Evolution of the Cytosolic Iron/Sulfur cluster Assembly machinery in Blastocystis sp. and other microbial eukaryotes

Anastasios D. Tsaousis\textsuperscript{1,2,*}, Eleni Gentekaki\textsuperscript{1,§}, Laura Eme\textsuperscript{1,§}, Daniel Gaston\textsuperscript{1} & Andrew J. Roger\textsuperscript{1,*}

Short title: The CIA machinery in microbial eukaryotes

1. Centre for Comparative Genomics and Evolutionary Bioinformatics, Dalhousie University, Department of Biochemistry and Molecular Biology, Halifax, NS, B3H 4R2, Canada
2. Laboratory of Molecular and Evolutionary Parasitology, School of Biosciences, University of Kent, Canterbury, CT2 7NJ, UK

* Corresponding authors: Anastasios D. Tsaousis (Laboratory of Molecular and Evolutionary Parasitology, School of Biosciences, University of Kent, Canterbury, CT2 7NJ, UK; email: tsaousis.anastasios@gmail.com) and Andrew J. Roger (Centre for Comparative Genomics and Evolutionary Bioinformatics, Dalhousie University, Department of Biochemistry and Molecular Biology, 5850 College Street, Halifax, Canada, B3H 4R2; email: andrew.roger@dal.ca); phone: +1 902 494 2881, fax: +1 902 494 1355

§: these authors made equal contributions to this manuscript
Abstract

The Cytosolic Iron/Sulfur cluster Assembly (CIA) machinery is responsible for the assembly of cytosolic and nuclear iron/sulfur clusters, cofactors that are vital for all living cells. This machinery is uniquely found in eukaryotes and consists of at least eight proteins in opisthokont lineages such as animals and yeast. We sought to identify and characterize homologues of the CIA system proteins in the anaerobic stramenopile parasite *Blastocystis* sp. NandII strain. We identified transcripts encoding six of the components – Cia1, Cia2, MMS19, Nbp35, Nar1, and a putative Tah18 – and showed that the last three of them localized to the cytoplasm of the cell using immunofluorescence microscopy, immuno-electron microscopy and subcellular fractionation. We then used comparative genomic and phylogenetic approaches to investigate the evolutionary history of these proteins. While most *Blastocystis* homologues branch with their eukaryotic counterparts, the putative *Blastocystis* Tah18 seems to have a separate evolutionary origin and therefore possibly a different function. Furthermore, our phylogenomic analyses revealed that all eight CIA components described in opisthokonts originated before the diversification of extant eukaryotic lineages and were likely already present in the Last Eukaryotic Common Ancestor (LECA). Nbp35, Nar1 Cia1 and Cia2 proteins have been conserved during the subsequent evolutionary diversification of eukaryotes and are present in virtually all extant lineages, whereas the other CIA proteins have patchy phylogenetic distributions. Cia2 appears to be homologous to SufT, a component of the prokaryotic SUF system, making this the first reported evolutionary link between the CIA and any other Fe/S biogenesis pathway. All of our results suggest that the CIA machinery is an ubiquitous biosynthetic pathway in eukaryotes, but its
apparent plasticity in composition raises questions regarding how it functions in non-model organisms and how it interfaces with various iron/sulfur cluster systems (i.e., ISC, NIF and/or SUF) found in eukaryotic cells.

Introduction

The assembly of iron/sulfur (Fe/S) clusters is considered one of the basic biosynthetic functions in all living cells. Both prokaryotic and eukaryotic organisms have at least one pathway dedicated to Fe/S cluster biosynthesis. In eukaryotes, these clusters are assembled by distinct biosynthetic pathways, which are localized in different compartments of the cell. The Sulfur Utilization Factors (SUF) system is typically found in plastid-bearing organisms and ensures maturation of apo-proteins within plastids, whereas this function is performed by the Iron/Sulfur Cluster (ISC) machinery in mitochondria and mitochondrion-related organelles (MRO). The ISC system is indirectly functionally linked with the Cytosolic Iron/sulfur cluster Assembly (CIA) machinery, which is involved in the maturation of cytosolic and nuclear apo-proteins (1). Many of these Fe/S cluster-bearing proteins are involved in key enzymatic activities such as DNA replication and repair, rRNA processing and telomere stability (2, 3). Consequently, depletions of components of the CIA system are lethal in *Saccharomyces cerevisiae* corroborating the significant role of this biosynthetic pathway in the cell (4-8).

Fe/S cluster biosynthetic machineries typically consist of five main parts: a desulfurase, an iron donor, an electron transfer mechanism, a scaffold, and Fe/S cluster transfer proteins. It is noteworthy that, while the SUF, NIF and ISC systems share several homologous components, no evolutionary link has yet been reported between
CIA system proteins and their counterparts in other Fe/S cluster machineries. Currently, it is known that the CIA pathway involves at least eight proteins in yeast and humans – Dre2, Tah18, Nbp35, Cfd1, Cia1, Cia2, Nar1 and MMS19 (Figure S1) – and that their function is dependent on the mitochondrial ISC machinery (9). The pathway starts when one of the two Fe/S clusters in Dre2 (human Ciapin1) is reduced by the diflavin reductase Tah18 (human Ndor1) (10). Nbp35 and Cfd1 then form a scaffold complex for assembling transiently-bound Fe/S clusters (11). In a later stage, mature Fe/S clusters are transferred to apo-proteins via Cia1 (human CIA01) and Nar1 (human IOP1) proteins, a process facilitated through the formation of a “CIA-targeting complex” with the recently discovered chaperone protein MMS19 (also known as MET18 in yeast) (9, 12, 13) and Cia2 (9, 12).

In most microbial eukaryotes the role of the CIA machinery in the maturation of the cytosolic and nuclear apo-proteins is unclear, as are its functional interactions with other Fe/S cluster maturation pathways, in particular in organisms bearing more than one cytosolic Fe/S system. For example, the microaerophiles Entamoeba and Mastigamoeba both possess functionally-reduced MROs that do not contain the typical ISC pathway. Instead, their genomes encode Nitrogen Fixation (NIF) system components that were acquired by lateral gene transfer (LGT) from ε-proteobacteria, and which localize both in the cytosol and in their organelles (14-16). Another example is found in Blastocystis, an obligate anaerobic parasite that encodes a functional fused version of the SufC and SufB proteins, SufCB, that was also acquired by LGT, in this case from a methanoarchaeal lineage, and functions in the cytosol of the organism (17). The presence of both this SUF
protein and components of the CIA Fe/S cluster biosynthetic pathway in *Blastocystis* [20] raises questions regarding their respective cytosolic roles.

In this manuscript we focus on the CIA system in *Blastocystis* sp. We show that this organism expresses homologues of six of the CIA system proteins – Nbp35, Nar1, Cia1, Cia2, Tah18, and MMS19 – and illustrate by immunomicroscopy and subcellular fractionation that three of them are localized in the cytosol. Using comparative genomic and phylogenetic approaches, we then investigated the distribution and evolutionary histories of the CIA components among eukaryotes.

**Materials & Methods**

*Blastocystis* culture and maintenance:

*Blastocystis* sp. NandII cultures were obtained from the American Type Culture Collection (ATCC) and maintained in Locke’s medium egg slants at 35.6°C in an anaerobic chamber.

Protein extraction and western blot:

*Blastocystis* cells were pelleted at 800 × g for 10 min and then washed twice in 1× Locke’s solution before suspended in 0.15 M NaCl. Cells were disrupted using ultrasonication (30s at 30s intervals, 6 cycles) (17, 18) on ice. Following disruption, protease inhibitor cocktail (Sigma; 10 μl) was added to the sample, which was then centrifuged at 10000 × g for 10 min at 4°C. Afterwards the supernatant was collected, and 10 μl of ice-cold nuclease buffer (20 mM Tris-HCl, pH 8.8 and 2 mM CaCl₂) along
with 10 μl of protease inhibitor cocktail were added. 30 μl of DNAase/RNAase mix (50 mM MgCl₂ and Tris-HCl 0.5 M, pH 7.0) was added and incubated on ice for 3 min. Subsequently, 10 μl of 3% SDS:10% mercaptoethanol was added and the mix was passed through a fine syringe. Samples were then stored at -20°C in NuPAGE LDS sample buffer along with 10 × sample reducing agent (Invitrogen). Depending on the amount of protein, 5 – 20 μl of the supernatant (~10%) was analyzed using a polyacrylamide mini gel.

Different cell fractions were isolated following procedures previously described (18, 19). Blastocystis cells (20 well-grown tubes) were harvested by centrifugation at 1200 × g for 10 min at 4°C. Cells were resuspended in Locke’s solution (pH 7.4) and pelleted again with the same speed and duration. Cells were then broken with 40 strokes in a 10 ml Potter-Elvehjem tissue homogenizer at 4°C in isotonic buffer (200 mM sucrose, pH 7.2, 30 mM phosphate, 15 mM β-mercaptoethanol, 30 mM NaCl, 0.6 mM CaCl₂ and 0.6 mM KCl). Broken cells were then diluted with isotonic buffer and then centrifuged at 700 × g for 10 min using a Sorvall RC-2B centrifuge to remove unbroken cells. The supernatant was collected and centrifuged at 5000 × g for 20 min to pellet the Large Granular Fraction (LGF) -where MROs are found – see 24, 25). The LGF was resuspended (washed) in isotonic buffer and pelleted as described above. Finally, all fractions were stored at -20°C in NuPAGE LDS sample buffer along with 10 × sample reducing agent (Invitrogen). Depending on the amount of protein, 5 – 20 μl of the supernatant was analyzed using a polyacrylamide mini gel.

For western blot analysis, proteins were transferred to Immuno-Blot polyvinylidene difluoride (PVDF) membranes (BioRad), visualized by Ponceau staining.
and then were blocked for an hour with 5% skimmed milk in TBS (10 mM Tris, 0.2 M NaCl, 0.2% BSA; pH 8.1) – Tween 0.1%. Membranes were washed with 0.5% skimmed milk in TBS – Tween 0.1% three times for 10 min. The primary antibodies (see below) were diluted in 1% skimmed milk in TBS– Tween 0.1% [anti-

Saccharomyces Nbp35 (6), 1:500; anti-

Saccharomyces Nar1 (8), 1:300; anti-

Saccharomyces Tah18 (10), 1:200] and applied to the membrane overnight at 4°C. Membranes were washed as before and incubated with the secondary antibody diluted in 1% skimmed milk in TBS – Tween 0.1%. Membranes were washed in TBS three times for 10 min, incubated with Enhanced ChemiLuminescence (ECL) reagent (GE Healthcare) and fluorescence was monitored by autoradiography.

Immuno-localization of CIA components in Blastocystis:

Blastocystis cells were fixed with 4% paraformaldehyde at 37°C for 15 min followed by 3 × 10 min washes with 1× TBS. Fixed cells were permeabilized with iced-cold acetone and washed 3 × 10 min with 1× TBS – 0.1% Triton. Fixed cells were incubated for 30 min with a blocking solution of 5% skimmed milk powder 1× TBS – 0.1% Triton solution (w/v). They were then rinsed with 0.5% milk/TBS – 0.1% Triton solution for 30 min. The cells were then incubated with a dilution of the antiserum (anti-

SufCB, 1:200; anti-Nbp35, 1:100; anti-Nar1, 1:100; anti-Tah18, 1:50) in 1% milk/ TBS – 0.1% Triton solution for overnight at 4°C. Three different dilutions of each antiserum were tested to determine optimal conditions. After three rinses in 1% milk/ TBS – 0.1% Triton, the slides were incubated with fluorescent dyes-labelled (Alexa 488 green and Alexa 594 red) secondary antibodies at a dilution of 1:200. Finally, the slides were
incubated with DRAQ5 (Cell Signaling Technology) stain for 5 min (dilution 1:1000) and washed 3 x 1min with 1 x TBS. Cover slips were mounted with anti-fade mounting medium (Vectashield) and observed under a laser scanning confocal microscope (Zeiss LSM 510 Meta) using a 100 × oil immersion-lens.

For the immuno-localization experiments using transmission electron microscopy, *Blastocystis* NandII cells were fixed and manipulated using the protocols described previously (17). The prepared grids were incubated in sodium borohydride (1 mg/ml) for 10 min followed by 10 min incubation in glycine buffer (30 mM of glycine in 0.1 M borate buffer, pH 9.6). Grids were then incubated in blocking solution (TBS buffer with 1% skim milk and 1% BSA) for 45 min followed by a quick rinse with TBS buffer. The grids were then incubated overnight with a dilution of the antiserum (anti-Nbp35, 1:20; anti-Nar1, 1:15; anti-Tah18, 1:5) in TBS buffer at 4°C. Three different dilutions of each antiserum were tested to determine optimal conditions. Grids were then rinsed three times (15 min each) in washing buffer (10 mM Tris, 0.3 M NaCl, 0.1% BSA; pH 8.1) followed by an hour incubation with secondary anti-rabbit antibody conjugated with 10 nm gold particles (Sigma) diluted in TBS buffer. Grids were then rinsed three times (15 min each) in washing buffer, incubated for 15 min in 2.5% glutaraldehyde, rinsed three times (3 min each) in distilled water and stained with 2% uranyl acetate and lead citrate. Samples were viewed with a JEOL JEM 1230 transmission electron microscope to determine quality before proceeding with immuno-labelling.

Transcriptomic data obtained by 454 pyrosequencing cDNA from *Blastocystis* sp. NandII.
Total RNA from *Blastocystis* cells was isolated using Trizol according to the manufacturer’s specifications with the following modification: following separation of the organic phase, the supernatant was collected and underwent a second round of Trizol extraction. Complementary DNA (cDNA) was constructed by Vertis Biotechnologies AG (Germany) and 454 pyrosequencing was performed by Genome Quebec in a 4SLX Titanium Platform. The Mira assembly program version 3.0 (20) was used to assemble the reads into contigs. A BLAST database was created using the resulting Mira contigs. The CIA proteins of *Blastocystis hominis* S7 strain (21) and *Saccharomyces cerevisiae* were used as seeds to perform a local tblastn search against the *Blastocystis* NandII contig database and extract the corresponding homologues. The sequences of *Blastocystis* sp. NandII transcripts are deposited in GenBank with the accession numbers KF438229 to KF438233 or Table S1.

Database searches and dataset assembly

Prokaryotic and eukaryotic homologues of the CIA system protein sequences were retrieved from GenBank using blastp and tblastn searches with *S. cerevisiae* and *Homo sapiens* as the initial seed query sequences (22). Identical and highly similar sequences were removed. Additional databases that were searched for eukaryotic homologues include the Joint Genome Institute (JGI, http://www.jgi.doe.gov), the Broad Institute (http://www.broadinstitute.org), the *Cyanidioschyzon merolae* Genome Project (http://merolae.biol.s.u-tokyo.ac.jp/), GiardiaDB (http://giardiadb.org), AmoebaDB (http://amoebadb.org), NemaGENETAG (http://elegans.imbb.forth.gr/nemagenetag) and the Wellcome Trust Sanger Institute (http://www.sanger.ac.uk). For a given protein,
absence of a homologue in a taxon was further verified by employing tblastn on the corresponding genomes using multiple queries as seeds. For organisms where the tblastn search did not result in any hits to a particular protein family, a profile HMM of the alignment was built and used to search the predicted proteomes using the hmmbuild and hmmsearch programs of the HMMER package, version 3.0 (http://hmmer.janelia.org).

Protein Domain identification

Protein domains were identified using SMART (http://smart.embl-heidelberg.de/) (23) and by performing HMMER searches against the PFAM 26.0 database (http://pfam.sanger.ac.uk) (24).

Multiple alignment and Phylogenetic analysis

Protein sequences were aligned with MAFFT v6.903b (25). Furthermore, the alignments were inspected by eye to detect cases of obviously misaligned regions. Subsequently, the alignments were masked to remove regions of ambiguous alignment using the Block Mapping and Gathering with Entropy (BMGE) version 1.1 using default parameters (26).

After trimming, the final alignments contained 59 taxa and 225 sites for the Cia1 protein, 78 taxa and 96 sites for Dre2, 87 taxa and 55 sites for MMS19, 81 taxa and 233 sites for Nar1, 95 taxa and 191 sites for the Nbp35/Cfd1 proteins, 110 taxa and 157 sites for Tah18, and 60 taxa and 98 sites for Cia2. Phylogenetic trees were constructed for each individual protein of the CIA machinery. Maximum likelihood (ML) trees were computed using RAxML version 7.2.8 (27) and the Le and Gascuel (LG) amino acid
substitution model (28). To account for rate heterogeneity across amino acid sites the gamma option was also implemented. For each protein dataset, bootstrap support was assessed from 100 bootstrap replicates that were subsequently mapped onto the best scoring ML tree. Phylogenies for the CIA components that display interesting patterns (e.g. Nbp35/Cfd1 and Tah18) are shown as main figures, whereas those with less notable evolutionary histories are included as supplemental figures.

**Topology testing**

In the case of Tah18, we tested whether various phylogenetic hypotheses could be significantly rejected by the data. We tested two alternate phylogenetic placements of putative Blastocystis spp. Tah18 sequences: (1) grouping with the four Blastocystis sequences of the “PNO clade” (sequences indicated by a white square, Figure 2) (case 1), or (2) grouping as monophyletic with other stramenopile Tah18 orthologues (sequences indicated by a black square, Figure 2) (case 2). We used the approximately unbiased (AU) test implemented in CONSEL (29). Since this test requires a large sample of “good” trees in addition to the test topologies to accurately estimate P values (30), 500 bootstrap trees were included in the analyses. The maximum-likelihood tree given a specific constraint (i.e. corresponding to a phylogenetic hypothesis) was obtained using the −g RAxML option with all other parameters set as previously described. Case (1) could not be rejected (P value = 0.228, likelihood = -51718.8), whereas case (2) was rejected (P value = 0.036, likelihood = -51781.3).

**Results and Discussion**
1. Characterization/localization of components of Blastocystis CIA machinery

In a previous study, we showed that three different Fe/S cluster biosynthetic machineries were likely present in the anaerobic parasite *Blastocystis* sp. (17). We localized and functionally characterized the mitochondrial Fe/S cluster (ISC) machinery and the cytosolic SUF machinery, but only limited data were available regarding the presence of the CIA machinery. Here, we report homologues of six CIA proteins (Nbp35, Nar1, Cia1, Cia2, Tah18, and MMS19) in our transcriptomic data. The homology of three of these proteins was further confirmed by the identification of conserved features/motifs shared with better-characterized eukaryotic homologues (e.g. human, yeast) (Figures S2 – S4).

We employed various methods to assess the localization of CIA components in *Blastocystis*. Using heterologous antibodies raised against the yeast proteins Nbp35 (6), Nar1 (8) and Tah18 (10), we demonstrated their specificity against the *Blastocystis* homologues in western blots of total protein extracts (Figure S5). To further confirm the localization of these proteins, we performed subcellular fractionations employing a two-step centrifugation method followed by a western blot assay. We used Nbp35, Nar1 and Tah18 antibodies against the corresponding proteins obtained from the fractions. In all cases, the antibodies were localized exclusively in the cytosolic fraction of the cell extracts (Figure S5).

Immuno-fluorescence microscopy demonstrated that *Blastocystis* Nbp35, Nar1 and Tah18 proteins localized in the cytosol of the cell; these proteins co-localized with the cytosolic protein SufCB as well (17) (Figure 1), and they did not co-localize with
Mitotracker, a dye that labels the MRO of the parasite (Figure S6). Moreover, immuno-gold transmission electron microscopy using the same antibodies against the CIA proteins revealed an abundance of gold particles in the cytosol and their virtual absence from the MRO, vacuole and nucleus (Figure 1d,e & Figures S7 – S9). A similar immuno-gold labelling pattern was demonstrated for the fused cytosolic SufCB protein of Blastocystis sp. (Figure 1e) (17).

This work, along with a previous study (17) demonstrated that key components of the ISC and CIA systems are encoded by genes in Blastocystis sp. NandII genome, and that they localize in this organism’s MROs and cytosol, respectively. However, Blastocystis sp. also has a cytosolic SUF system (17) - a machinery never before described in the eukaryotic cytoplasmic compartment (1). With these two cytosolic Fe/S cluster biogenesis systems, Blastocystis sp. resembles Entamoeba, and Mastigamoeba, both of which possess novel cytosolic NIF systems (16, 31) in addition to CIA components (presumed to be cytosolic in these organisms) (Table 1). For these organisms it is unclear how Fe/S cluster biogenesis is ‘partitioned’ between the two cytosolic systems and whether or not they function independently or in a coordinated fashion. Currently, it is not possible to genetically manipulate some of these organisms, making it difficult to study these systems in more detail.

2. Distribution and evolutionary history of the CIA machinery homologues in other microbial eukaryotes

Although most of the eight CIA system components present in yeast are essential for its viability (9), our survey of Blastocystis spp. data (21, 32) allowed the identification
of only five (potentially six, see discussion below about Tah18) of these proteins. Consequently, it is unclear how this pathway functions in Blastocystis and whether other eukaryotes also lack some of the components. To address this question, we carried out a phylogenomic analysis of each individual component.

a. DRE2

Dre2 is an essential gene required for the assembly of virtually all cytosolic and nuclear target Fe/S proteins in yeast and localizes mostly in the cytosol and partially in the mitochondrial inter-membrane space (IMS) (33). The latter seemed to suggest that Dre2 acts before the rest of the CIA components; an observation further supported by additional experimental evidence (10). Interestingly, Dre2 itself is a Fe/S cluster containing protein, whose maturation seems to occur independently of the CIA system (10). This suggests a putative functional dependence of Dre2 (and thus the CIA system) on another Fe/S cluster assembly pathway.

Dre2 has a highly conserved C-terminus corresponding to a eukaryote-specific CIAPIN1 domain that contains two Fe/S cluster-binding motifs: CX$_2$CXC and CX$_2$CX$_2$CX$_2$C while the N-terminus is more divergent (Figure S10). In some chlorophytes/embryophytes and metazoans, the N-terminus contains a methyltransferase domain (Figure S10) supporting the recently proposed hypothesis that this part of Dre2 is homologous to the S-adenosylmethionine (SAM) methyltransferase protein family (34). Although the interaction of Dre2 and Tah18 has been shown to be indispensable for yeast survival, a few parasitic taxa (including Blastocystis, Entamoeba, Giardia and Trichomonas) appear to be lacking a Dre2 homologue (Table 1). Apart from these
absences (presumably because of secondary loss), the phylogenetic analysis was consistent with vertical inheritance of Dre2 within eukaryotes from LECA (Figure S11). The taxonomic distribution of the SAM domain seems to suggest that Dre2 in LECA had a SAM – CIAPIN1 domain composition. It is possible that eukaryotic sequences in which a SAM domain was not detected have evolved beyond recognition since most of them have a N-terminal sequence that aligns well with (and thus seems homologous to) the SAM domain of green plants/metazoa sequences.

b. TAH18

Tah18 is a reductase with binding motifs for flavin mononucleotide (FMN), flavin adenine dinucleotide (FAD) and nicotinamide adenine dinucleotide (NAD) cofactors (Figure 2). Tah18 forms a complex with Dre2 that functions as an electron transfer chain by transferring two reducing equivalents from NADPH to Dre2 (10, 34, 35). Experimental evidence suggest that the Tah18/Dre2 complex disassociates under oxidative stress conditions and as a result Tah18 re-targets to mitochondria in yeast (35), leading to the speculation that Tah18 promotes apoptosis under oxygen stress (34). Phylogenetic analyses show that Tah18 belongs to a large multi-protein family, most members of which share the canonical FMN-FAD-NAD domain organization and are involved in redox reactions. The pyruvate-NADPH oxidoreductase (PNO) and NADPH cytochrome P450 reductase (CPR) sub-families are the closest homologues of Tah18 (Figure 2). Our analyses show that the putative Blastocystis Tah18 sequences (Figure 2) do not group with the rest of the Tah18 stramenopile homologues (black square, Figure 2), but instead branch with the PNO sequences. While most eukaryotes possess an
orthologue of Tah18, the protein seems to be absent in Entamoeba (Table 1) and potentially in Giardia whose homologues are unclearly related to Tah18 or CPR (Figure 2). There seems to be a similar phylogenetic profile of presence/absence of canonical Tah18 and Dre2 (Table 1) that is consistent with them forming a functional ‘module’ in the CIA system.

c. NBP35/CFD1

Nbp35 and Cfd1 are homologous Mrp-like proteins and members of the P-loop-NTPase family, which also includes the more distantly related mitochondrial Fe/S apoprotein IND1 (36). Both Nbp35 and Cfd1 proteins possess a Walker A box for ATP hydrolysis and two conserved C-terminal cysteines that serve as the binding sites of Fe/S clusters (11) (Figure S4). Nbp35 has an N-terminal ferredoxin-like domain that Cfd1 lacks (7) (Figure S4). While Nbp35 is universally distributed among eukaryotes, Cfd1 has a by far more patchy distribution: it is found only in opisthokonts, amoebozoa, discicristate excavates and the cryptophyte Guillardia theta (Table 1). These results are in contrast to a recent publication (37) claiming the universal presence of Cfd1 protein among eukaryotes (with the exception of the Plantae clade). As the authors of the aforementioned paper used a reciprocal blast approach and not a phylogenetic analysis, we suspect their approach might have led to a misidentification of the two paralogous families.

In yeast, Cfd1 and Nbp35 form a hetero-tetramer that serves as the scaffold of the cytosolic Fe/S cluster biogenesis (11, 38) and absence of either of these two proteins in yeast results in cell death (11). However, in Viridiplantae, which lack Cfd1, it has been
shown that Nbp35 forms a homo-dimer that delivers Fe/S clusters to apo-proteins (39).
This is possibly how Nbp35 functions in Blastocystis sp. and in other organisms that lack the Cfd1 orthologue.

Our phylogenetic analysis shows that Nbp35 and Cfd1 proteins are paralogues and the duplication event that gave rise to them most likely preceded eukaryotic diversification, and thus that these two paralogues existed in LECA (Figure 3a). In the Nbp35 subtree, stramenopiles (including Blastocystis), cluster with Viridiplantae and Emiliania huxleyi, albeit with weak support. However, most members of this group share an insertion in the C-terminal region of the protein, ranging in length from 2-45 amino acids and fairly conserved in sequence (Figure 3b). A few taxa within this group appear to have lost part or nearly all of the insertion secondarily (e.g., Volvox and Chlamydomonas). The observed branching pattern of Nbp35 is at odds with current conceptions of eukaryotic phylogeny (40) that place stramenopiles with alveolates and Rhizaria. A potential explanation is that the gene has been laterally transferred between stramenopiles, Viridiplantae and Emiliania, although, if this is correct, the numbers and directions of transfer events could not be determined from the poorly resolved phylogeny. However, the lack of the first of the two conserved central cysteine residues of the C-terminal motif [CPXC] in Blastocystis and Viridiplantae Nbp35 sequences supports the possibility of a recent transfer event involving these two lineages (Figure S4).

Finally, phylogenetic reconstructions also revealed that bacterial ApbC proteins are the closest prokaryotic homologues of Nbp35/Cfd1. Interestingly, experimental evidence suggests that the ApbC of Salmonella enterica functions in Fe/S cluster biogenesis pathways (41) and homologues from several archaea functionally complement
ApbC mutants of *S. enterica* (42). In addition, biochemical analyses of bacterial ApbC homologues showed that they bind and transfer Fe/S clusters to a yeast apo-protein *in vitro* (40). Collectively, these results suggest that ApbC/Nbp35/Cfd1 homologues were involved in Fe/S cluster biosynthesis prior to the diversification of the three domains of life.

d. NAR1

Nar1 belongs to the hydrogenase-related protein family, however it is not involved in hydrogen production (8). Even though Nar1 is predominantly localized in the cytosol, it also appears to be partly membrane-associated in yeast (5). The protein acts late in the pathway where it facilitates transfer of Fe/S clusters to target proteins. All Nar1 homologues have a highly conserved N-terminus ferredoxin-like domain CX₅CX₂CX₅C and a CX₅CX₅CX₃C C-terminal motif, the so-called “H-cluster”, that has a likely role in Fe/S cluster coordination (8) ([Figure S3](#)). This protein has a wide distribution among eukaryotes ([Table 1](#)) and phylogenetic analysis shows that all Nar1 homologues cluster together with strong bootstrap support ([Figure S12](#)), and represent paralogues of the hydrogenase-like proteins found in anaerobic eukaryotes. In the Nar1 subtree, *Blastocystis* branches within the alveolates and not with other stramenopiles, however, this relationship is not strongly supported ([Figure S12](#)).

e. CIA1

Cia1 is a W40-repeat protein responsible for the last step of the Fe/S cluster assembly process. WD40 proteins are usually known to coordinate the assembly of
multiprotein complexes by functioning as a docking site for other proteins. As such, Cia1 interacts with Nar1 and facilitates the transfer of Fe/S clusters to target proteins (4, 43). In yeast, depletion of Cia1 resulted in the loss of activities of cytosolic and nuclear Fe/S-cluster-bearing enzymes (4). The essential role of this protein in the CIA system is further supported by its universal presence in all eukaryotes (Table 1) and by the strict conservation of the 7-bladed beta-propeller structure (43) across all homologues (Figure S2). The Cia1 phylogeny is poorly resolved but does not appear to indicate any lateral gene transfer between eukaryotic lineages (Figure S13).

f. MMS19

MMS19 is the most recently discovered component of the CIA machinery (9, 12). This chaperone forms a complex with Nar1 and Cia1 components. This protein is patchily distributed among eukaryotes, suggesting it has been lost numerous times along various lineages (Table 1). However, a high degree of divergence was observed between homologous sequences. We thus cannot exclude the possibility that other eukaryotic homologues have diverged beyond recognition, and that MMS19 is universally present, but unrecognizable at the sequence level in many eukaryotes (see Materials and Methods). In particular, the large degree of sequence divergence observed specifically in homologues from Blastocystis spp., alveolates and excavates suggests that the protein may have a different function in these lineages. Although the phylogeny is poorly resolved, the presence of MMS19 homologues in most organisms of all major groups, as well as the absence of obvious LGT between eukaryotic lineages suggests that this component already existed in LECA (Figure S14). As no homologues of this protein
have been identified in prokaryotes, this component likely evolved after the divergence of
eukaryotes from their prokaryotic ancestor, although here again the rapid rate of
evolution of this gene might preclude the identification of prokaryotic homologues.

g. CIA2

The Cia2 protein family has been recently identified as acting at the end of the
CIA pathway (9, 12). In animals, two paralogous members of this family are present,
Cia2A and Cia2B. Cia2B associates with Cia1 and MMS19 to form the so-called “CIA-
targeting complex” which binds directly to the coordinating cysteines of Fe/S clusters of
most nuclear/cytosolic apo-proteins that are matured by the complex (44). This suggests
that Cia2B has a role in connecting donor and acceptor Fe/S proteins (Stehling et al.,
2012). In contrast, Cia2A seems to be specifically involved in cellular iron homeostasis
through the maturation of iron regulatory protein 1 and the stabilization of iron regulatory
protein 2 (45).

Phylogenetic analyses showed that virtually all eukaryotes, including Blastocystis,
possess at least one homologue of the Cia2 family (Figure S15). The phylogenetic tree is
poorly supported overall but seems to show a basal split between two monophyletic
groups. The first one includes the human Cia2B homologue and contains sequences from
all eukaryotic lineages (with very few exceptions), suggesting that this protein was
already present in LECA. The second group, which includes the human Cia2A, displays a
much narrower taxonomic distribution with orthologues present only in some lineages of
animals, Amoebozoa, ciliates, Trypanosoma and red algae. Given the very weak support
for this clade and its sparse taxonomic distribution, the existence of a second paralogue in
these lineages can be explained by three different hypotheses: (i) the two paralogues were already present in LECA and have been independently lost many times along the eukaryotic tree; (ii) the duplication event occurred in one of these eukaryotic lineages and the gene has subsequently been laterally transferred to the other lineages represented in the “Cia2A clade”; (iii) several duplication events occurred independently and the second copy of each lineage artefactually groups with the others in the tree.

Proteins of the Cia2 families are composed of a unique DUF59 domain and represent the only protein carrying this domain in eukaryotic organisms (with the exception of the HCF101 protein, discussed below). Cia2 possess homologues in both archaea and bacteria. Surprisingly, the closest prokaryotic homologues can be found among bacteria and correspond to a DUF59 containing protein called SufT. The function of SufT is unknown, but members of this family are commonly found in operons for the of iron/sulfur cluster biosynthesis SUF system. This is particularly noteworthy since, to our knowledge, it is the first reported evidence for an evolutionary link between the SUF and the CIA pathways.

Other bacterial and archaeal homologues are found in operons associated with phenylacetic acid degradation pathways and are represented by the PaaD protein (46, 47). PaaD is essential in vivo, and is thought to be involved in the maturation of the other Fe-S cluster-baring members of the operon (48). Finally, the only other eukaryotic protein carrying a DUF59 domain is the plastid HCF101 protein (high chlorophyll fluorescence 101). HCF101 is essential for the accumulation of two [4Fe-4S] containing chloroplast proteins (49). Interestingly, this protein is a fusion of a DUF59 domain, a member of the...
3. On the origin and evolution of the CIA machinery

Our phylogenetic analyses suggest that all of the eight known components of the CIA system originated before the diversification of extant eukaryotic lineages and that LECA had a complete and functional CIA pathway, likely resembling the one found in most animals and fungi. While some components have been universally conserved during eukaryotic evolution (i.e. Nbp35, Nar1, Cia1 and Cia2), a few of the CIA proteins, such as Cfd1 and MMS19, display a very patchy distribution (Table 1). According to the corresponding phylogenetic trees, this distribution is more likely to be explained by multiple loss events than by lateral gene transfers (see above). This presence/absence pattern also has no obvious correlation with organismal lifestyle. However, we can speculate that it may correlate with the complement of Fe/S cluster containing proteins and the type of Fe/S clusters needed and assembled through the CIA pathway in these organisms. By contrast, the presence/absence pattern of Tah18 and Dre2 seems to have a clearer correlation with organismal lifestyle. With the exception of Cryptosporidium species, all of the protists that live in hypoxic environments and contain mitochondrion-related organelles (i.e., Entamoeba, Mastigamoeba, Trichomonas, Giardia and Blastocystis) lack Dre2 and a canonical Tah18 orthologue (Table 1). This is unexpected because the NADPH-Tah18-Dre2 electron transfer chain was shown (in yeast) to be required early in the biosynthetic pathway for the incorporation of Fe-S cluster into all proteins carrying stable Fe/S clusters (i.e., not scaffold proteins), including the CIA Fe/S
proteins Nbp35 and Nar1 (10). Consequently, it seems like another protein (or pair of proteins) carries out the Tah18-Dre2 functions when the latter are absent. For example, the *bona fide* Tah18 orthologue, which seems to have been lost in some protists, could have been replaced by another member of the superfamily showing the same FMN-FAD-NAD domain organization (e.g., as found in *Blastocystis* and *Giardia*). In contrast, Dre2 could have been replaced by a non-homologous protein, since no paralogue of this component can be found in the aforementioned protists. This is consistent with the fact that eukaryotic difflavin reductases can be coupled with various electron-accepting proteins (e.g., cytochrome P450, methionine synthase, nitric oxide synthase 1, etc.). It is thus possible that Tah18 is transferring electrons to an analogue of Dre2 (i.e., a non-homologous, but functionally similar protein). It is interesting to correlate the loss of Dre2 with the presence in some of these anaerobes of recently acquired unique cytosolic Fe/S biosynthetic proteins; NifS/NifU in the cases of *Entamoeba* and *Mastigamoeba* and a SufCB fusion protein in the case of *Blastocystis*. It is possible that the NIF and SUF components have a Dre2-like role in the electron transfer chain and therefore interface with the remainder of CIA system in these organisms.

In addition, the Tah18-Dre2 complex putatively interfaces with the mitochondrial ISC system in yeast, and is implicated in processing its product, compound X (1). Compound X is thought to be a sulfur-containing compound of as-yet unknown nature that serves as the sulfur donor for the CIA pathway. It is produced by the mitochondrion-localized ISC assembly machinery and exported from the mitochondrial matrix to the cytosol. In *Giardia* mitosomes, the only proteins carrying Fe/S clusters are the components of the ISC Fe/S cluster assembly machinery itself. Consequently, one of the
main roles of mitosomes could be to export compound X to other cellular compartments (50). The potentially essential role of compound X for the proper functioning of the CIA pathway is another line of evidence that a Tah18-Dre2 analogue must exist in this organism.

The various patterns of presence/absence of components across eukaryotes suggest that while there are a number of ‘core’ essential components of the CIA system (e.g. Nbp35, Cia1, Cia2 and Nar1); the remaining set of proteins appears to be more evolutionarily plastic. Since this pathway has only been extensively studied in a few opisthokont lineages, further investigations focusing on microbial eukaryotes will likely provide us with insights into other novel components involved in the CIA system that may functionally replace some of the less conserved known CIA proteins. In any case, a broader study of the CIA system and its target apo-proteins across eukaryotic diversity should illuminate both universal ancestral and lineage-specific functions of this essential pathway.

Acknowledgments

This work, EG and DG were supported by an operating grant MOP-62809 from the Canadian Institutes of Health Research awarded to AJR. ADT was supported by a Marie Curie international outgoing fellowship and LE was supported by a postdoctoral fellowship from the Tula Foundation. AJR acknowledges support from the Canadian Institute for Advanced Research Program in Integrated Microbial Biodiversity. We would like to thank Prof. Roland Lill for providing us with the antibodies against Nbp35, Nar1 and Tah18 proteins. We would also like to thank Mary Ann Trevors for her
assistance on the Electron Microscopy and Stephen Whitefield for his guidance on using the Cellular & Molecular Digital Imaging facility. Finally, we are thankful to the three anonymous reviewers for valuable comments.

Figure legends:

Figure 1: Immuno-localization of the Nbp35, Nar1 and Tah18 in Blastocystis sp.

a. Cellular localization of the Nbp35 and SufCB in Blastocystis cells. i. Rabbit anti-yeast Nbp35 antibody (1:200) detects Blastocystis Nbp35 protein. ii. Localization of Blastocystis SufCB antibody. iii. DRAQ5 staining the nuclei of Blastocystis. iv. Differential interference contrast (DIC) image of the cells used for immunofluorescence. v. Overlapping of the previous images showing the localization pattern of these proteins.

b. Cellular localization of components of the Nar1 and SufCB in Blastocystis cells. i. Rabbit anti-yeast Nar1 antibody (1:100) detects Blastocystis Nbp35 protein. ii. Localization of Blastocystis SufCB antibody. iii. DRAQ5 staining the nuclei of Blastocystis. iv. Differential interference contrast (DIC) image of the cells used for immunofluorescence. v. Overlapping of the previous images showing the localization pattern of these proteins.

c. Cellular localization of components of the Tah18 and SufCB in Blastocystis cells. i. Rabbit anti-yeast Tah18 antibody (1:100) detects the Blastocystis Nbp35 protein. ii. Localization of Blastocystis SufCB antibody. iii. DRAQ5 staining the nuclei of Blastocystis. iv. Differential interference contrast (DIC) image of the cells used for immunofluorescence. v. Overlapping of the previous images showing the localization pattern of these proteins.
pattern of these proteins. d. Immuno-gold microscopy image demonstrating the localization pattern of the anti-Nbp35 antibody. A higher magnification image is shown in **Figure S7**. e. Densities of immuno-gold particle labelling in different compartments of *Blastocystis* cells suggest that Nbp35, Nar1, Tah18 along with the previously published SufCB [20] proteins are mainly localized in the cytosol of the parasite.

Figure 2:

**Tah18 phylogeny:** A maximum likelihood phylogeny of the cytosolic Fe/S cluster assembly protein Tah18 and its two closest homologues NADPH cytochrome P450 reductase and pyruvate ferredoxin oxidoreductase. Numerical values on the branches represent statistical support in the form of bootstrap values and the scale bar indicates the branch-length in terms of substitutions/site. Only bootstrap support values greater than 50 are shown. Thick branches highlight clades whose members have common protein domain structure. Black and white squares indicate the two alternative phylogenetic branching positions of the putative *Blastocystis* spp. Tah18 orthologues considered in AU-tests (see Materials and Methods). Letters underneath the branches indicate the number and order of domains as those are presented at the bottom of the figure. Protein domains are individually depicted in taxa that possess numbers of domains different than most members of their clade.

Figure 3:

**Nbp35/Cfd1 phylogeny and alignment features:** a. A maximum likelihood phylogeny of the cytosolic Fe/S cluster assembly proteins Nbp35 and Cfd1 and bacterial
homologues. Statistical support values shown are as described for Figure 2. The Nbp35 and Cfd1 groups are indicated and correspond with the presence versus absence of the ferredoxin domain at their N-termini respectively. Members of the clade marked with the black dot share a variable-length insertion in the carboxyl terminus of the Nbp35 protein. b. A portion of the amino acid sequence alignment depicting the taxa sharing the insertion in the Nbp35 protein.

Table 1: Distribution of homologues of the CIA machinery among publicly available genomes or expressed sequence tag data from eukaryote taxa. The presence (+) and absence (-) of homologues in complete genomes are indicated while “?” illustrates the absence from incomplete genome or transcriptomic data. In addition, “*+” indicates the presence of a homologous sequence in which at least one domain is missing; “%” corresponds to tah18 homologues of which the phylogenetic placement in the tree unclear (see main text); Finally, “#” indicates the presence of a potential but highly divergent homologue.

References:


<table>
<thead>
<tr>
<th>Supergroup</th>
<th>Species</th>
<th>dre2</th>
<th>tah18</th>
<th>nph35</th>
<th>cfd1</th>
<th>nar1</th>
<th>cia1</th>
<th>cia2b</th>
<th>mms19</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>Apicomplexa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>