCenter/5_AG/genomes/nori); *Cyanophora paradoxa* (Rutgers University; http://cyanophora.rutgers.edu/cyanophora). Finally, we also used the conceptual translation of the transcriptome of *Porphyridium purpureum* downloaded from the NCBI.

The HMM profile of the subgroup QCR1/QCR2/QCRTB2, derived from an orthologous group obtained in the MCL-1.1 run, was used to search for homologous sequences in a combined database of our 68 organisms with HMMER. To select QCR-like sequences, we drew a negative logarithmic plot of HMMER E-values using R (http://www.R-project.org) and set the E-value cutoff to 10^{-107}, *i.e.*, just before the first undisputable drop in $-\log_{10}$(E-value). The 178 retained sequences were aligned with MUSCLE [version 3.8.31] (Edgar, 2004), which resulted in a better alignment than Clustal Omega on this particular protein family. We then enriched this alignment with the four partial orthologous sequences from *E. gracilis* and *E. longa* identified by HaMStR.

Before phylogenetic analysis, the latter alignment and the MUSCLE alignment of the MCL-1.1 orthologous group were both filtered to eliminate poorly aligned positions and mostly incomplete sequences using the Bio-MUST-Core software package (D.B., unpublished). Briefly, positions due to insertions in less than 50% of the sequences were discarded. Gblocks 0.91b (Castresana, 2000) was then used with loose parameters to further filter the least reliably aligned positions. Finally, the sequences having more than 90% missing characters with respect to the longest sequence were discarded. The two prepre-processed alignments were submitted to phylogenetic inference using PhyML as above, and statistical support was estimated through the analysis of 100 bootstrap pseudo-replicates (Felsenstein, 1985). Tree rooting was done manually in Seaview [version 4.4.2] (Gouy et al., 2010), while ladderization and taxonomic colouring were automatically conducted using Bio-MUST-Core. Final trees were displayed and arranged in FigTree [version 1.4.0] (http://tree.bio.ed.ac.uk/software/figtree). It is worth mentioning that a similar procedure was used to ensure that apparently Euglenozoa-specific subunits were indeed restricted to the latter group.

All *in silico* analyses were carried out on a desktop workstation running Bio-Linux 6 (http://nebc.nerc.ac.uk/tools/bio-linux) (Field et al., 2006).

Results

1. In silico analysis of the subunit composition of respiratory-chain complexes in Euglenozoa

Additional, putatively taxon-specific, subunits of respiratory-chain complexes often share two attributes that make them difficult to be reliably identified in the genomes of non-model organisms. First, they are poorly conserved [*e.g.*, NDUFC2/B14.5b, NDUFA7/B14.5a, NDUFA3/B9 (Cardol, 2011; Huynen et al., 2009)], which rules out simple BLAST searches and calls for iterative (*e.g.*, PSI-BLAST) or HMM-based (*e.g.*, HMMER) approaches. Second, they are regularly recruited from large multi-gene families with many paralogous groups, the evolution of which cannot be untangled without broadly sampled phylogenetic trees (Gabaldon and Koonin, 2013; Koski and Golding, 2001). For

example, in CI, NDUFA11 is a TIM17/22-like protein, CAG9 is a gamma carbonic anhydrase, NDUFA2 features a mitochondrial ribosomal protein L51/S25/CI-B8 domain, NDUFAB1 is an acyl-carrier protein, and NDUFB9 belongs to the LYR protein family while in CIII, QCR1/2 both belong to the peptidase family M16. Consequently, when searching for the orthologues of known additional subunits in a target genome, it may be hard to ensure that a given gene, identified with a weak PSI-BLAST hit, actually corresponds to the reference subunit used as the query for the search, and not to a paralogous gene of the same family, potentially fulfilling a different function in a different cell compartment. In this work, we used a carefully designed phylogenetic pipeline to ensure maximal specificity when identifying orthologous mitochondrial proteins in five Euglenozoa (*E. gracilis*, *E. longa*, *T. brucei*, *L. major* and *D. papillatum*) and in a heterolobosean outgroup, *Naegleria gruberi* (Figure 1). This higher reliability is at the expense of lower sensitivity, which means that we probably failed to identify genuine orthologous genes, some of them previously reported, *e.g.*, in *N. gruberi* (Cardol, 2011). However, we argue our approach is sound in the context of a study aimed at expanding the taxonomic distribution of reportedly taxon-specific subunits.

To this end, we compiled an inventory of 99 subunits constituting the traditional complexes of the respiratory chain in model eukaryotes, using *Saccharomyces cerevisiae*, *Homo sapiens* (both Opisthokonta) and *Arabidopsis thaliana* (Viridiplantae = green plants) as reference organisms. In addition, we also included the 84 apparently specific subunits described in *T. brucei* and/or *T. cruzi* (Acestor et al., 2011; Morales et al., 2009; Panigrahi et al., 2008; Zikova et al., 2008; Zikova et al., 2009) (Supplemental Table 1). In the following, TC and TB notations refer to subunits described in *T. cruzi* and *T. brucei*, respectively. We then collected the complete proteomes of six bacteria (*Escherichia coli* K-12 MG1655, *Mesorhibozium loti* MAFF303099, *Rickettsia prowazekii* Rp22, *Nostoc sp.* PCC 7120, *Prochlorococcus marinus* MIT9303, *Microcoleus chthonoplastes sp.* PCC 7420) and 15 broadly sampled eukaryotes (*Homo sapiens*, *Bos taurus*, *Saccharomyces cerevisiae*, *Komagataella pastoris*, *Dictyostelium purpureum, Cyanidioschyzon merolae, Chlamydomonas reinhardtii*, *Volvox carteri*, *Micromonas sp.*, *Arabidopsis thaliana*, *Phaeodactylum tricornutum*, *Tetrahymena thermophile, N. gruberi*, *T. brucei* and *L. major*). To this database, we added private 454 genomic contigs and public EST contigs of *E. gracilis*, all translated in the frame that yielded the longest predicted protein fragment (see Materials and Methods for details).

In order to confidently cluster orthologous respiratory complex subunits in our complete proteomes, we first carried out an all-*vs*-all comparison of our database sequences using BLASTP. Application of the OrthoMCL pipeline to the resulting BLAST reports allowed us to generate between 35,800 and 49,522 orthologous groups, by varying the inflation parameter (I) from 1.1 to 1.5. These groups only contained orthologous and in-paralogous genes (due to terminal duplications), with out-paralogous

genes (due to ancestral duplications) automatically sorted out into different groups. This unbiased global approach spared us the need to choose 99 specific E-value thresholds for separating orthologues from paralogues in each individual case, as done in more conventional database mining studies. For each of these five MCL runs, we annotated the orthologous groups by comparison with our inventory of reference proteins (again using BLASTP). This revealed that the value of "I" leading to the best delineation of the different subunits was 1.2. Among the 605 annotated orthologous groups of the MCL-1.2 run, we carefully selected one group for each subunit, generally the one including both the reference protein(s) and the largest sample of species. These 99 groups were then aligned, cleared of ambiguously aligned positions and of incomplete sequences (because short sequences hinder analyses), and finally used to compute phylogenetic trees (Supplemental Figures 1 and 2; see also Figure 3 below). In each of these trees, we identified and analysed the subtree corresponding to the mitochondrial paralogue (by opposition to the plastid paralogue) by its proximity to the two α-proteobacteria and/or by its inclusion of the reference protein(s). However, as the sequences in our *E. gracilis* database had only been translated in a single frame and were sometimes very short, each mitochondrial subgroup was re-analysed with HaMStR (based on HMM profiles and applying the best-reciprocal hit (BRH) orthology criterion) to find additional orthologous sequences in our *E. gracilis* nucleotide database (using six-frame translation) and in the incomplete public EST databases of the secondarily non-photosynthetic *E. longa* and of the free-living diplonemid *D. papillatum*. Altogether, this allowed us to identify a total of 84 mitochondrial subunits in euglenids (*E. gracilis* and *E. longa*), 8 in *D. papillatum* and 118 in kinetoplastids (*T. brucei* and *L. major*). Note that all the subunits specifically described in trypanosomes were also found in *L. major*. This is why, in the following, we refer to them as "apparently kinetoplastid-specific subunits". Furthermore, 60 subunits were found for the heterolobosean *N. gruberi*, which we selected as outgroup (Figure 2). Accession numbers and identified sequences in excavate representatives are given in supplemental files 1-7.

1.1 Complex I

We found 19 of the 45 conventional subunits of the eukaryotic CI in *E. gracilis*. The majority of these subunits were either core subunits (ND1/4/5, NDUFS1-3/7/8 and NDUFV1/2) that trace back to the α-proteobacterial ancestor of the mitochondrion or among the additional subunits whose primary structure is the best conserved of CI (NDUFA7/9/12/13, NDUFAB1, NDUFB7/11, NDUFC1, and CAG9). The CAG9 group corresponds to gamma carbonic anhydrases, two of which are present in Euglena (CAG1/2). Most of these 19 proteins were also identified in *N. gruberi*. Out of the 34 additional CI subunits described in *T. brucei* as potentially kinetoplastid-specific (Acestor et al., 2011; Panigrahi et

al., 2008), we found 14 subunits in *E. gracilis* (NDTB1, 2, 5, 6, 11, 12, 17, 18, 22, 25, 28, 29, 31 and 34) (Figure 2). Three of these subunits were also identified *in D. papillatum* (NDTB26/29/32) and 10 in *N. gruberi* (NDTB2/4/8/11/12/17-20/22). Finally, we concluded that NDTB25 is actually homologous to a conventional subunit (NDUFB7), that NDTB5 is paralogous to a conventional subunit (NDUFA9), and that NDTB2 and NDTB3 are paralogous to each other.

1.2 Complex II

CII is typically described as a four-subunit complex in mammal/fungal species. In contrast, *T. cruzi* presents a set of 11 subunits, of which only three are identical to the conventional subunits SDH1 and the two fragments of SDH2 (SDH2N and SDH2C) (Gawryluk and Gray, 2009; Morales et al., 2009). Four of the *T. cruzi*-specific subunits (SDHTC3/6/8/9) and two other additional subunits (SDHTB1/2) have been identified in *T. brucei* (Acestor et al., 2011). Similarly, we found the canonical SDH1 and SDH2 subunits in *E. gracilis* (the latter also split in two parts, SDH2N and SDH2C), as well as eight apparently kinetoplastid-specific subunits (SDHTC3/4/6-10 and SDHTB1) (Figure 2). In addition, SDHTC8 and SDHTC11 were identified in *D. papillatum*, while only SDH1 and SDH2 were found in *N. gruberi*.

1.3 Complex III

Among the 10 classical CIII subunits, eight were found in *E. gracilis* (QCR1, QCR2, RIP1, COB, CYT1, QCR6, QCR7 and QCR10) and six in *N. gruberi* (QCR1, QCR2, RIP1, COB, CYT1 and QCR7) (Figure 2). In *T. brucei*, three additional apparently specific subunits (QCRTB1, QCRTB2 and QCRTB3) have been described (Acestor et al., 2011), two of which were found in *E. gracilis* (QCRTB1/2). Interestingly, we noticed that in the MCL-1.1 run, QCRTB2 fell in an orthologous group of M16 peptidases also containing QCR1 and QCR2. Thus, to confirm that QCRTB2 was indeed homologous to QCR1/QCR2, we aligned all the sequences of the MCL group and used the resulting alignment for inferring a phylogenetic tree (Figure 3A). This tree allowed us to separate the sequences belonging to the subgroup QCR1/QCR2/QCRTB2 from the rest of the M16 peptidase family. To determine the evolutionary relationships between QCR1, QCR2 and QCRTB2, we built a HMM profile of this subgroup and used it to search a larger database enriched of 47 broadly sampled eukaryotic proteomes (in addition to the 21 used above; see Materials and Methods for a complete list of organisms). QCR-like sequences matching the HMM profile were then aligned, supplemented with four short sequences of *E. gracilis* and *E. longa* identified by HaMStR, and finally submitted to

phylogenetic analysis. From the obtained tree (Figure 3B), we concluded that QCRTB2 was likely the result of a duplication of the QCR1 gene having taken place in the common ancestor of Euglenozoa.

1.4 Complex IV

Among the 13 mammalian CIV subunits (Kadenbach et al., 1983), we identified seven subunits in *E. gracilis* (COX1-3/5A/5B/6B/8A), and only three in *N. gruberi*, all encoded in the mitochondrial genome (Figure 2). Among the 15 apparently kinetoplastid-specific subunits described in *T. brucei* (Zikova et al., 2008), nine were found in *E. gracilis* (COXTB1/2/4-6/8/10/12/16), one in the partial database of *D. papillatum* (COXTB6) and one in the complete proteome of *N. gruberi* (COXTB2).

1.5 ATP synthase

Eukaryotic F_1F_o-ATP synthases of mammal, fungal and land plant model organisms (Collinson et al., 1994; Heazlewood et al., 2003; Velours and Arselin, 2000) usually consist of ~20 subunits, seven of which were found in *E. gracilis* (α, β, χ, δ, ε, OSCP, c), one in *D. papillatum* (subunit c) and seven in *N. gruberi* (α, β, χ, δ, OSCP, a, c). Among the 15 apparently specific subunits described in *T. brucei* (Zikova et al., 2009), we identified eight subunits in *E. gracilis* (ATPTB1/3/4/6/7/10/12 and p18) and one in *D. papillatum* (ATPTB2) (Figure 2).

1.6 Alternative pathways

Alternative pathways described in trypanosomes are similar to the alternative pathways described in other protists, green plants and fungi: an alternative oxidase (AOX or TAO), a type-II NADH dehydrogenase (NDA or TAD) (Acestor et al., 2011; Fang and Beattie, 2002) and a glyceraldehyde-3-phosphate dehydrogenase (G3PD), the activity of which compensates for the absence of oxidative phosphorylation during the bloodstream stage (Chaudhuri et al., 2006). Our analyses demonstrated that trypanosomal alternative oxidase and type-II NADH dehydrogenase are homologous to those of the other organisms. While we recovered an AOX and a G3PD in *E. gracilis*, we did not find a NDA (Figure 2). All three of these enzymes were found in *N. gruberi*.

2. Proteomic analysis of the subunit composition of the respiratory-chain complexes in E. gracilis

We sought to robustly validate our *in silico* predictions by determining whether the proteins identified by data mining were *bona-fide* components of the respiratory-chain complexes in *E.*

gracilis. To identify growth conditions where *E. gracilis* accumulated more mitochondria and respiratory complexes, we compared various combinations of light intensity (darkness, low light or medium light) and availability of an exogenous carbon source [absence or presence of acetate (17mM or 60mM) or ethanol (20mM or 200mM)]. For each culture, dark *in vivo* respiratory rates were then estimated (Supplemental Table 2). Oxygen consumption per cell was greater in the presence of high concentrations of acetate or ethanol, independent of light availability. The maximal value (~1.0–1.5 µmoles O_2. h^{-1}. 10^{-6} cells) was in good accordance with a previous report (Buetow, 1961). Lowest respirations rates (~0.1 µmoles O_2. h^{-1}. 10^{-6} cells) were observed in the absence of exogenous carbon sources.

We next tested the effect of increasing concentrations of classical potent respiratory-complex inhibitors on the *in vivo* respiratory rate. Maximal inhibition (I_M; percentage of total respiration) and half maximal inhibitory concentration (IC_{50}, µM) were determined (Supplemental Figure 3). Rotenone (CI, I_M = 61 ± 8 %; IC_{50}= ~20 µM), antimycin A (CIII, I_M = 71 ± 7 %; IC_{50} = <1 µM), cyanide (CIV, I_M = 89 ± 4 %; IC_{50} = ~20 µM) and oligomycin (CV, I_M = 62 ± 5 %; IC_{50}= ~8 µM) inhibited respiration, suggesting that the corresponding complexes might participate to respiration. In contrast, SHAM and myxothiazol, which classically inhibit AOX and alternative CIII, respectively, did not significantly inhibit respiration. Specific enzyme activities were also measured on crude total membrane fractions using the highest inhibitor concentrations tested *in vivo*. These analyses confirmed that rotenone-sensitive NADH:duroquinone oxidoreductase (CI, 24 ± 6 nmol NADH.min^{-1}.mg^{-1} prot), antimycin A-sensitive succinate:cytochrome *c* oxidoreductase (CII+III, 9 ± 5 nmol cyt *c*.min^{-1}.mg^{-1} prot), cyanide-sensitive cytochrome *c* oxidase (CIV, 58 ± 12 nmol cyt *c*.min^{-1}.mg^{-1} prot) and oligomycin-sensitive ATP synthase (CV, 90 ± 40 nmol ATP.min^{-1}.mg^{-1} prot) were all present in *E gracilis*.

To analyze the composition of respiratory-chain complexes, we solubilized membrane complexes in their native form and separated them by BN-PAGE. The effect of increasing concentrations of two mild non-ionic detergents (n-dodecyl maltoside and digitonin) on protein solubilisation was evaluated. For cells grown in the dark in the presence of high concentrations of acetate, six major bands ranging from ~200 kDa to ~2 MDa could be visualized (Figure 4A). Bands 2 (~1.5 MDa) and 4 (~500 kDa) sometimes appeared as doublets. While band 3 (~900 kDa) was recovered in higher abundance with digitonin, bands 4, 5 (~400 kDa) and 6 (~200 kDa) were more abundant with n-dodecyl maltoside. We then compared the distribution of these bands in mitochondrial extracts of cells cultivated in the growth conditions yielding the highest respiratory rates (*i.e.*, high concentrations of acetate or ethanol in the dark or in the light). However, light + ethanol conditions were avoided since they led to the well-known phenomenon of cell bleaching due to catabolite

repression of chloroplast development in *E. gracilis* (Monroy and Schwartzbach, 1984). Bands 1, 2, 4, and 5 were present in every condition, while all six bands were more abundant in extracts of dark-grown cells (Figure 4B). Further, in-gel staining aimed at highlighting ATPase and NADH dehydrogenase activities allowed us to hypothesize that band 1 comprised ATP synthase (CV) and band 2, CI (Figure 4C,D). Immunodetection on western blot with antibodies against the β subunit confirmed the presence of this CV subunit in band 1 (Figure 4E).

The subunits of each complex separated by BN-PAGE were then tentatively resolved by SDS-PAGE and stained with Coomassie blue. The starting material was a purified mitochondrial fraction from cells grown in the dark in the presence of 200 mM ethanol. As shown in Figure 5A, at least 18, 22, 8, and 10 protein bands could be visualized for bands 1, 2, 4 and 5, respectively. The same pattern of spots was also obtained for a crude membrane extract of cells grown under low light in the presence of 60 mM acetate (data not shown). The most prominent spots were excised out of the gel and analysed by tandem mass spectrometry (MS/MS). The anonymous fragmented proteins of *E. gracilis* obtained in MS analyses were annotated using BLASTP queries against GenBank (nr) and our own protein predictions (Figure 1). Forty-seven of the 58 analysed protein spots matched a protein in our database and 30 corresponded to an annotated protein in GenBank (Figure 5A, Supplemental Table 3). In band 1, six of the classical eukaryotic ATP synthase subunits were found (α, β, γ, δ, ε and OCSP), in addition to four of the subunits described in *T. brucei* (ATPTB1/4/12 and p18). In band 2, we found 5 canonical CI subunits (NDUFS2/3, NDUFV1, NDUFA12/13, and 2 gamma carbonic anhydrases CAG1/2) and 3 subunits described in *T. brucei* CI (NDTB2/12/17). Two additional proteins, glyceraldehyde-3-phosphate dehydrogenase (G3PD) and chaperone protein HSP40 (DNAJ), were also identified. Band 4 corresponded to CIII with 4 identified subunits (QCR1/7, RIP1, QCRTB1) and band 5 to CIV with 4 known subunits (COX3/6B, COXTB4/5). Band 6 could unfortunately not be resolved into well-defined spots. A partial subunit pattern (7 bands) of band 3 was obtained from a different experiment (Figure 5B). Interestingly, this corresponded to a mixture of CIII (500 kDa) and CIV (400 kDa), which are probably associated *in vivo* in a supercomplex (900 kDa).

In a second approach, bands 2–6 from BN-PAGE were directly excised and each band was analysed by liquid chromatography-mass spectrometry (LC-MS) to identify possible additional subunits (see Supplemental File 8). Additional subunits could be identified for CI (band 2; NDUFS1/7/8, NDUFA8/9 and NDTB5/29), CIII (bands 3 and 4; CYT1, COB, QCR9), CIV (bands 3 and 5; COX1/2, COXTB2) and CII (band 6; SDH1 and SDH2N/C, SDHTC3 and SDHTC9). Surprisingly, the NDUFA8 subunit was found by proteomics but not by *in silico* analysis. This is explained by the observation that the orthologous group that annotated the protein fragment was composed only of euglenozoan sequences, for which the first BLASTP hit was the NDUFA8 subunit of *Ciona intestinalis*, but with a high (thus not

significant) E-value (0.053). In accordance with this observation, a multiple alignment built from reference NDUFA8 sequences (from *A. thaliana*, *H. sapiens* and *C. intestinalis*) showed that the NDUFA8 subunits of Euglenozoa were poorly conserved, while retaining the majority of the characteristic cysteine residues (Figure 6). As for the NDUFA8 subunit, the QCR9 subunit was only found by proteomics. This protein was previously identified by N-terminal sequencing as a peculiar component of CIII in *E. gracilis* (Cui et al., 1994). We could not identify any similar sequences in other eukaryotes. Thus, we propose that the Euglena QCR9 does not exhibit any conserved domain and does not correspond to the canonical QCR9 subunit described in other species.

A summary of the proteins identified in this study is given in Figure 2.

IV. Discussion

1. Similar subunit composition of respiratory-chain complexes in kinetoplastids and euglenids

Respiratory-chain complexes have a dual genetic origin, most subunits being encoded in the nucleus, while a few subunits of prokaryotic origin remain encoded by the mitochondrial genome (Gabaldon and Huynen, 2005; Saccone et al., 2006). Compared with their prokaryotic counterparts, mitochondrial respiratory-chain complexes have acquired additional subunits, although most of these are not directly involved in the catalytic activity of the enzyme complexes. During the last decade, many studies have highlighted the peculiarities of respiratory-chain complexes in trypanosomes (Panigrahi, Zikova et al. 2008, Zikova, Panigrahi et al. 2008, Morales, Mogi et al. 2009, Zikova, Schnaufer et al. 2009, Acestor, Zikova et al. 2011), in which the majority of the identified subunits are apparently unrelated to those described in green plants, mammals or fungi. Trypanosomes are part of Euglenozoa, a eukaryotic phylum first described three decades ago and grouping kinetoplastids and euglenids (Cavalier-Smith, 1981). At the same time, a secondary symbiotic origin of euglenids chloroplast was proposed (Gibbs, 1981). More recently, Euglenozoa were extended to also include diplonemids (Simpson, 1997). The mitochondrial respiratory chain of *E. gracilis* is well studied (Bronstrup and Hachtel, 1989; Buetow, 1961; Castro-Guerrero et al., 2005; Castro-Guerrero et al., 2004; Moreno-Sánchez et al., 2000; Sharpless and Butow, 1970a, b; Wilson and Danforth, 1958) but the protein composition of its complexes has only been investigated in a couple of studies (Bronstrup and Hachtel, 1989; Cui et al., 1994). In the present work, we addressed this gap by combining a conservative phylogenetic pipeline for *in silico* data mining of complete proteomes with proteomic analyses performed under multiple conditions. It is worth noting that some low molecular mass subunits may probably not be visualised by the Coomassie-blue staining, and that some hydrophobic subunits may have been lost during the second electrophoresis step.

Moreover, such proteins contain only few tryptic cleavage sites, thus generating a limited number of peptides in the mass range suitable for MS analysis. In particular, none of the seven hydrophobic core subunits, which are usually mitochondrially encoded (ND1-6, ND4L), could be detected, and which suggests that despite the multiple and exhaustive analysis presented here that the complete composition of the Euglena respiratory apparatus remains to be achieved. Nonetheless, at this stage, the respiratory-chain complexes I–V of *E. gracilis* consist of at least 92 different proteins, 79 of which are shared with trypanosomes and 34 of which, according to our analyses, appear to be restricted to Euglenozoa.

From these data, we can draw three main conclusions: (i) the peculiar subunit composition of the respiratory-chain complexes described in trypanosomes originated in the common ancestor that they share with euglenids and, for a few subunits, in the common ancestor of Discoba, since some were also identified in the heterolobosean *N. gruberi*; (ii) these data reinforce the idea that the additional subunits found in trypanosomes are genuine components of respiratory complexes of these parasites, and; (iii) contrarily to recent suggestions (Panigrahi et al., 2008), the additional subunits shared with euglenids are unlikely to play any specific role in the parasitic lifestyle of trypanosomes. Thus, beside canonical subunits shared by all eukaryotic groups, respiratory complexes I–V in Euglenozoa also possess specific subunits that either have diverged beyond recognition or have been recruited early in the evolution of the lineage. Two scenarios may account for the recruitment of new subunits: (i) acquisition of xenologous genes by lateral gene transfer, either of bacterial or eukaryotic origin, or (ii) retargeting of cytosolic proteins to the mitochondria, possibly after duplication of the corresponding genes. In both cases, these novel subunits would have assembled to form a new scaffold around the canonical core of respiratory complexes. From our *in silico* analyses, it is not possible to differentiate between these two models, as we generally failed to identify homologues (neither paralogues or orthologues) for Euglenozoa–specific genes, even after thorough HMM searches against the 68-species database built for studying the evolution of QCR1/QCR2/QCRTB2, as well as against another database of about 400 complete proteomes representative of prokaryotic diversity (data not shown). The exceptions here may be NDTB2/3/4/8/10/11/17/18/22/29, SDHTB2, QCRTB1/2 and COXTB6/16, all of which possess conserved domains (but not full-length architectures) found in other eukaryotic groups. Thus, whatever their ultimate origin, Euglenozoa-specific genes are likely to have undergone extensive evolutionary divergence.

2. Peculiar features of respiratory-chain complexes in E. gracilis

Our study also highlights some interesting features at the level of individual respiratory-chain complexes in *E. gracilis*.

2.1. Complex I

The approximate molecular mass of CI in *E. gracilis* is ~1.5 MDa. In contrast, in all eukaryotes investigated so far, including *T. brucei* (Acestor, Zikova et al. 2011), the mammal *Bos taurus* (Carroll et al., 2006), the yeast *Yarrowia lipolytica* (Angerer et al., 2011), the green alga *Chlamydomonas reinhardtii* (Cardol et al., 2004) and the amoeba *Acanthamoeba castellanii* (Gawryluk et al., 2012), CI has a molecular mass of ~900-1000 kDa. In green plants, mammals and fungi, CI has however been found in association with dimeric CIII and/or CIV, leading to supercomplexes of >1.5 MDa (Cardol et al., 2008; Dudkina et al., 2005; Schagger and Pfeiffer, 2000). However, in the present study, no evidence for the presence of CIII or CIV subunits could be obtained in the band corresponding to CI in *E. gracilis*. Compared with CI in fungi and mammals, CI in green plants (Cardol et al., 2004; Sunderhaus et al., 2006), amoeba, and presumably other eukaryotes (Gawryluk and Gray, 2010) comprises additional subunits belonging to the γ–carbonic anhydrase (CAG) protein family. Two CAG proteins have also been specifically identified in CI of *E. gracilis*. In green plants, it has been shown that these proteins form a matrix-exposed domain (Sunderhaus et al., 2006), the presence of which does not substantially increase the molecular mass of CI. G3PD and DNAJ enzymes have also been found in association with CI in *E. gracilis*. Interestingly, a DNAJ homologue was recently identified in CI of *T. brucei* (Acestor et al., 2011). DNAJ/HSP40 (heat shock protein 40) primarily stimulates ATPase activity of HSP70 chaperones (Qiu et al., 2006) and these proteins are broadly distributed among eukaryotic lineages. Although we cannot completely rule out the possibility that DNAJ/HSP40 and G3PD are non-mitochondrial contaminations of the mitochondrial fraction (see section 3 of Materials and Methods), our study brings evidence of a potential interaction between this cohort of proteins and respiratory CI. We recently proposed that some proteins might use CI as an anchoring point and that these interactions might be rather specific to limited groups or species (Cardol, 2011): acyl carrier protein (NDUFAB1/ACPM) and deoxyribonucleoside kinase-like subunit (NDUFA10/42 kDa) in mammals, galactono-lactone dehydrogenase in land plants (Klodmann et al., 2010), or rhodanese in *Y. lipolytica* (Angerer et al., 2011). Similarly, G3PD and DNAJ could be specifically associated with CI in *E. gracilis*. Thus, the apparent higher molecular mass of *E. gracilis* CI probably does not result from an abnormal molecular mass of the canonical subunits or from association with other respiratory complexes, but might rather be the consequence of interactions with proteins bearing functions that are not directly related to CI activity.

2.2. Complex II

CII (succinate dehydrogenase) was found to be more abundant in heterotrophic conditions compared to mixotrophic conditions. This is in good agreement with a previous report (Brown and Preston, 1975). CII usually comprises four subunits (SDH1-4): a flavoprotein (SDH1), an iron-sulfur (Fe-S) protein (SDH2) and two membrane subunits (SDH3/SDH4) providing ligands to heme *b* and a reduction site for ubiquinone. As previously pointed out (Morales et al., 2009), membrane anchor subunits are highly divergent between bacteria, mammals and other eukaryotes, and thus are difficult to identify with standard BLAST searches (Acestor et al., 2011). Candidates for *T. cruzi* and *T. brucei* SDH3 and SDH4 were proposed based on the presence of quinone/heme-binding motifs, predicted *trans*-membrane domains and similar protein sizes, but a direct comparison with these proteins did not reveal any obvious predicted structural or motif similarity (Acestor et al., 2011; Morales et al., 2009). This further illustrates the difficulty to assess orthology for small hydrophobic membrane proteins [see also Discussion in (Cardol, 2011)]. Additional subunits were also found in trypanosomes (Acestor et al., 2011; Morales et al., 2009), some of them also present in *E. gracilis*. However, recruitment of taxon-specific subunits is not limited to Euglenozoa, since unrelated additional subunits have also been described in *A. thaliana* (Millar et al., 2004).

2.3. Complex III

Unlike the four other complexes involved in the electron transfer chain, CIII of trypanosomes has a very small set of additional subunits (QCRTB1, QCRTB2 and QCRTB3), one of which (QCRTB2) is a paralogue of the QCR1 core subunit. In *E. gracilis*, seven of the ten classical CIII subunits were found (QCR1/2/6/7, RIP1, COB, and CYT1), along with two of the three subunits described in trypanosomes (QCRTB1/2). It is tempting to speculate that the switch from antimycin A-sensitive CIII activity to myxothiazol-sensitive CIII activity (the so-called *bc*1-bypass) (Moreno-Sánchez et al., 2000; Sharpless and Butow, 1970b) could be due to the differential expression of core 1 paralogues in *E. gracilis* whose presence would modify the affinity of cytochrome *b* Qo and Qi sites for myxothiazol and antimycin A, respectively.

2.4. Complex IV

A preliminary analysis of CIV subunit composition suggested an atypical subunit composition of the enzyme, when compared to mammalian cytochrome *c* oxidase (Bronstrup and Hachtel, 1989). In the

present work, the presence of four classical subunits (COX1/2/3/6B) was confirmed, along with at least nine of the sixteen subunits described in *T. brucei* (COXTB2/4/5/16). These additional subunits probably constitute a different scaffold compared to the one described in mammals (Tsukihara et al., 1996); this may explain why *E. gracilis* cytochrome *c* oxidase activity was low with heterologous bovine cytochrome *c550* and 35-fold higher with the homologous *E. gracilis* cytochrome *c558* (Bronstrup and Hachtel, 1989). Euglena CIII and CIV were also found to form a supercomplex of ~900 kDa that is stabilized in the presence of the mild non-ionic detergent digitonin. This association of CIII and CIV has been previously described in *S. cerevisiae* and *B. taurus* (Cruciat et al., 2000; Schagger and Pfeiffer, 2000).

2.5. Complex V

ATP synthase of *E. gracilis* also exhibited an unusually large molecular mass (> 2 MDa). The dimeric nature of ATP synthase is now largely acknowledged in many organisms, and in mammals and yeasts, the dimer has a molecular mass of ~1.2 MDa and is thought also to be responsible for shaping mitochondrial cristae (Davies et al., 2012; Velours and Arselin, 2000; Walker et al., 1991). Beyond mammals, fungi and flowering plants, the first organism where an unusual structure and subunit composition of the mitochondrial ATP synthase was found was the green alga *C. reinhardtii* and its colourless relative *Polytomella sp.* (~1.6-1.7 MDa) (Dudkina et al., 2006; Vazquez-Acevedo et al., 2006). More recently, unusual subunit composition for the ATP synthase was also reported in the ciliate (alveolates) *T. thermophila* (Balabaskaran Nina et al., 2010), and in *T. brucei* (Zikova et al., 2009). In these species, as in *E. gracilis*, only canonical subunits involved in the F_1 catalytic head (α, β, OSCP), the rotor (a, c ring) and the central axis (δ, ε, γ) have been found. In contrast, subunits involved in the peripheral stator (b/ATP4, d, e, f, g, etc.) are missing and have been replaced by new sets of subunits that have no counterparts in other lineages. In *E. gracilis*, we identified seven of the 14 new subunits discovered in *T. brucei*, for which no homologues could be identified outside Euglenozoa. This strongly suggests that, as we earlier proposed in the case of mitochondrial ATP synthase from chlorophycean green algae (see Discussion in Lapaille et al., 2010), the recruitment of new subunits might be concomitant to the loss of mitochondrial genes for ATP synthase proteins (including subunit b/ATP4 gene) in the course of mitochondrial gene relocation into the nucleus.

In conclusion, our study of the mitochondrial respiratory chain in *E. gracilis* provides evidence that the additional subunits of the different complexes described in trypanosomes are not specific to kinetoplastids, but rather are at least shared with other Euglenozoa. Consequently, their presence

cannot be explained by the parasitic lifestyle of Trypanosomatidae, as many Euglenozoa are non-parasitic (*e.g.* photosynthetic euglenids and free-living diplonemids). However, since many of these subunits are not found beyond Euglenozoa (or Discoba), these differences may explain the biochemical peculiarities observed for the respiratory-chain complexes of kinetoplastids and euglenids.

Acknowledgments. We thank M. Radoux and G. Gain for technical help. This work was supported by University of Liège (SFRD-11/05 to P.C. and SFRD-12/04 to D.B), the Fonds National de la Recherche Scientifique (an Incentive Grant for Scientific Research MIS F.4520, FRFC 2.4597.11; FRFC 2.4567.11; CDR J.0138.13) and FRS-FNRS/CONACyT B330/123/11 (Belgium-Mexico). E.P. is supported by the Belgian FRIA F.R.S.-FNRS, P.C. is Research Associate of F.R.S.-FNRS.

Figure Legends

Figure 1. Scheme of the computational pipeline for *in silico* and proteomic identifications.

Figure 2. Subunits of respiratory-chain complexes and proteins of alternative pathways found within euglenids (*E. gracilis* and *E. longa*), kinetoplastids (*T. brucei*, *T. cruzi* and *L. major*) and Heterolobosea (*N. gruberi*). [#,$]At the exception of Euglena NDUFA8 (see text for details and Figure 6) and QCR9, all proteins were identified by bioinformatic analyses (see supplemental files 1 and 5). Underlined proteins were found in *E. gracilis* by proteomic analyses (see supplemental Table 3 and supplemental file 8). Proteins marked in bold and italics are specific to Euglenozoa. For each sector of the diagram, the proteins are grouped by respiratory complex. Proteins marked with * were also found in diplonemids (*D. papillatum*; see supplemental file 6). [$]Euglena QCR9 protein is not homologous to the canonical QCR9 subunit found in mammals and fungi (see text for details).

Figure 3. Phylogenetic relationships between QCR1, QCR2 and QCRTB2. A. Maximum-likelihood tree (LG+F+Γ_4 model) of the MCL group obtained with I = 1.1 for the QCR1, QCR2 and QCRTB2 reference sequences (169 sequences x 207 amino acid positions). gamma: γ-proteobacteria, Virid.: Viridiplantae, Rhodo.: Rhodophyta, Opist.: Opisthokonta, Amoeb.: Amoebozoa, Stram.: Stramenopiles, Heter.: Heterolobosea, Trypa.: Trypanosomatidae, Eugle.: Euglenozoa, and Alveo.: Alveolata. **B**. Phylogenetic tree showing the paralogy between QCR1, QCR2 and QCRTB2. Black: α-proteobacteria, green: Viridiplantae, red: Rhodophyta, blue: Opisthokonta, violet: Amoebozoa,

brown: Stramenopiles, orange: Excavates, deep pink: Alveolata, yellow: Haptophyceae, grey: Rhizaria, azure: Glaucocystophyceae and pink: Cryptophyta. Branches marked by // and //// were reduced to one half and one quarter of their length, respectively. Bootstrap support values ≥ 50% are shown above the corresponding nodes.

Figure 4. Coomassie-Blue stained images of respiratory-chain complexes from *Euglena gracilis* separated by BN-PAGE. A. 100 µg of protein from crude membrane fraction of cells grown in the dark in presence of 60 mM acetate were solubilized by treatment with n-dodecyl-maltoside (nDM) or digitonine (Dig). Detergent concentrations are given in % (w/v) above the lane. **B**. Crude mitochondrial fraction solubilized by treatment with 1% (w/v) nDM. Ac, 60 mM acetate; Et: 200 mM ethanol; DK: dark; LL: low light: ML: medium light. The different complexes are identified by coloration and/or mass spectrometry analysis of their constituents. C,D,E. Proteins from crude membrane fraction of *Chlamydomonas reinhardtii* (Cr) and *Euglena gracilis* (Eg) cells grown in low light in presence of 60 mM acetate. 1.5% (w/v) detergent. Upper part of BN gels stained for NADH dehydrogenase activity using NBT as an electron acceptor (**C**), for ATPase activity (**D**) or immunoblotted with ATP synthase β subunit antibody (**E**).

Figure 5. Two-dimensional resolution of the mitochondrial protein complexes from *Euglena gracilis*. A. BN gel lane loaded with 500 µg of protein was cut out and placed horizontally for subsequent resolution of the protein complexes into their respective components on Tricine-SDS-PAGE. The main complexes on the first dimension BN-PAGE (Figure 4A,B) are indicated at the top of the gel. Coomassie blue-stained image. The numbered spots correspond to polypeptides that were subject to MS/MS analysis. The corresponding sequences are given in Supplemental File 8 and Supplemental Table 3. UP, database match to an unknown protein; n.i., no annotation due to the lack of a significant score in MS analysis. B. Comparison of partial spot pattern of CIII (Band 4), CIV (Band 5) and supercomplex III + IV (Band 3).

Figure 6. Clustal-Omega alignment between NDUFA8 sequences of *Arabidopsis thaliana*, *Homo sapiens* and *Ciona intestinalis* and the potential NDUFA8 sequences of *Trypanosoma brucei*, *Leishmania major* and *Euglena gracilis*. *: conserved characteristic cysteines of NDUFA8. Alignment formatted with Jalview [version 2.7] (Waterhouse et al., 2009).

Supplemental Table 1. Inventory of subunits constituting the respiratory-chain complexes in model eukaryotes.

Supplemental Table 2. Respiratory rates of *Euglena* cells.

Supplemental Table 3. Mass spectrometry analysis of protein spots.

Supplemental Figure 1. Preliminary phylogenetic trees of MCL groups OGMCL11165 and OGMCL11434 annotated by beta and ND5 reference sequences, respectively. A. Maximum-likelihood tree (LG+F+Γ_4 model) of the MCL group OGMCL11165 obtained with I = 1.2 and including the beta reference sequence (33 sequences x 491 amino acid positions). B. Maximum-likelihood tree (same model) of the MCL group OGMCL11434 obtained with I = 1.2 and including the ND5 reference sequence (29 sequences x 638 amino acid positions). Along with the tree shown in Supplemental Figure 2, a total of 99 preliminary trees were inferred and analysed by hand to select the subtree corresponding to the mitochondrial paralogue. Pmar: *Prochlorococcus marinus* MIT9303, NPCC: *Nostoc* sp. PCC 7120, Mcht: *Microcoleus chthonoplastes* sp. PCC 7420, Ecol: *Escherichia coli* K-12 MG1655, Rpro: *Rickettsia prowazekii* Rp22, Mlot: *Mesorhibozium loti* MAFF303099, Btau: *Bos taurus*, Hsap: *Homo sapiens*, Kpas: *Komagataella pastoris*, Scer: *Saccharomyces cerevisiae*, Dpur: *Dictyostelium purpureum*, Ptri: *Phaeodactylum tricornutum*, Tthe: *Tetrahymena thermophila*, Cmer: *Cyanidioschyzon merolae*, Atha: *Arabidopsis thaliana*, Vcar: *Volvox carteri*, Crei: *Chlamydomonas reinhardtii*, MRCC: *Micromonas* sp. RCC299, Ngru: *Naegleria gruberi*, Lmaj: *Leishmania major*, Tbru: *Trypanosoma brucei*, and Egra: *Euglena gracilis*.

Supplemental Figure 2. Preliminary phylogenetic trees of MCL group OGMCL10071 annotated by NDUFAB1 reference sequence. A. Maximum-likelihood tree (LG+F+Γ_4 model) of the MCL group OGMCL10071 obtained with I = 1.2 and including the NDUFAB1 reference sequence, along with distantly related sequences corresponding to ribosomal protein L21 and chlorophyll a/b binding protein (171 sequences x 182 amino acid positions). B. Maximum-likelihood tree (same model) of the more limited subgroup including the NDUFAB1 reference sequence (30 sequences x 114 amino acid positions). Pmar: *Prochlorococcus marinus* MIT9303, NPCC: *Nostoc* sp. PCC 7120, Mcht: *Microcoleus chthonoplastes* sp. PCC 7420, Ecol: *Escherichia coli* K-12 MG1655, Rpro: *Rickettsia prowazekii* Rp22, Mlot: *Mesorhibozium loti* MAFF303099, Btau: *Bos taurus*, Hsap: *Homo sapiens*, Kpas: *Komagataella pastoris*, Scer: *Saccharomyces cerevisiae*, Dpur: *Dictyostelium purpureum*, Ptri: *Phaeodactylum tricornutum*, Tthe: *Tetrahymena thermophila*, Cmer: *Cyanidioschyzon merolae*, Atha: *Arabidopsis thaliana*, Vcar: *Volvox carteri*, Crei: *Chlamydomonas reinhardtii*, MRCC: *Micromonas* sp. RCC299, Ngru: *Naegleria gruberi*, Lmaj: *Leishmania major*, Tbru: *Trypanosoma brucei*, and Egra: *Euglena gracilis*.

Supplemental Figure 3. Sensitivity of oxygen consumption in the dark to classical potent inhibitors of mitochondrial respiration. Measurements were performed in presence of salicyl hydroxamic acid or SHAM (A), potassium cyanide (B), oligomycin (C), rotenone (D), myxothiazol (E) or antimycine A (F). All measurements were performed in triplicate and data are presented in mean ± SD.

Supplemental Files 1-8: Amino acid sequences of respiratory-chain subunits in surveyed Euglenozoa.

References

Acestor, N., Zikova, A., Dalley, R.A., Anupama, A., Panigrahi, A.K., Stuart, K.D., 2011. *Trypanosoma brucei* mitochondrial respiratome: composition and organization in procyclic form. Mol Cell Proteomics 10, M110 006908.

Adl, S.M., Simpson, A.G., Lane, C.E., Lukes, J., Bass, D., Bowser, S.S., Brown, M.W., Burki, F., Dunthorn, M., Hampl, V., Heiss, A., Hoppenrath, M., Lara, E., Le Gall, L., Lynn, D.H., McManus, H., Mitchell, E.A., Mozley-Stanridge, S.E., Parfrey, L.W., Pawlowski, J., Rueckert, S., Shadwick, R.S., Schoch, C.L., Smirnov, A., Spiegel, F.W., 2012. The revised classification of eukaryotes. J Eukaryot Microbiol 59, 429-493.

Angerer, H., Zwicker, K., Wumaier, Z., Sokolova, L., Heide, H., Steger, M., Kaiser, S., Nubel, E., Brutschy, B., Radermacher, M., Brandt, U., Zickermann, V., 2011. A scaffold of accessory subunits links the peripheral arm and the distal proton-pumping module of mitochondrial complex I. Biochem J 437, 279-288.

Balabaskaran Nina, P., Dudkina, N.V., Kane, L.A., van Eyk, J.E., Boekema, E.J., Mather, M.W., Vaidya, A.B., 2010. Highly divergent mitochondrial ATP synthase complexes in *Tetrahymena thermophila*. PLoS Biol 8, e1000418.

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W., 2013. GenBank. Nucleic Acids Res 41, D36-42.

Bradford, M.M., 1976. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal Biochem 72, 248-254.

Bronstrup, U., Hachtel, W., 1989. Cytochrome *c* oxidase of *Euglena gracilis*: purification, characterization, and identification of mitochondrially synthesized subunits. J Bioenerg Biomembr 21, 359-373.

Brown, G.E., Preston, J.F., 1975. Changes in mitochondrial density and succinic dehydrogenase activity in *Euglena gracilis* as a function of the dependency on light for growth. Arch Microbiol 104, 233-236.

Buetow, D.E., 1961. Ethanol Stimulation of Oxidative Metabolism in *Euglena gracilis*. Nature 190, 1196-1196.

Burger, G., Gray, M.W., Forget, L., Lang, B.F., 2013. Strikingly bacteria-like and gene-rich mitochondrial genomes throughout jakobid protists. Genome Biol Evol 5, 418-438.

Cardol, P., 2011. Mitochondrial NADH:ubiquinone oxidoreductase (complex I) in eukaryotes: a highly conserved subunit composition highlighted by mining of protein databases. Biochim Biophys Acta 1807, 1390-1397.
Cardol, P., Boutaffala, L., Memmi, S., Devreese, B., Matagne, R.F., Remacle, C., 2008. In *Chlamydomonas*, the loss of ND5 subunit prevents the assembly of whole mitochondrial complex I and leads to the formation of a low abundant 700 kDa subcomplex. Biochim Biophys Acta 1777, 388-396.
Cardol, P., Matagne, R.F., Remacle, C., 2002. Impact of mutations affecting ND mitochondria-encoded subunits on the activity and assembly of complex I in *Chlamydomonas*. Implication for the structural organization of the enzyme. J Mol Biol 319, 1211-1221.
Cardol, P., Vanrobaeys, F., Devreese, B., Van Beeumen, J., Matagne, R.F., Remacle, C., 2004. Higher plant-like subunit composition of mitochondrial complex I from *Chlamydomonas reinhardtii*: 31 conserved components among eukaryotes. Biochim Biophys Acta 1658, 212-224.
Carroll, J., Fearnley, I.M., Skehel, J.M., Shannon, R.J., Hirst, J., Walker, J.E., 2006. Bovine complex I is a complex of 45 different subunits. J Biol Chem 281, 32724-32727.
Castresana, J., 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17, 540-552.
Castro-Guerrero, N.A., Jasso-Chavez, R., Moreno-Sánchez, R., 2005. Physiological role of rhodoquinone in *Euglena gracilis* mitochondria. Biochim Biophys Acta 1710, 113-121.
Castro-Guerrero, N.A., Krab, K., Moreno-Sánchez, R., 2004. The alternative respiratory pathway of *Euglena* mitochondria. J Bioenerg Biomembr 36, 459-469.
Cavalier-Smith, T., 1981. Eukaryote kingdoms: seven or nine? Biosystems 14, 461-481.
Chaudhuri, M., Ott, R.D., Hill, G.C., 2006. Trypanosome alternative oxidase: from molecule to function. Trends Parasitol 22, 484-491.
Collinson, I.R., Runswick, M.J., Buchanan, S.K., Fearnley, I.M., Skehel, J.M., van Raaij, M.J., Griffiths, D.E., Walker, J.E., 1994. Fo membrane domain of ATP synthase from bovine heart mitochondria: purification, subunit composition, and reconstitution with F1-ATPase. Biochemistry 33, 7971-7978.
Cruciat, C.M., Brunner, S., Baumann, F., Neupert, W., Stuart, R.A., 2000. The cytochrome *bc*1 and cytochrome *c* oxidase complexes associate to form a single supracomplex in yeast mitochondria. J Biol Chem 275, 18093-18098.
Cui, J.Y., Mukai, K., Saeki, K., Matsubara, H., 1994. Molecular cloning and nucleotide sequences of cDNAs encoding subunits I, II, and IX of *Euglena gracilis* mitochondrial complex III. J Biochem 115, 98-107.
Davies, K.M., Anselmi, C., Wittig, I., Faraldo-Gomez, J.D., Kuhlbrandt, W., 2012. Structure of the yeast F1Fo-ATP synthase dimer and its role in shaping the mitochondrial cristae. Proc Natl Acad Sci U S A 109, 13602-13607.
Duby, F., Matagne, R.F., 1999. Alteration of dark respiration and reduction of phototrophic growth in a mitochondrial DNA deletion mutant of *Chlamydomonas* lacking cob, nd4, and the 3' end of nd5. Plant Cell 11, 115-125.
Dudkina, N.V., Eubel, H., Keegstra, W., Boekema, E.J., Braun, H.P., 2005. Structure of a mitochondrial supercomplex formed by respiratory-chain complexes I and III. Proc Natl Acad Sci U S A 102, 3225-3229.
Dudkina, N.V., Sunderhaus, S., Braun, H.P., Boekema, E.J., 2006. Characterization of dimeric ATP synthase and cristae membrane ultrastructure from *Saccharomyces* and *Polytomella* mitochondria. FEBS Lett 580, 3427-3432.
Dyall, S.D., Brown, M.T., Johnson, P.J., 2004. Ancient invasions: from endosymbionts to organelles. Science 304, 253-257.
Ebersberger, I., Strauss, S., von Haeseler, A., 2009. HaMStR: profile hidden markov model based search for orthologs in ESTs. BMC Evol Biol 9, 157.
Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32, 1792-1797.

Fang, J., Beattie, D.S., 2002. Novel FMN-containing rotenone-insensitive NADH dehydrogenase from *Trypanosoma brucei* mitochondria: isolation and characterization. Biochemistry 41, 3065-3072.
Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution, 783-791.
Field, D., Tiwari, B., Booth, T., Houten, S., Swan, D., Bertrand, N., Thurston, M., 2006. Open software for biologists: from famine to feast. Nat Biotechnol 24, 801-803.
Fischer, S., Brunk, B.P., Chen, F., Gao, X., Harb, O.S., Iodice, J.B., Shanmugam, D., Roos, D.S., Stoeckert, C.J., Jr., 2011. Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. Curr Protoc Bioinformatics Chapter 6, Unit 6 12 11-19.
Gabaldon, T., Huynen, M.A., 2005. Lineage-specific gene loss following mitochondrial endosymbiosis and its potential for function prediction in eukaryotes. Bioinformatics 21 Suppl 2, ii144-150.
Gabaldon, T., Koonin, E.V., 2013. Functional and evolutionary implications of gene orthology. Nat Rev Genet 14, 360-366.
Gawryluk, R.M., Chisholm, K.A., Pinto, D.M., Gray, M.W., 2012. Composition of the mitochondrial electron transport chain in *Acanthamoeba castellanii*: structural and evolutionary insights. Biochim Biophys Acta 1817, 2027-2037.
Gawryluk, R.M., Gray, M.W., 2009. A split and rearranged nuclear gene encoding the iron-sulfur subunit of mitochondrial succinate dehydrogenase in Euglenozoa. BMC Res Notes 2, 16.
Gawryluk, R.M., Gray, M.W., 2010. Evidence for an early evolutionary emergence of gamma-type carbonic anhydrases as components of mitochondrial respiratory complex I. BMC Evol Biol 10, 176.
Gibbs, S.P., 1981. The chloroplasts of some algal groups may have evolved from endosymbiotic eukaryotic algae. Ann N Y Acad Sci 361, 193-208.
Gouy, M., Guindon, S., Gascuel, O., 2010. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 27, 221-224.
Gray, M.W., Burger, G., Lang, B.F., 1999. Mitochondrial evolution. Science 283, 1476-1481.
Gray, M.W., Lang, B.F., Burger, G., 2004. Mitochondria of protists. Annu Rev Genet 38, 477-524.
Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59, 307-321.
Hampl, V., Hug, L., Leigh, J.W., Dacks, J.B., Lang, B.F., Simpson, A.G., Roger, A.J., 2009. Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". Proc Natl Acad Sci U S A 106, 3859-3864.
Harris, E.H., 1989. The *Chlamydomonas* Sourcebook. Academic Press, San Diego.
Heazlewood, J.L., Whelan, J., Millar, A.H., 2003. The products of the mitochondrial orf25 and orfB genes are FO components in the plant F1FO ATP synthase. FEBS Lett 540, 201-205.
Huang, X., Madan, A., 1999. CAP3: A DNA sequence assembly program. Genome Res 9, 868-877.
Huynen, M.A., de Hollander, M., Szklarczyk, R., 2009. Mitochondrial proteome evolution and genetic disease. Biochim Biophys Acta 1792, 1122-1129.
Kadenbach, B., Jarausch, J., Hartmann, R., Merle, P., 1983. Separation of mammalian cytochrome *c* oxidase into 13 polypeptides by a sodium dodecyl sulfate-gel electrophoretic procedure. Anal Biochem 129, 517-521.
Klodmann, J., Sunderhaus, S., Nimtz, M., Jansch, L., Braun, H.P., 2010. Internal architecture of mitochondrial complex I from *Arabidopsis thaliana*. Plant Cell 22, 797-810.
Koski, L.B., Golding, G.B., 2001. The closest BLAST hit is often not the nearest neighbor. J Mol Evol 52, 540-542.
Lang, B.F., Burger, G., O'Kelly, C.J., Cedergren, R., Golding, G.B., Lemieux, C., Sankoff, D., Turmel, M., Gray, M.W., 1997. An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. Nature 387, 493-497.
Lapaille, M., Escobar-Ramirez, A., Degand, H., Baurain, D., Rodriguez-Salinas, E., Coosemans, N., Boutry, M., Gonzalez-Halphen, D., Remacle, C., Cardol, P., 2010. Atypical subunit composition of the chlorophycean mitochondrial F1FO-ATP synthase and role of Asa7 protein in stability and oligomycin resistance of the enzyme. Mol Biol Evol 27, 1630-1644.

Le, S.Q., Gascuel, O., 2008. An improved general amino acid replacement matrix. Mol Biol Evol 25, 1307-1320.
Li, L., Stoeckert, C.J., Jr., Roos, D.S., 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13, 2178-2189.
Martin, W., Muller, M., 1998. The hydrogen hypothesis for the first eukaryote. Nature 392, 37-41.
Mego, J.L., Farb, R.M., 1974. Alcohol dehydrogenases of *Euglena gracilis*, strain Z. Biochim Biophys Acta 350, 237-239.
Michalecka, A.M., Svensson, A.S., Johansson, F.I., Agius, S.C., Johanson, U., Brennicke, A., Binder, S., Rasmusson, A.G., 2003. *Arabidopsis* genes encoding mitochondrial type II NAD(P)H dehydrogenases have different evolutionary origin and show distinct responses to light. Plant Physiol 133, 642-652.
Millar, A.H., Eubel, H., Jansch, L., Kruft, V., Heazlewood, J.L., Braun, H.P., 2004. Mitochondrial cytochrome *c* oxidase and succinate dehydrogenase complexes contain plant specific subunits. Plant Mol Biol 56, 77-90.
Monroy, A.F., Schwartzbach, S.D., 1984. Catabolite repression of chloroplast development in *Euglena*. Proc Natl Acad Sci U S A 81, 2786-2790.
Morales, J., Mogi, T., Mineki, S., Takashima, E., Mineki, R., Hirawake, H., Sakamoto, K., Omura, S., Kita, K., 2009. Novel mitochondrial complex II isolated from *Trypanosoma cruzi* is composed of 12 peptides including a heterodimeric Ip subunit. J Biol Chem 284, 7255-7263.
Moreno-Sánchez, R., Covian, R., Jasso-Chavez, R., Rodriguez-Enriquez, S., Pacheco-Moises, F., Torres-Marquez, M.E., 2000. Oxidative phosphorylation supported by an alternative respiratory pathway in mitochondria from *Euglena*. Biochim Biophys Acta 1457, 200-210.
Moreno-Sánchez, R., Raya, J.C., 1987. Preparation of coupled mitochondria from *Euglena* by sonication. Plant Science 48, 151-157.
Panigrahi, A.K., Zikova, A., Dalley, R.A., Acestor, N., Ogata, Y., Anupama, A., Myler, P.J., Stuart, K.D., 2008. Mitochondrial complexes in *Trypanosoma brucei*: a novel complex and a unique oxidoreductase complex. Mol Cell Proteomics 7, 534-545.
Pullman, M.E., Penefsky, H.S., Datta, A., Racker, E., 1960. Partial resolution of the enzymes catalyzing oxidative phosphorylation. I. Purification and properties of soluble dinitrophenol-stimulated adenosine triphosphatase. J Biol Chem 235, 3322-3329.
Qiu, X.B., Shao, Y.M., Miao, S., Wang, L., 2006. The diversity of the DnaJ/Hsp40 family, the crucial partners for Hsp70 chaperones. Cell Mol Life Sci 63, 2560-2570.
Remacle, C., Baurain, D., Cardol, P., Matagne, R.F., 2001. Mutants of *Chlamydomonas reinhardtii* deficient in mitochondrial complex I: characterization of two mutations affecting the nd1 coding sequence. Genetics 158, 1051-1060.
Saccone, C., Lanave, C., De Grassi, A., 2006. Metazoan OXPHOS gene families: evolutionary forces at the level of mitochondrial and nuclear genomes. Biochim Biophys Acta 1757, 1171-1178.
Schagger, H., Pfeiffer, K., 2000. Supercomplexes in the respiratory chains of yeast and mammalian mitochondria. EMBO J 19, 1777-1783.
Schagger, H., von Jagow, G., 1991. Blue native electrophoresis for isolation of membrane protein complexes in enzymatically active form. Anal Biochem 199, 223-231.
Sharpless, T.K., Butow, R.A., 1970a. An inducible alternate terminal oxidase in *Euglena gracilis* mitochondria. J Biol Chem 245, 58-70.
Sharpless, T.K., Butow, R.A., 1970b. Phosphorylation sites, cytochrome complement, and alternate pathways of coupled electron transport in *Euglena gracilis* mitochondria. J Biol Chem 245, 50-57.
Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., Thompson, J.D., Higgins, D.G., 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7, 539.
Simpson, A.G.B., 1997. The Identity and Composition of the Euglenozoa. Arch. Protistenkd. 148, 10.
Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G., Korf, I., Lapp, H., Lehvaslaiho, H., Matsalla, C., Mungall, C.J., Osborne, B.I., Pocock, M.R., Schattner, P., Senger, M., Stein, L.D., Stupka, E., Wilkinson, M.D., Birney, E., 2002. The Bioperl toolkit: Perl modules for the life sciences. Genome Res 12, 1611-1618.

Sunderhaus, S., Dudkina, N.V., Jansch, L., Klodmann, J., Heinemeyer, J., Perales, M., Zabaleta, E., Boekema, E.J., Braun, H.P., 2006. Carbonic anhydrase subunits form a matrix-exposed domain attached to the membrane arm of mitochondrial complex I in plants. J Biol Chem 281, 6482-6488.
Tsukihara, T., Aoyama, H., Yamashita, E., Tomizaki, T., Yamaguchi, H., Shinzawa-Itoh, K., Nakashima, R., Yaono, R., Yoshikawa, S., 1996. The whole structure of the 13-subunit oxidized cytochrome *c* oxidase at 2.8 A. Science 272, 1136-1144.
Turmel, M., Gagnon, M.C., O'Kelly, C.J., Otis, C., Lemieux, C., 2009. The chloroplast genomes of the green algae *Pyramimonas*, *Monomastix*, and *Pycnococcus* shed new light on the evolutionary history of prasinophytes and the origin of the secondary chloroplasts of euglenids. Mol Biol Evol 26, 631-648.
Van Aken, O., Giraud, E., Clifton, R., Whelan, J., 2009. Alternative oxidase: a target and regulator of stress responses. Physiol Plant 137, 354-361.
Vazquez-Acevedo, M., Cardol, P., Cano-Estrada, A., Lapaille, M., Remacle, C., Gonzalez-Halphen, D., 2006. The mitochondrial ATP synthase of chlorophycean algae contains eight subunits of unknown origin involved in the formation of an atypical stator-stalk and in the dimerization of the complex. J Bioenerg Biomembr 38, 271-282.
Velours, J., Arselin, G., 2000. The *Saccharomyces cerevisiae* ATP synthase. J Bioenerg Biomembr 32, 383-390.
Villavicencio-Queijeiro, A., Vazquez-Acevedo, M., Cano-Estrada, A., Zarco-Zavala, M., Tuena de Gomez, M., Mignaco, J.A., Freire, M.M., Scofano, H.M., Foguel, D., Cardol, P., Remacle, C., Gonzalez-Halphen, D., 2009. The fully-active and structurally-stable form of the mitochondrial ATP synthase of *Polytomella* sp. is dimeric. J Bioenerg Biomembr 41, 1-13.
Walker, G., Dorrell, R.G., Schlacht, A., Dacks, J.B., 2011. Eukaryotic systematics: a user's guide for cell biologists and parasitologists. Parasitology 138, 1638-1663.
Walker, J.E., Lutter, R., Dupuis, A., Runswick, M.J., 1991. Identification of the subunits of F1F0-ATPase from bovine heart mitochondria. Biochemistry 30, 5369-5378.
Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M., Barton, G.J., 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics 25, 1189-1191.
Wilson, B.W., Danforth, W.F., 1958. The extent of acetate and ethanol oxidation by *Euglena gracilis*. J Gen Microbiol 18, 535-542.
Yang, Z., 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol 10, 1396-1401.
Zerbino, D.R., Birney, E., 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18, 821-829.
Zhao, S., Burki, F., Brate, J., Keeling, P.J., Klaveness, D., Shalchian-Tabrizi, K., 2012. *Collodictyon*--an ancient lineage in the tree of eukaryotes. Mol Biol Evol 29, 1557-1568.
Zikova, A., Panigrahi, A.K., Uboldi, A.D., Dalley, R.A., Handman, E., Stuart, K., 2008. Structural and functional association of *Trypanosoma brucei* MIX protein with cytochrome *c* oxidase complex. Eukaryot Cell 7, 1994-2003.
Zikova, A., Schnaufer, A., Dalley, R.A., Panigrahi, A.K., Stuart, K.D., 2009. The F(0)F(1)-ATP synthase complex contains novel subunits and is essential for procyclic *Trypanosoma brucei*. PLoS Pathog 5, e1000436.

Figure 1 (revised)
Click here to download high resolution image

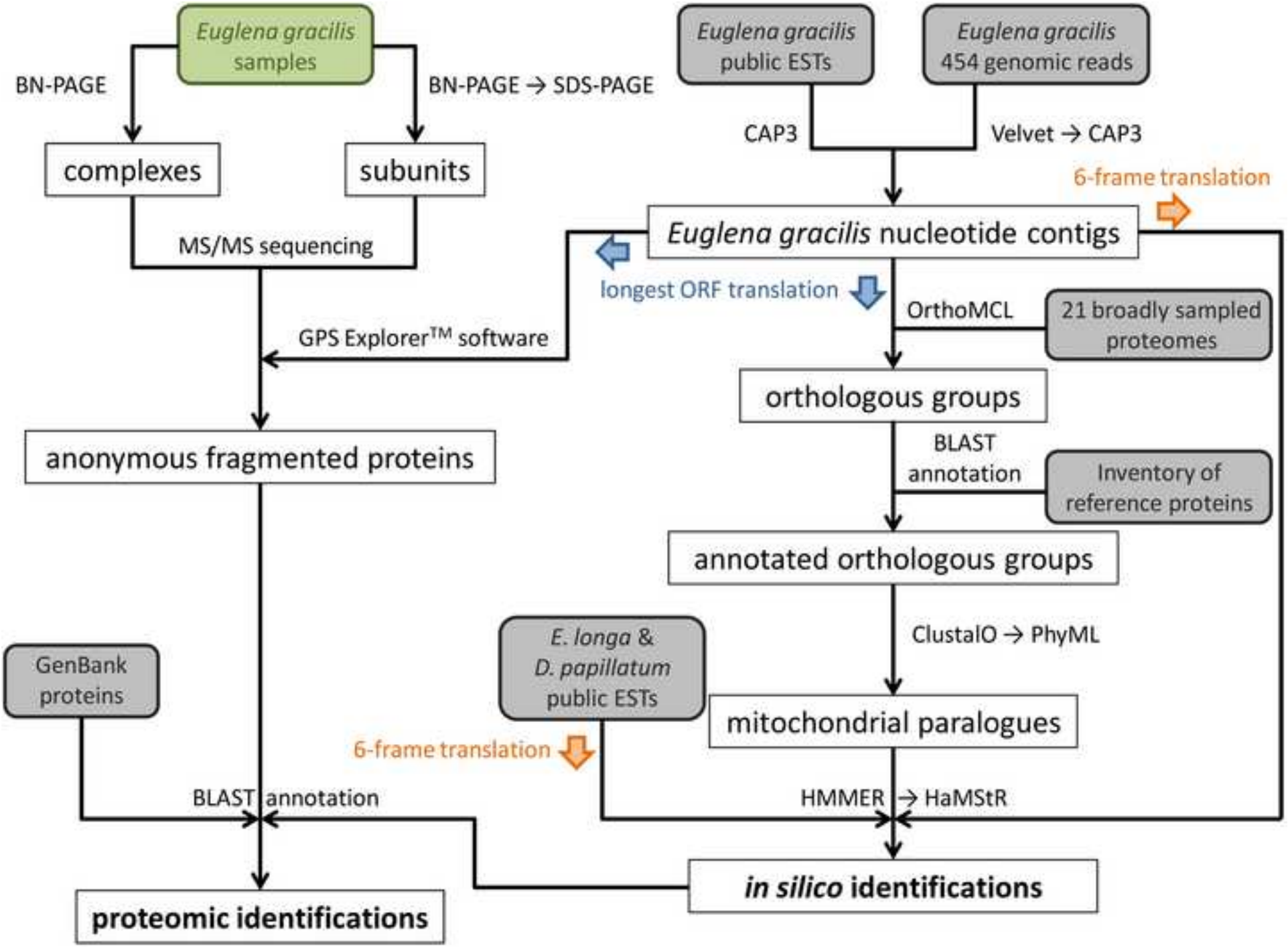


Figure 2 (revised)
Click here to download high resolution image

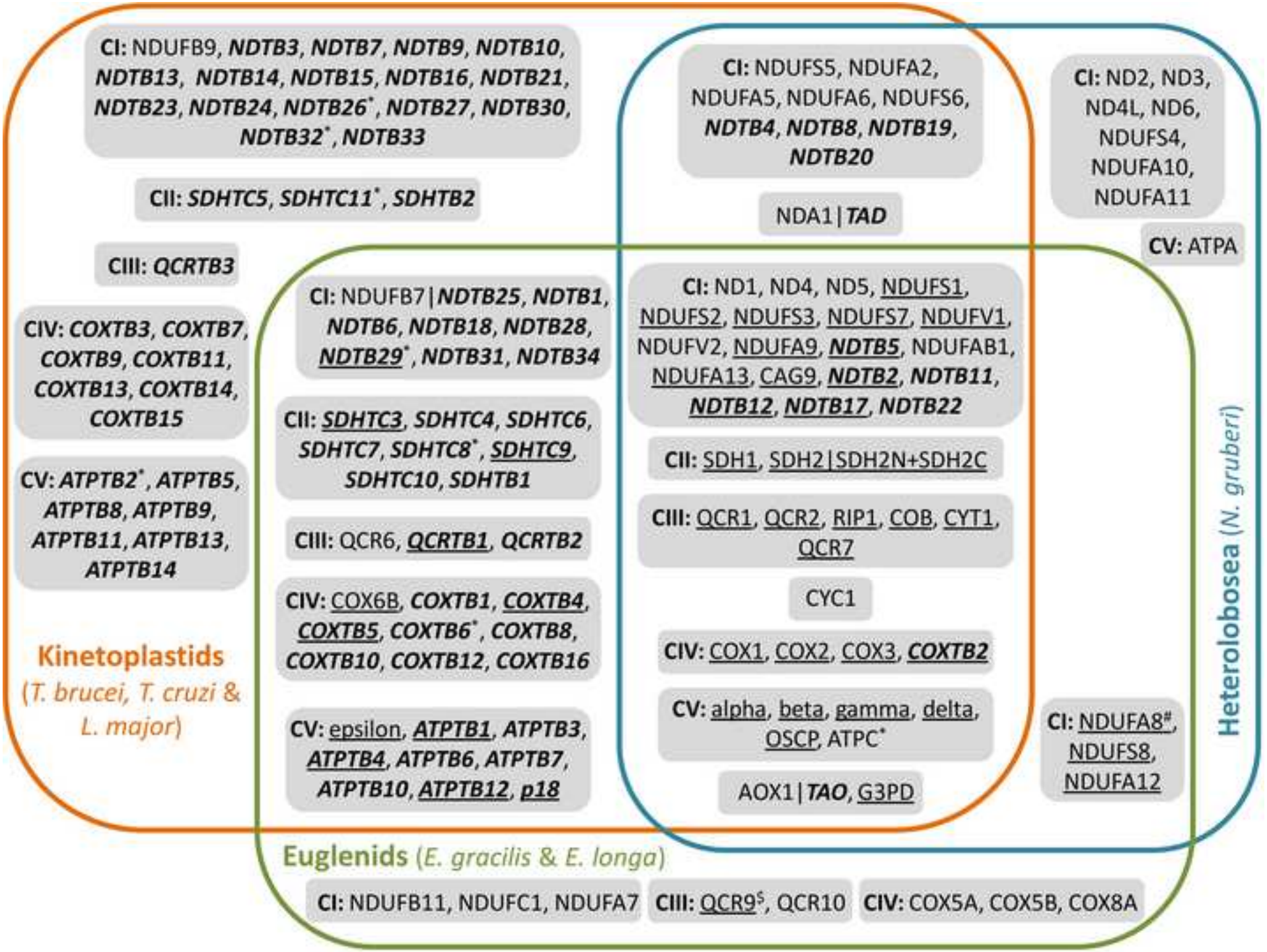


Figure 3

Click here to download high resolution image

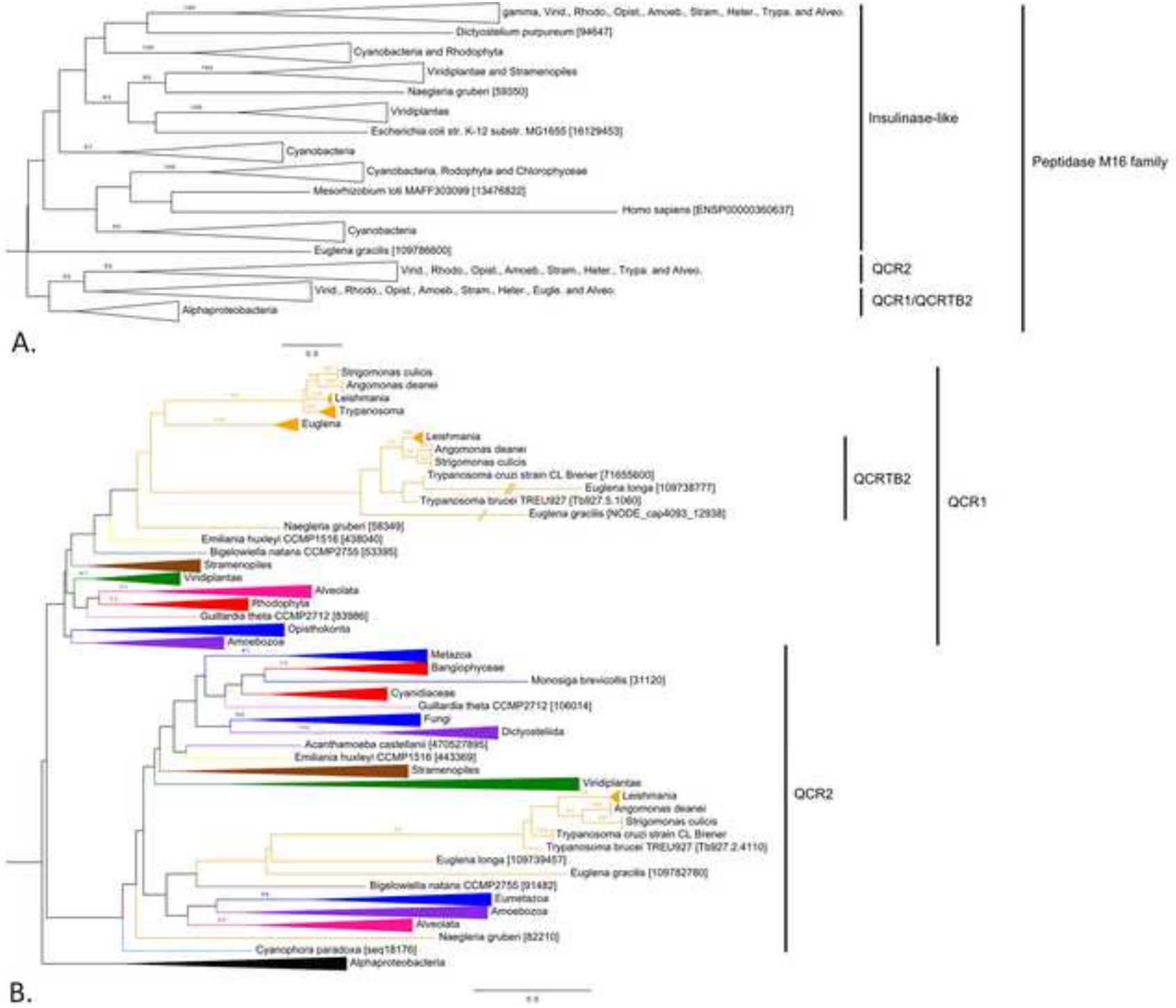


Figure 4

Click here to download high resolution image

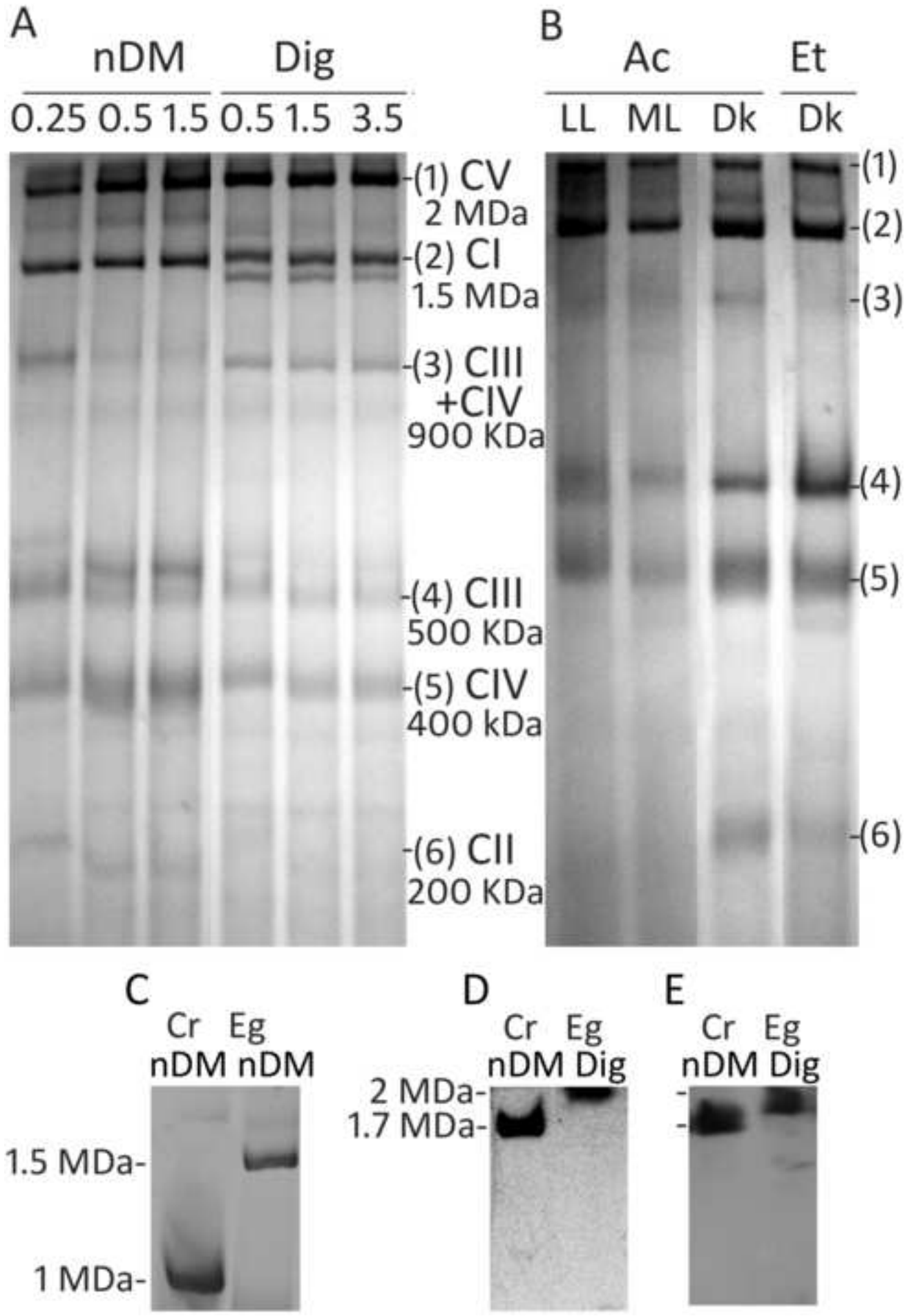


Figure 5

Click here to download high resolution image

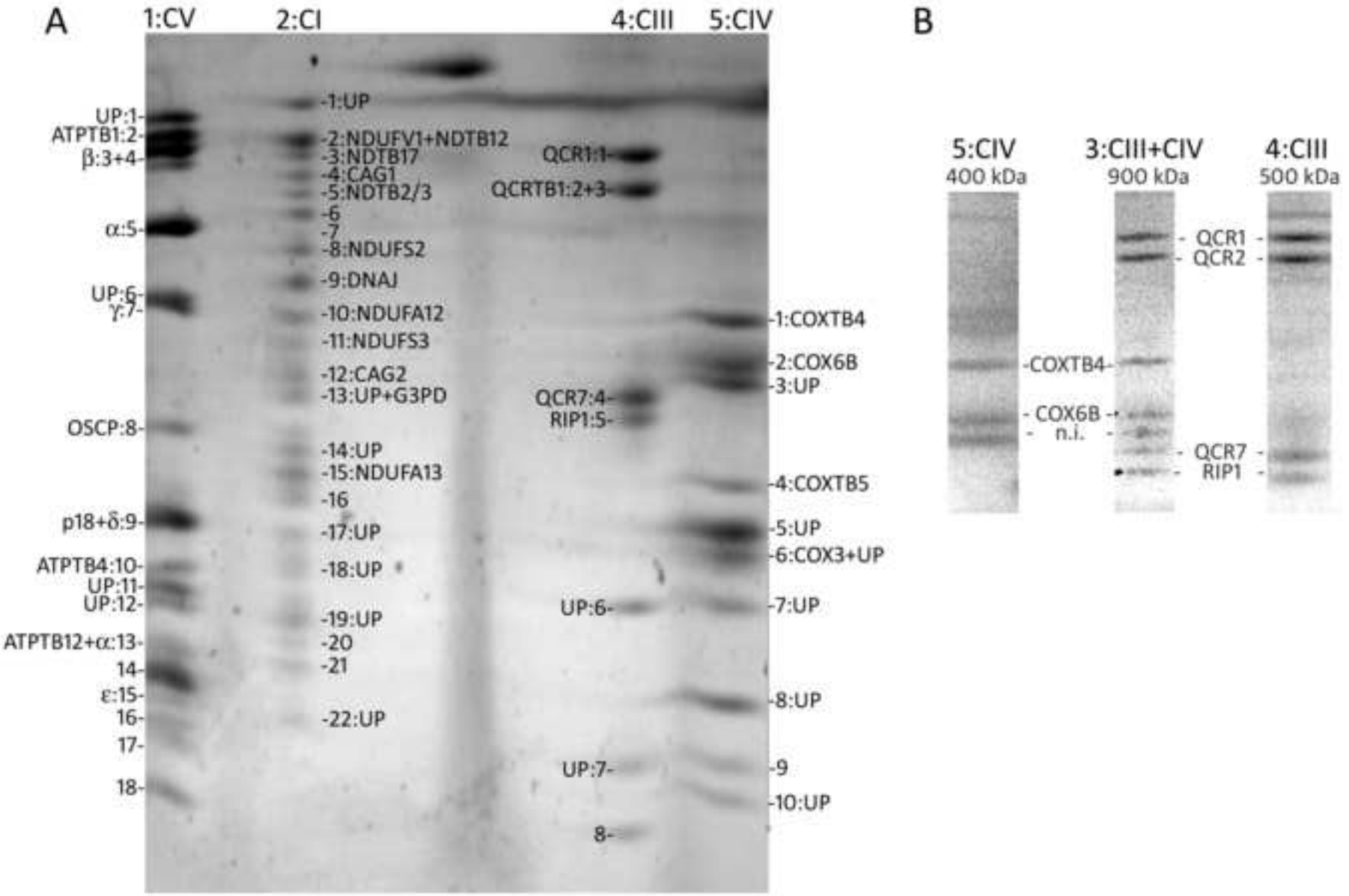
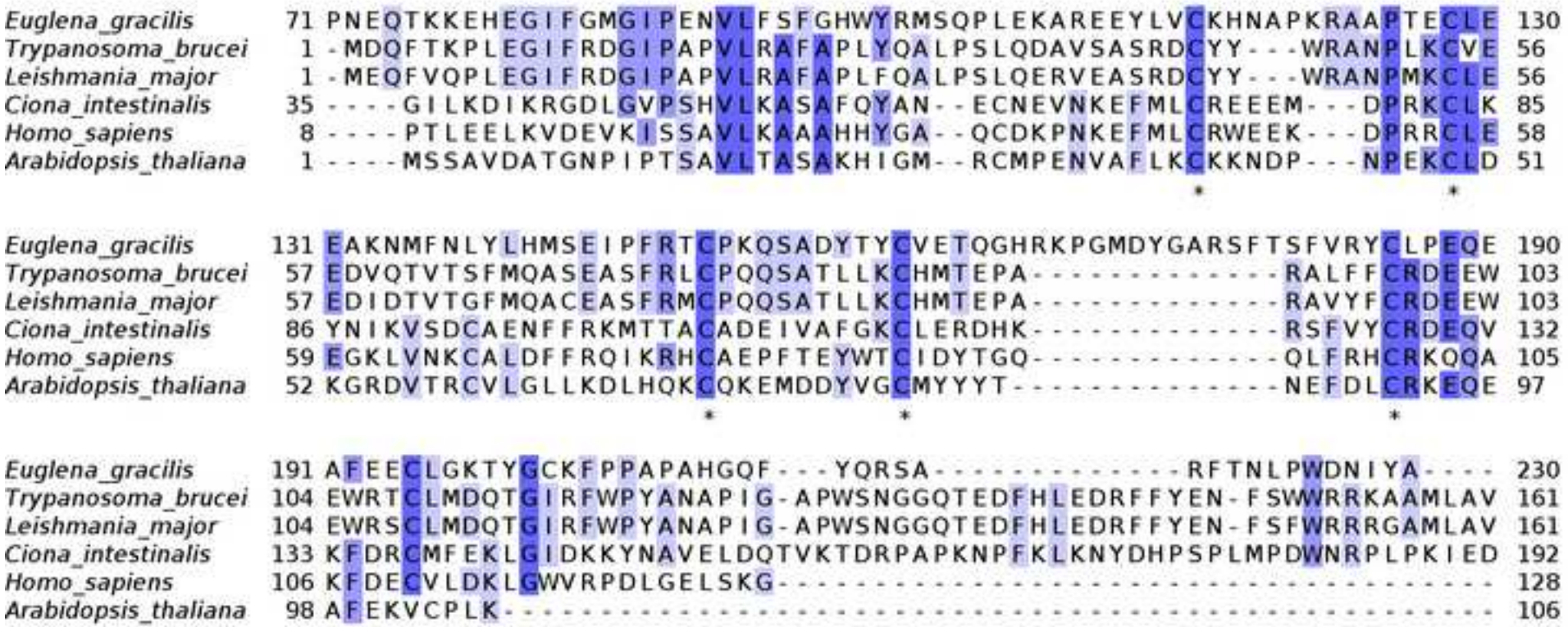


Figure 6

Click here to download high resolution image



Supplementary Table 1 (revised)

Click here to download Supplementary material: REVISED_TABLE_S1_PEREZ.xls

Supplementary Table 2 (revised)

Click here to download Supplementary material: REVISED_TABLE_S2_PEREZ.doc

Supplementary Table 3 (revised)

Click here to download Supplementary material: REVISED_TABLE_S3_PEREZ.xls

Supplementary Figure 1 (revised)

Click here to download Supplementary material: REVISED_FIGURE_S1_PEREZ.tif

Supplementary Figure 2 (revised)

Click here to download Supplementary material: REVISED_FIGURE_S2_PEREZ.tif

Supplementary Figure 3 (revised)

Click here to download Supplementary material: REVISED_FIGURE_S3_PEREZ.tif

Supplementary file 1 (revised)

Click here to download Supplementary material: Supplemental file 1 Egra_ID-MCL_bioinfo_genes.fasta

Supplementary file 2 (revised)

Click here to download Supplementary material: Supplemental file 2 Lmaj_ID-MCL_bioinfo_genes.fasta

Supplementary file 3(revised)

Click here to download Supplementary material: Supplemental file 3 Ngru_ID-MCL_bioinfo_genes.fasta

Supplementary file 4(revised)

Click here to download Supplementary material: Supplemental file 4 Tbru_ID-MCL_bioinfo_genes.fasta

Supplementary file 5(revised)

Click here to download Supplementary material: Supplemental file 5 Egra_hamstr_bioinfo_new_names_genes.fasta

Supplementary file 6 (revised)

Click here to download Supplementary material: Supplemental file 6 Dpap_hamstr_bioinfo_new_names_genes.fasta

Supplementary file 7 (revised)

Click here to download Supplementary material: Supplemental file 7 Elon_hamstr_bioinfo_new_names_genes.fasta

Supplementary file 8 (revised)

Click here to download Supplementary material: Revised_Supplemental file 8 Egra_ID-spots-C_proteomic_genes.fasta