

School of Doctoral Studies in Biological Sciences
University of South Bohemia in České Budějovice
Faculty of Science

Evolution of the Heme Biosynthetic Pathway in
Eukaryotic Phototrophs

Ph.D. Thesis

Mgr. Jaromír Cihlář

Supervisor: Prof. Ing. Miroslav Oborník, Ph.D.
Biology Centre CAS v.v.i., Institute of Parasitology

České Budějovice 2018

This thesis should be cited as: Cihlář J., 2018. Evolution of the Heme Biosynthetic Pathway in Eukaryotic Phototrophs. Ph.D. Thesis Series, University of South Bohemia, Faculty of Science, School of Doctoral Studies in Biological Sciences, České Budějovice, Czech Republic.

Annotation

This thesis is devoted to the evolution of the heme biosynthetic pathway in eukaryotic phototrophs with particular emphasis on algae possessing secondary and tertiary red and green derived plastids. Based on molecular biology and bioinformatics approaches it explores the diversity and similarities in heme biosynthesis among different algae. The core study of this thesis describes the heme biosynthesis in *Bigeloviella natans* and *Guillardia theta*, algae containing a remnant endosymbiont nucleus within their plastids, in dinoflagellates containing tertiary endosymbionts derived from diatoms – called dinotoms, and in *Lepidodinium chlorophorum*, a dinoflagellate containing a secondary green plastid. The thesis further focusses on new insights in the heme biosynthetic pathway and general origin of the genes in chromerids the group of free-living algae closely related to apicomplexan parasites.

Declaration [in Czech]

Prohlašuji, že svoji disertační práci jsem vypracoval samostatně pouze s použitím pramenů a literatury uvedených v seznamu citované literatury. Prohlašuji, že v souladu s § 47b zákona č. 111/1998 Sb. v platném znění souhlasím se zveřejněním své disertační práce, a to v nezkrácené podobě elektronickou cestou ve veřejně přístupné části databáze STAG provozované Jihočeskou univerzitou v Českých Budějovicích na jejích internetových stránkách, a to se zachováním mého autorského práva k odevzdanému textu této kvalifikační práce. Souhlasím dále s tím, aby toutéž elektronickou cestou byly v souladu s uvedeným ustanovením zákona č. 111/1998 Sb. zveřejněny posudky školitele a oponentů práce i záznam o průběhu a výsledku obhajoby kvalifikační práce. Rovněž souhlasím s porovnáním textu mé kvalifikační práce s databází kvalifikačních prací Theses.cz provozovanou Národním registrem vysokoškolských kvalifikačních prací a systémem na odhalování plagiátů.

České Budějovice, 19.01.2018

.....
Jaromír Cihlář

This thesis originated from a partnership of Faculty of Science, University of South Bohemia, and Institute of Parasitology, Biology Centre CAS v.v.i., supporting doctoral studies in the Molecular biology and genetics study program.



Přírodovědecká
fakulta
Faculty
of Science

Jihočeská univerzita
v Českých Budějovicích
University of South Bohemia
in České Budějovice

Financial support

This work was financially supported by following grants and supporting institutions:

- Grant Agency of the Czech Republic project P501-12-G055
- Grant Agency of the Czech Republic project 15-17643S
- Grant Agency of the Czech Republic project P506-12-1522

Acknowledgements:

I would like to express my immense gratitude to my supervisor Miroslav Oborník for giving me the opportunity to work on such an interesting topic. I sincerely appreciate his enthusiasm, support, and last but not least his patient guidance. I would also like to thank all the members of Laboratory of Evolutionary Protistology for their inspiring advices, help, and especially for being always great colleagues. My special thanks belong to Zoltán Füssy for excellent collaboration and for proofreading.

Furthermore, I would like to express many thanks to my beloved Olinka for her love, patience and support.

I am immensely thankful to my family: my mother, father and my grandmother for their constant faith in me and for the support of all kinds.

List of papers and author's contribution

The thesis is structured based on the following papers:

Cihlář J, Füssy Z, Horák A, Oborník M. (2016) Evolution of the Tetrapyrrole Biosynthetic Pathway in Secondary Algae: Conservation, Redundancy and Replacement. *PLoS One* 11(11):e0166338.

J. Cihlář and Z. Füssy contributed equally to this work. Jaromír performed heme pathway in silico targeting predictions and phylogenetic analyses of B. natans and G. theta, and participated in production of visualizations, writing and reviewing the manuscript.

Woo YH, Ansari H, Otto TD, Klinger CM, Kolisko M, Michálek J, Saxena A, Shanmugam D, Tayyrov A, Veluchamy A, Ali S, Bernal A, del Campo J, **Cihlář J**, Flegontov P, Gornik SG, Hajdušková E, Horák A, Janouškovec J, Katris NJ, Mast FD, Miranda-Saavedra D, Mourier T, Naeem R, Nair M, Panigrahi AK, Rawlings ND, Padron-Regalado E, Ramaprasad A, Samad N, Tomčala A, Wilkes J, Neafsey DE, Doerig C, Bowler C, Keeling PJ, Roos DS, Dacks JB, Templeton TJ, Waller RF, Lukeš J, Oborník M, Pain A. (2015) **Chromerid genomes reveal the evolutionary path from photosynthetic algae to obligate intracellular parasites. *eLife* 4:e06974.**

J. Cihlář performed DNA and RNA extractions, library preparation and sequencing, and heme pathway in silico targeting predictions and phylogenetic analyses.

Burki F, Flegontov P, Oborník M, **Cihlář J**, Pain A, Lukeš J, Keeling PJ. (2012) **Re-evaluating the green versus red signal in eukaryotes with secondary plastid of red algal origin. *Genome Biology Evolution* 4(6):626-35.**

J. Cihlář participated in the identification of contamination by higher plant sequences, and with M. Oborník in the evaluation of phylogenies.

Miroslav Oborník, the supervisor of this Ph.D. thesis and co-author of all presented papers, approves the contribution of Jaromír Cihlář in these papers as described above.

.....
Prof. Ing. Miroslav Oborník, Ph.D.

Table of contents

Review	1
Evolution of the heme biosynthetic pathway in eukaryotic phototrophs	
1 Introduction	1
1.1 Importance of heme in living organisms	2
2 Biosynthesis of heme	4
2.1 Biosynthesis of heme in eukaryotes and the significance of the endosymbiotic gene transfer.....	5
2.2 Biosynthesis of heme in primary heterotrophs.....	6
2.3 Biosynthesis of heme in phototrophic eukaryotes.....	7
2.3.1 Primary phototrophs.....	7
2.3.2 Other phototrophs	13
2.3.2.1 Biosynthesis of heme in algae with secondary and tertiary plastids derived from the red lineage	15
2.3.2.2 Biosynthesis of heme in algae with secondary plastids derived from the green lineage.....	21
3 Conclusion	25
References	26
 Paper I	 39
Evolution of the Tetrapyrrole Biosynthetic Pathway in Secondary Algae: Conservation, Redundancy and Replacement	

Paper II.....	70
Chromerid genomes reveal the evolutionary path from photosynthetic algae to obligate intracellular parasites	
Paper III.....	112
Re-evaluating the green versus red signal in eukaryotes with secondary plastid of red algal origin	
Final conclusion.....	123
Curriculum vitae	126

Evolution of the heme biosynthetic pathway in eukaryotic phototrophs

Jaromír Cihlář

Faculty of science, University of South Bohemia; Institute of Parasitology, Biology centre ASCR, České Budějovice, Czech Republic.

Abstract:

Tetrapyrroles such as heme and chlorophyll are involved in energy consumption (oxidative phosphorylation) and fixation (photosynthesis) and therefore are crucial components for cell metabolism. Biosynthetic pathway responsible for tetrapyrrole production is present in all three domains of cellular life. In eukaryotes, the pathway was shaped by past endosymbioses and consecutive endosymbiotic and non-endosymbiotic gene transfers which turned it into a true mosaic of enzymes with diverse origins and often with different subcellular localization. This thesis discusses the current view on the heme biosynthesis in eukaryotic phototrophs, how the pathway evolved during the history of plastid endosymbioses and how it adapted in response to evolutionary events.

1 Introduction

Tetrapyrrole biosynthetic pathway is undoubtedly one of the most important enzymatic pathways in living organisms. Besides heme, the pathway provides precursors for the synthesis of other crucial tetrapyrroles (Frankenberg et al., 2003) (Fig. 1). In phototrophs, the substantial part of protoporphyrin IX, the last precursor of heme before iron chelation, is employed in the synthesis of chlorophylls, photosynthetic pigments essential for the capturing of light energy (Mochizuki et al., 2010). Uroporphyrinogen III, in eubacteria and Archaea, is needed for the production of cobalamin (vitamin B12), which is also an essential cofactor for enzymes involved in DNA synthesis and energy metabolism in all organisms (Roth et al., 1996; Scott and

Roessner, 2002). Furthermore, uroporphyrinogen III is also a precursor for the formation of siroheme, which bacteria and plants use for nitrogen and sulfur fixation (Frankenberg et al., 2003). The ability to synthesize (or acquire) tetrapyrroles is therefore inherently linked to life. This can be exemplified by the most widely spread tetrapyrrole, the heme.

1.1 Importance of heme in living organisms

Heme functions as a cofactor in numerous enzymes. Heme-binding proteins can be found in cells of nearly all living organisms, with most of them being heme autotrophs capable for *de novo* heme synthesis (Yin and Bauer, 2013). Heme auxotrophs, for instance trypanosomatids and many other parasites, must seek heme in their hosts (Kořený et al., 2009; Kořený et al., 2013). Analogously, predators (e.g. the nematode *Caenorhabditis elegans*) obtain heme from their prey (Sinclair and Hamza, 2015). In every way, heme is essential for survival of vast majority of living organisms. However, there are a few exceptions: pathogenic and anaerobic bacteria (Decker et al., 1970) and single known aerobic eukaryote *Phytomonas serpens* (Kořený et al., 2012) are able to survive without heme.

Heme consists of four porphyrin rings (which is how tetrapyrroles earned their name) attached to one another in a cyclic manner coordinated with divalent iron ion (Frankenberg et al., 2003) (Fig.1). Functional hemoproteins are assembled through interactions of the iron ion with various apo-proteins. Hemoproteins have many different biological activities primarily determined by the ability of heme to exist either in oxidized ferric (Fe^{3+}) or reduced ferrous (Fe^{2+}) state that enables them to accept and donate electrons from/to various compounds (Frankenberg et al., 2003). Such defined conformational plasticity is employed by the cytochromes, the most abundant group of hemoproteins that participate in diverse redox reactions in the cell. Although in aerobic organisms the main importance of cytochromes lies in the electron transport during oxidative phosphorylation and photosynthesis (Beale and Weinstein, 1991), cytochromes also respond to oxidative stress (Chelikani et al., 2004; Bonifacio et al., 2011) or detoxify drugs (Schenkman and Jansson, 2003; Anzenbacher and Anzenbacherová, 2001). As the main component of hemoglobin, heme is also responsible for the transport of diatomic gasses (O_2 and CO_2) in red blood cells (Chen

et al., 2008). Alternatively, other proteins benefit from interactions with heme which gives them the ability to sense biologically important gases such as oxygen, carbon monoxide, and nitric oxide (Jain and Chan, 2003). As the nitric oxide-sensory part of guanylyl cyclase, heme also plays an important role in signal transduction (Hou et al., 2006). Although most of the intracellular heme is bound to proteins, probably due to its toxicity, presence of “uncommitted” or “free” heme was reported (Ponka, 1999). Intracellular heme reversibly binds to numerous transcription factors (Creusot et al., 1989; Sassa and Nagai, 1996; Zhang et al., 1998; Zhang and Hach, 1999) and ion channels (Tang et al., 2003; Horrigan et al., 2005).

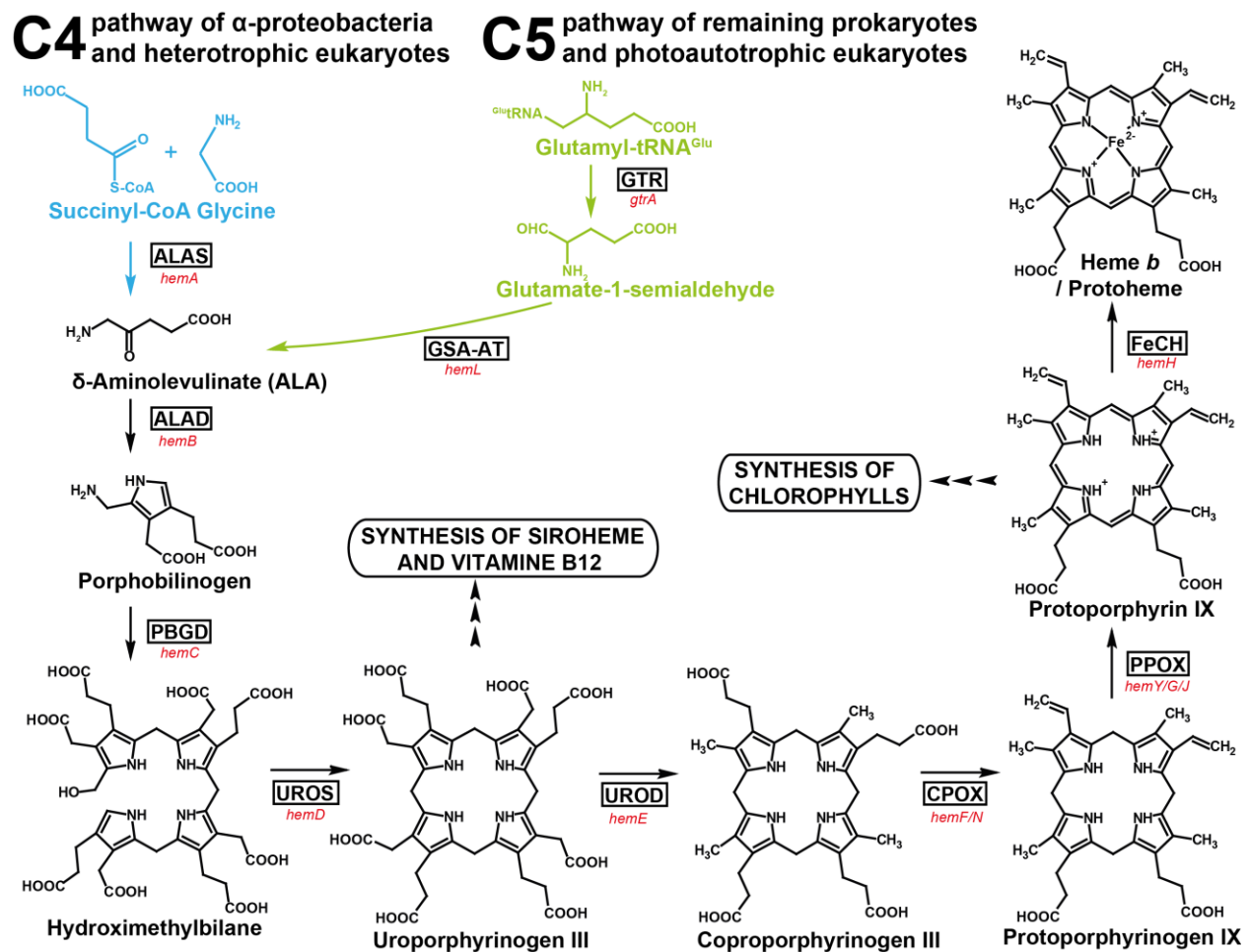


Fig. 1. General scheme of heme biosynthesis pathway. The difference is shown between C4 and C5 pathways for the synthesis of ALA. ALA is synthesized in the plastid via the C5 pathway in most eukaryotic phototrophs while in primarily heterotrophic eukaryotes, chromerids, and apicomplexans, it is produced in the mitochondrion using the C4 (Shemin) pathway. Enzymes abbreviations are explained in the text. Gene names are shown in red.

2 Biosynthesis of heme

Most organisms synthesize heme exclusively under aerobic conditions. It is mainly because of two oxidative steps involved in the pathway, catalyzed by coproporphyrinogen oxidase (CPOX) and protoporphyrinogen oxidase (PPOX), require oxygen for their function (Hainemann et al., 2008). However, the oxygen-independent counterparts of both above-mentioned enzymes were identified within genomes of some bacterial species (Panek and O'Brian, 2002; Frankenberg et al., 2003). Furthermore, several hemoproteins have been found in anaerobic organisms by Pereira and colleagues (1998) suggesting the possible appearance of heme pathway even before the emergence of oxygen in the atmosphere. Such hemoproteins might have equipped ancient anaerobic organisms with the ability to sense diatomic gases and to avoid nitrosative and oxidative stress (Poole and Hughes, 2000; Green et al., 2009).

Heme biosynthesis is a rather complex process which requires a coordinated activity of up to nine different enzymes (Fig. 1). At the same time, this multi-step pathway is exceedingly conserved in all three domains of cellular life, though, there are differences in the synthesis of the δ -aminolevulinic acid (ALA), a common precursor of all tetrapyrroles (Fig. 1). All prokaryotes, except α -proteobacteria (Beale, 1999; Panek and O'Brian, 2002), use the so-called C5 pathway for ALA synthesis, where the five-carbon glutamate bound to its tRNA is used as a precursor for the synthesis (hence the name C5). Consequently, most photosynthetic eukaryotes that inherited their heme pathways from cyanobacterial plastid predecessors also use the C5 pathway. The C5 pathway consists of two enzymes: glutamyl-tRNA reductase (GTR) and glutamate 1-semialdehyde aminotransferase (GSA-AT) progressively converting glutamyl-tRNA^{Glu} into ALA. In comparison, α -proteobacteria use so-called C4 (or Shemin) pathway (Jordan and Shemin, 1972; Ferreira and Gong, 1995; Duncan et al., 1999), where ALA is synthesized through the condensation of glycine and succinyl-CoA, catalyzed by a single enzyme ALA synthase (ALAS). Consequently, most heterotrophic eukaryotes also use the C4 pathway adopted from the α -proteobacterial predecessors of mitochondria, where the pathway is located.

The consecutive steps of heme biosynthetic pathway are common to all organisms that are able to synthesize heme (Heinemann et al., 2008, Layer et al., 2010) and lead to the ultimate step of the pathway catalyzed by ferrochelatase (FeCH). The association of iron with the porphyrin ring of protoporphyrin IX in this step forms heme *b* (or protoheme) as the most abundant type of heme in respiratory cytochromes, photosynthetic electron transport chain components, peroxidases, plant cytochrome P-450, and other oxidative enzymes (Beale and Weinstein, 1991). Furthermore, heme *b* serves as a precursor for other types of heme that are component parts of mitochondrial and/or plastidial cytochromes. Heme *a* is present exclusively as a prosthetic group of cytochromes of the respiratory complex IV (cytochrome *c* oxidase) and is derived from heme *b* via sequential modification catalyzed by heme *o* synthase and heme *a* synthase (Swenson et al., 2016). Another important prosthetic group of both mitochondrial complex III (cytochrome *b/c1*) and plastidial cytochrome *b6f* is the heme *c*. It is also present in soluble cytochrome *c* and therefore the synthesis of heme *c* is closely intertwined with the synthesis of cytochrome *c* and represents the most complex modification of heme (Kranz et al., 2009).

2.1 Biosynthesis of heme in eukaryotes and the significance of the endosymbiotic gene transfer

In comparison with prokaryotes, the eukaryotic heme biosynthesis is somewhat more complex. This complexity follows the evolutionary history of eukaryotes that was vastly influenced by multiple endosymbioses with bacterial and phototrophic eukaryotic cells. These endosymbiotic processes resulted in the rise of semi-autonomous organelles such as mitochondria and plastids (Vothknecht and Soll, 2007). By extension, endosymbiosis led to different spatial organization among cellular compartments and to mosaic evolutionary origins of heme pathway enzymes in eukaryotic heterotrophs and autotrophs (Oborník and Green, 2005).

It is believed that in the early stage of an endosymbiosis, both involved cells had their own heme biosynthetic pathway. During later stages many genes from the endosymbiont were transferred to the host nucleus by the endosymbiotic gene transfer (EGT) (Martin et al., 1998; Delwiche, 1999; Martin et al., 2002; Timmis et al., 2004;

Oborník and Green, 2005; Jiroutová et al., 2010), allowing for enhanced host control over the organelle and reduced functional redundancy of cellular biochemistry. In this manner, some of the endosymbiont genes eventually replaced original host genes and took over their function, which resulted in a mosaic character of eukaryotic heme biosynthesis pathway (Oborník and Green, 2005). In most cases, enzymes targeted to a given organelle originated from the bacterial ancestor of the respective organelle. In this logic, enzymes of cyanobacterial origin should be located in a plastid, enzymes originating from mitochondria (α -proteobacteria) should be located to mitochondria, and enzymes originating from the host nucleus should be of cytosolic location. Besides experimental evidence, localization of particular enzymes can be predicted based physicochemical properties of their N-terminus, where an organellar targeting presequence might be found, while cytosolic enzymes would lack such presequence. With that being said, in eukaryotes, the placement of enzymes does not always correspond to their origins (Kořený et al., 2011; Cihlář et al., 2016), which will be discussed in the following sections.

2.2 Biosynthesis of heme in primary heterotrophs

Heme biosynthesis is in primarily heterotrophic eukaryotes located in two different cellular compartments (Dailey et al., 2005) (Fig. 2A). The biosynthesis starts in mitochondria where ALA is synthesized by ALAS via C4 pathway (Jordan and Shemin, 1972). The putative origin of the nuclear gene encoding for ALAS is in mitochondrion, because it is only present in α -proteobacteria and primarily heterotrophic eukaryotes. None of the eukaryotic ALAS examined so far have ever been proven to possess a predictable mitochondrial targeting presequence (termed the mitochondrial transit peptide, mTP); Still ALAS has never been found outside of the mitochondria (Kořený et al., 2011). This is very likely due to the fact that ALAS utilizes succinyl-CoA, a product of the mitochondrial TCA cycle, as the initial substrate. ALA is then transported to the cytosol where ALA dehydratase (ALAD), porphobilinogen deaminase (PBGD; also called hydroxymethylbilane synthase), uroporphyrinogen synthase (UROS), and uroporphyrinogen decarboxylase (UROD) are located. ALA export from the mitochondrion is probably coupled with glycine import in order to balance the amount of

glycine for ALA synthesis (Chiabrando et al., 2014; Guernsey et al., 2009; Hamza and Dailey, 2012).

The re-location of subsequent steps of tetrapyrrole synthesis to cytosol probably follows the attempts to reduce oxidative stress caused by the highly reactive intermediates of heme biosynthesis in the mitochondrion (Vavilin and Vermaas, 2002). Oxidative stress posed on the mitochondrion is already high enough because of the free radicals generated by electrons escaping from the respiratory chain (Chen et al., 1993; Turrens, 2003; Murphy, 2009). There is an interesting difference between localization of coproporphyrinogen oxidase (CPOX) among heterotrophs: this enzyme is located inside mitochondria of animals, oomycetes (heterotrophic stramenopiles) and some fungi. It means that the genes coding for the CPOX, at some point of the endosymbiotic process, independently acquired N-terminal pre-sequences required for a mitochondrial targeting. On the other hand, CPOX of other eukaryotic heterotrophs represented by some fungi (e.g. yeast), rhizarians, some apicomplexans, and ciliates is likely to be located to the cytosol. The last two steps of heme pathway catalyzed by two closely interacting enzymes, the protoporphyrinogen oxidase (PPOX) and the ferrochelatase (FeCH) (Ferreira et al., 1988), are again located in mitochondria of heterotrophs (Oborník and Green, 2005; Kořený and Oborník, 2011). Such organization probably reflects the highest need of heme in cytochromes of the mitochondrial respiratory chain. Furthermore, it is also important for the regulation of the pathway since heme is known to mediate the feedback inhibition of ALA synthesis (Czarnecki and Grimm, 2012; Masuda and Fujita, 2008).

Regardless of their mitochondrial/cytosolic localization, ALAS and uroporphyrinogen synthase (UROS) seem to be the only two enzymes originating from the α -proteobacterial/mitochondrial genome, while genes for the other enzymes more likely originate from the eukaryotic exosymbiont (host) nucleus (Oborník and Green, 2005; Kořený et al., 2011; Kořený and Oborník, 2011).

2.3 Biosynthesis of heme in phototrophic eukaryotes

2.3.1 Primary phototrophs

The evolution of heme biosynthesis in photosynthetic eukaryotes was largely influenced by the process of primary endosymbiosis involving a primary heterotrophic eukaryote (primary host, exosymbiont) and a cyanobacterium (primary endosymbiont). This has resulted in the emergence of primary plastids in Archaeplastida, the eukaryotic supergroup containing glaucophytes, rhodophytes (red algae), chlorophytes (green algae) and higher plants (Rodríguez-Ezpeleta et al., 2005; Archibald, 2009; Ponce-Toledo et al., 2017). During the primary endosymbiogenesis, the cyanobacterial genes coding for enzymes of heme biosynthetic pathway were transferred into the host nucleus. Hence, they are expressed in the cytosol and their protein products need to be delivered back to the place of their action (Fig. 2B). The entire pathway is bound to the plastid stroma.

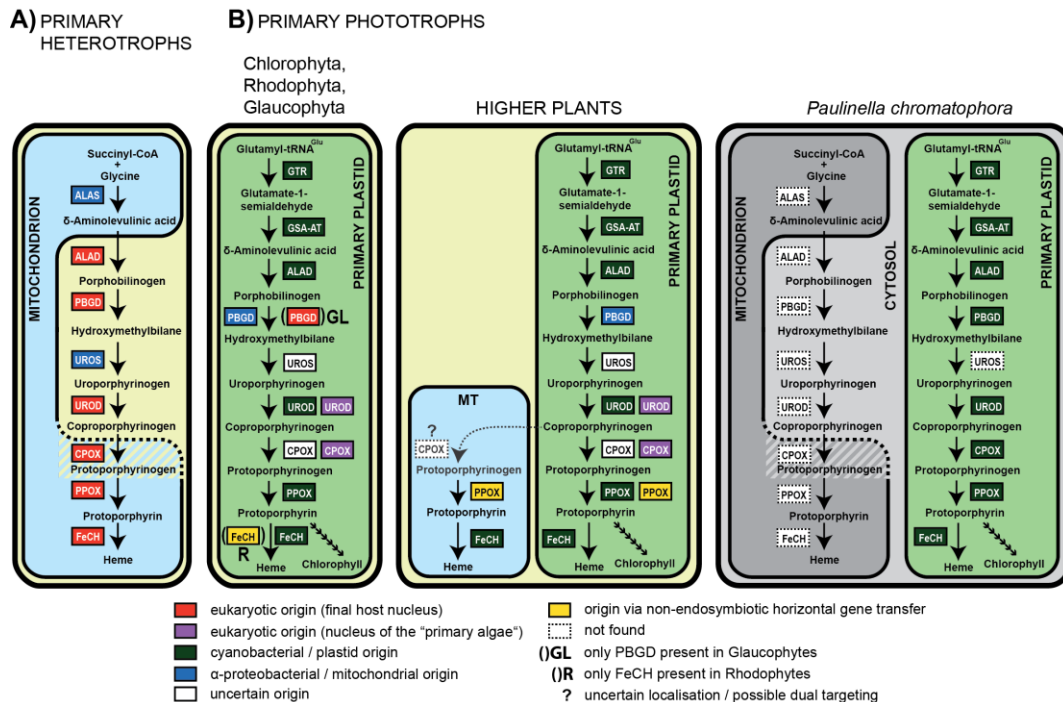


Fig. 2. Origins and subcellular locations of the enzymes of heme biosynthetic pathway in primary phototrophs. Enzymes abbreviations are explained in the text. Colored rectangles indicate the respective origins of genes. Schemes of primary heterotroph and phototroph heme pathways are based on Kofený and Oborník (2011). Origin of glaucophyte PBGD and heme pathway of *Paulinella chromatophora* were revealed by phylogenetic analyses (see text for details).

Indeed, the cyanobacterial origin of most enzymes involved in heme biosynthesis in primary phototrophs is supported by phylogenetic analyses (Oborník and Green, 2005; Kořený et al., 2011; Kořený and Oborník, 2011; Cihlář et al., 2016). Just like cyanobacteria, Archaeplastida synthesize ALA via C5 pathway and genes coding for GTR, GSA-AT, ALAD, PPOX and FeCH show strong affiliation to their cyanobacterial counterparts. Although the cyanobacterial origin of UROS genes is not well supported by phylogenetic analyses, they are very likely derived from the endosymbiont as well. In phylogenetic analyzes, all photosynthetic eukaryotes group together and some analyses place this group close to cyanobacteria (Kořený and Oborník, 2011; Cihlář et al., 2016). Some of the genes encoding enzymes of the heme pathway are duplicated in most primary phototrophs. One of two copies of UROD and CPOX genes are likely of cyanobacterial origin. The other copies of UROD and CPOX do not cluster with cyanobacteria, but instead they seem to represent the original eukaryotic genes because of their grouping with heterotrophic eukaryotes. This indicates that they were retargeted to the plastid by acquiring the N-terminal pre-sequence and complement with the original cyanobacterial enzymes in a plastid heme pathway. Another similar example is found in red algae as they differ from other Archaeplastida by possessing a FeCH that clusters among bacteria. It is likely that the original cyanobacterial gene was replaced by a bacterial gene via non-edosymbiotic horizontal gene transfer (Kořený and Oborník, 2011).

PBGD is the last enzyme to mention and represents a special case. Most plastid PBGDs are clearly related to α -proteobacterial homologs (Oborník and Green, 2005; Kořený and Oborník, 2011; Kořený et al., 2012; Cihlář et al., 2016). It is possible that this gene originated in the mitochondrion of early Archaeplastida and was transferred to the eukaryotic nucleus of an ancestor of Archaeplastida during mitochondrial symbiogenesis. Curiously, this enzyme is missing from eukaryotic heterotrophs. Three genes encoding PBGD were supposedly present in the early primary alga, particularly of eukaryotic (exosymbiont), mitochondrial and cyanobacterial origins. In the lineage of glaucophytes, the mitochondrial PBGD was lost and the cyanobacterial gene was replaced by the gene originating from the exosymbiont (host) nucleus. In contrast, plants, primary green and red algae retained the mitochondrial (α -proteobacterial)

PBGD, while eukaryotic and plastid genes were lost (Cihlář et al., unpublished). Another possibility is that the cyanobacterial gene was never transferred to the nucleus and a copy of the preexisting nuclear-encoded gene was retargeted from the mitochondrion to the plastid (Oborník and Green, 2005). It needs to be stressed that there have probably been several attempts for endosymbiosis in the evolutionary history of eukaryotic phototrophs and those mitochondria and plastids we see today are the latest ones which persisted until now (Keeling et al., 2015). Accordingly, gene transfer and protein targeting have, probably, already been introduced with previous unstable organelles thus allowing host genes (the case of glaucophyte PBGD) or previously acquired foreign genes (the case of PBGD of remaining Archaeplastida) to be targeted to the current plastid endosymbiont (Larkum et al., 2007; Keeling, 2013).

Higher plants seem to have retained remnants of the secondary host (exosymbiont) pathway (Fig. 2B). In 1993, Smith et al. demonstrated PPOX activity in mitochondrial fractions from pea leaves. Later it was reported that one of GFP-tagged CPOX proteins is exclusively targeted to the mitochondria of maize, suggesting a possible collocation of the last three enzymatic steps of the heme biosynthesis in both chloroplast and mitochondrion (Williams et al., 2006). On the other hand, mitochondrial CPOX activity in higher plants has not been observed by Smith et al. (1993) and Santana et al. (2002) and no further evidence for the presence of CPOX in plant mitochondria has been published since then. Therefore, it was suggested that portion a of protoporphyrinogen IX can be exported from plastids and transferred to mitochondria, correcting the previous assumption that only the last two enzymes (PPOX, FeCH) are converting mitochondrial protoporphyrinogen IX to protoheme (Lermontova et al., 1997; Chow et al., 1998) (Fig. 2B).

There are two genes coding for PPOX (PPOX1 and PPOX2) in higher plants (Narita et al., 1996; Lermontova et al., 1997). As for phylogenetic origin, PPOX1 shows affinity with cyanobacteria and it is shared with other phototrophic eukaryotes. PPOX2, on the other hand, has been found exclusively in land plants and seems to have been acquired horizontally from bacteria (Kořený and Oborník, 2011; Kořený et al., 2011). Unlike PPOX1, which is targeted to the plastid, PPOX2 was experimentally localized to

the mitochondrion (Lermontova et al., 1997; Watanabe et al., 2001). Similarly, there are two types of FeCH present in higher plants with distinct tissue specific and development-dependent expression pattern (Hey et al., 2016). While type I FeCH (FC1) is expressed mainly in non-photosynthetic (root) tissues and is induced upon stress, type II FeCH (FC2) serves for heme synthesis in photosynthetic tissues (leaves and shoots) (Chow et al., 1998; Singh et al., 2002, Nagai et al., 2007). Although it has been shown that the majority of FeCH activity in plants is associated with the plastid, there was still some FeCH activity detectable in mitochondria (Cornah et al., 2002; Masuda et al., 2003). Based on *in vitro* experiments Chow and colleagues (1997); Suzuki and colleagues (2002) reported a dual-targeting of FC1 into both plastids and mitochondria. However, others failed to confirm the mitochondrial localization of FC1 (Lister et al., 2001; Masuda et al., 2003; Heazlewood et al., 2004; Huang et al., 2009). For the time being, the possible location of FeCH in plant mitochondria remained ambiguous. However, there is no obvious reason for the presence of PPOX in plant mitochondria in the absence of FeCH that would prevent the accumulation of protoporphyrin IX, which can be harmful to the cell. Quite the contrary, it would be advantageous for plants to import protoporphyrinogen IX into mitochondria and synthesize heme there in non-photosynthetic respiring tissues (Watanabe et al., 2001). Indeed, recently published work on tobacco provided new experimental evidence for dual localization of FC1, because a transgenic FC1 protein was immunologically shown to be present in mitochondria (Hey et al., 2016). Moreover, when overexpressed, the FeCH activity in mitochondria significantly increased in comparison with wild-type plants, suggesting that FC1 contributes to mitochondrial heme biosynthesis (Hey et al., 2016).

In primary phototrophs the establishment of a plastid tetrapyrrole biosynthesis inevitably led to functional redundancy with the mitochondrial/cytosolic pathway. Then again, some enzymes of mitochondrial and cytosolic origin have been relocated to the plastid. The production (and need) of chlorophyll is much higher than the production (and need) of heme in photosynthetic cells (Papenbrock et al., 1999), so it was advantageous to utilize pathway entirely localized to the plastid. Moreover, heme is also needed for the synthesis of plastidial cytochromes and light-harvesting pigments, so, a plastid is the place where the majority of heme is utilized. This mostly led to gradual loss

of the mitochondrial/cytosolic pathway in primary phototrophs. The placement of the entire pathway in plastid is again important for its regulation. However, the regulation in photoautotrophs that synthesize both, heme and chlorophyll, in the same compartment, is a rather complex process that involves negative feedback loops from both branches in order to inhibit the synthesis of ALA (Czarnecki and Grimm, 2012; Vavilin and Vermaas, 2002). A smaller portion of heme required by other cellular compartments, especially by the respiratory chain cytochromes in mitochondria is probably transported outside of the plastid by yet unknown mechanisms.

The emergence of plastids in Archaeplastida has long been considered the only case of primary endosymbiosis with cyanobacteria. Based on available evidence, this event has been dated between 900 to 1,600 million years ago (Hedges et al., 2004; Yoon et al., 2004; Shih and Matzke, 2013). Nonetheless, the evidence for the occurrence of another, more recent primary endosymbiotic event has been reported (Martin et al., 2005; Yoon et al., 2006; Nowack et al., 2008). *Paulinella chromatophora* is a freshwater filose thecamoeba of cercozoan affiliation that possesses a cyanobacterial photosynthetically active organelle referred to as the chromatophore (Nowack et al., 2008). No paper has been published about the biosynthesis of heme in *P. chromatophora*. Still, we managed to identify short sequence homologs of the C5 pathway enzymes (GTR and GSA-AT) and almost all subsequent enzymes of the heme pathway, except UROS, within the chromatophore genome. All the enzymes belong within the cyanobacterial clade with a strong affiliation to *Synechococcus* spp. and *Prochlorococcus* spp. This means that genes of the chromatophore heme pathway have not been yet transferred to the nucleus of *Paulinella*, which is consistent with a rather recent endosymbiosis in *P. chromatophora*, dated between 60 and 200 MYA (Marin et al., 2005; Yoon et al., 2009; Nowack et al., 2016), and that there must be also a host pathway present in *Paulinella* which would supply heme for remaining cellular compartments (Fig. 2B). However, we do not know how exactly the amoebic host heme pathway look like, since there is no data available so far. One can assume though that it is not significantly different from mitochondrial-cytosolic pathways present in other heterotrophic eukaryotes.

2.3.2 Other phototrophs

The ability of photosynthesis among eukaryotes is not confined exclusively to Archaeplastida, and there are many other eukaryotic genera from different supergroups which were endowed with plastids. These plastids were likely acquired through the process of complex endosymbioses (secondary, tertiary or other advanced events), when a primary (or a complex) alga was engulfed by a heterotrophic eukaryote (e.g. Archibald, 2009; Oborník et al., 2009; Keeling, 2013) (Fig. 3). Secondary algae appear in three eukaryotic supergroups: Chlorarachniophytes (Rhizaria) (Fig. 5B) and photosynthetic euglenids (Excavata) (Fig. 5C) acquired their plastids by the engulfment of a green alga in two independent endosymbiotic events (Rogers et al., 2007). In contrast, cryptophytes, alveolates, stramenopiles, and haptophytes (the “CASH taxa”) (Fig. 4) acquired their plastid by an engulfment of a red alga, initially proposed to be in a single endosymbiotic event (Lane and Archibald, 2008; Sanchez-Puerta and Delwiche, 2008). This was also the basic presumption of the so-called Chromalveolate hypothesis (Cavalier-Smith, 1999). The monophyly of red-derived complex plastids is supported by the phylogenies based on plastid encoded genes. However, plastids in the red lineage are not morphologically identical. Plastids of cryptophytes contain a remnant nucleus of the engulfed alga, while other red-derived plastids do not. Plastids of dinoflagellates are three-membrane-bound, other plastids are bound with four, etc. Accordingly, phylogenies based on genes from host nuclear genomes are often incompatible with this scenario (Janouškovec et al., 2010; Gould et al., 2015), testifying for multiple secondary endosymbiotic events in CASH taxa involving a red algal endosymbiont (Falkowski et al., 2004; Bodył et al., 2009) or even suggesting that the plastid was transferred horizontally between at least some of these lineages (Petersen et al., 2014).

Furthermore, there are other examples of tertiary and/or quaternary endosymbiotic events in the evolution of phototrophic eukaryotes (Delwiche, 1999; Keeling, 2010; Archibald, 2015) related to dinoflagellates that are prone to the numerous plastid replacements. Some dinoflagellates replaced their ancestral peridinin-pigmented plastids with plastids originating from serial secondary or tertiary endosymbioses (Inagaki et al., 2000; Saldarriaga et al., 2001; Ishida and Green 2002).

For example a group of dinoflagellates called dinotoms (e.g. *Durinskia baltica*, *Glenodinium foliaceum*) (Chesnick et al, 1996; Imanian and Keeling, 2007) harbor two distinct endosymbionts: the original peridinin plastid is supplemented by the newly obtained diatom tertiary endosymbiont (Hehenberger et al., 2014) (Fig. 4E). *Lepidodinium chlorophorum* is the only example of dinoflagellate where the original red-derived plastid was replaced by a secondary green plastid (Elbrachter and Schnepf, 1996; Takishita et al., 2008) (Fig. 5A).

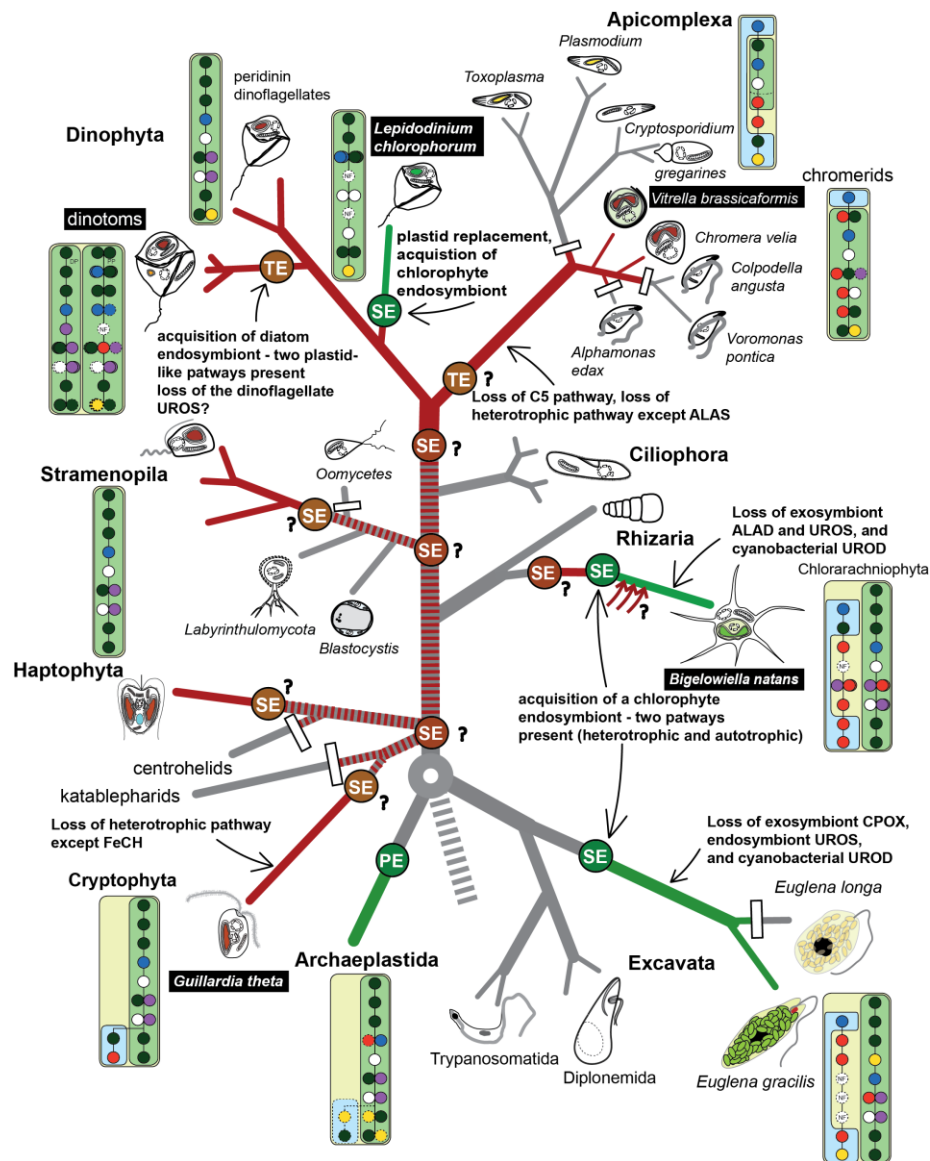


Fig. 3. A simplified scheme of the evolution of the tetrapyrrole biosynthesis pathway in eukaryotic phototrophs adopted from Cihlář et al., 2016 (study II of this thesis) with an emphasis on the models from this review (black bars). Simplified schemes of heme biosynthetic pathways of each group are explained in the text and in figures 2, 4, and 5.

2.3.2.1 Biosynthesis of heme in algae with secondary and tertiary plastids derived from the red lineage

As inferred from phylogenies (Kořený et al., 2011; Cihlář et al., 2016), there are few differences in heme biosynthesis in haptophytes, stramenopiles, and most dinoflagellates (Butterfield et al., 2013), when compared to the primary algae (Fig. 4). Genes of the heme pathway were transferred from the nucleus of the engulfed alga into the host nucleus where they have substituted original host genes with a complete set of enzymes from the algal endosymbiont. On the other hand, there are also several exceptions. For example, a new subfamily of putative GSA-AT was identified within dinoflagellates with N-terminal pre-sequences putatively targeting them to the peridinin plastid. The new subfamily clusters with proteobacteria and therefore likely originates from non-endosymbiotic gene transfer (Cihlář et al., 2016). Likewise, it has been shown that FeCH of apicomplexans cluster with proteobacteria suggesting its possible origin from non-endosymbiotic gene transfer (Sato and Wilson, 2003; Kořený et al., 2011). FeCH of the same origin was later identified in other algae belonging to the Alveolata group (dinoflagellates and chromerids) (Kořený et al., 2011; Cihlář et al., 2016). Based on phylogenetic analyses, it seems that apicomplexans and chromerids acquired this enzyme independently from dinoflagellates that probably kept the bacteria-related FeCH which resulted from non-endosymbiotic gene transfer earlier in the evolution of rhodophytes. It is worth to mention that apicomplexans and some dinoflagellates possess only the bacterial-derived FeCH, while other dinoflagellates and chromerids possess also an additional enzyme of a cyanobacterial origin (Kořený et al., 2011, Cihlář et al., 2016).

Cryptophytes share a similar organization of heme pathway present in the aforementioned algae, which was shown on the example of *Guillardia theta* (Cihlář et al., 2016) (Fig. 4D). However, in contrast with most algae of the CASH group, some genes originating from the primary endosymbiotic event have been duplicated, such as the cyanobacterial genes for GTR and UROD, and CPOX copy inherited from the primary endosymbiont nucleus. Interestingly, *G. theta* also possesses an additional FeCH of eukaryotic (host) origin, which is putatively targeted to the mitochondrion, like

its close homologs in heterotrophic eukaryotes. The same location of FeCH was also found in land plants. In this case, however, it is a protein of bacterial origin (Kořený and Oborník, 2011; Kořený et al., 2011), which is also likely to be dual-targeted to both the mitochondrion and the plastid (Hey et al., 2016). Moreover, in land plants, the mitochondrial FeCH is also coupled with PPOX, together contributing to mitochondrial heme biosynthesis. There is no obvious reason related to heme biosynthesis for *G. theta* to retain the mitochondrial ferrochelatase, especially if PPOX is missing from the mitochondrion. It is evident that protoporphyrinogen IX, the substrate for the ferrochelatase, would also not be available unless it is transported from the plastid. This could also mean that mitochondrial FeCH of cryptophytes has been evolutionarily conserved because of its, so far undescribed, role in the biology of the organism. Therefore, we thoroughly examined the protein sequence of *G. theta* PPOX, which clearly possesses a signal peptide suggesting its localization in the plastid. Judging from results of additional *in silico* predictions (Mitoprot II, TargetP) it is possible that PPOX is dually targeted to both the plastid and the mitochondrion. This would mean that *G. theta* is the only alga bearing plastids of red provenience that possesses two heme pathways (although the mitochondrial one is incomplete). Taking into account the presence of nucleomorph in cryptophyte plastids, such an arrangement could have resulted from independent and more recent plastid acquisition in cryptophytes (Burki et al., 2016; Cihlář et al., 2016).

Other algae with plastids derived from the red lineage have more complex mechanisms of heme biosynthesis. There are two redundant heme pathways present in dinotoms that seem to supply tetrapyrroles in parallel to both independent symbiotic partners (Hehenberger et al., 2014; Cihlář et al., 2016) (Fig. 4E). While one pathway is located in the tertiary diatom endosymbiont and the respective enzymes cluster together with sequences from free-living diatoms, the other represents the original pathway of the peridinin plastid, which also supplies heme to cytosol and mitochondrion, and conversely the respective enzymes group together with other dinoflagellate sequences (Cihlář et al., 2016).

SECONDARY / TERTIARY PHOTOTROPHS WITH RED DERIVED PLASTIDS

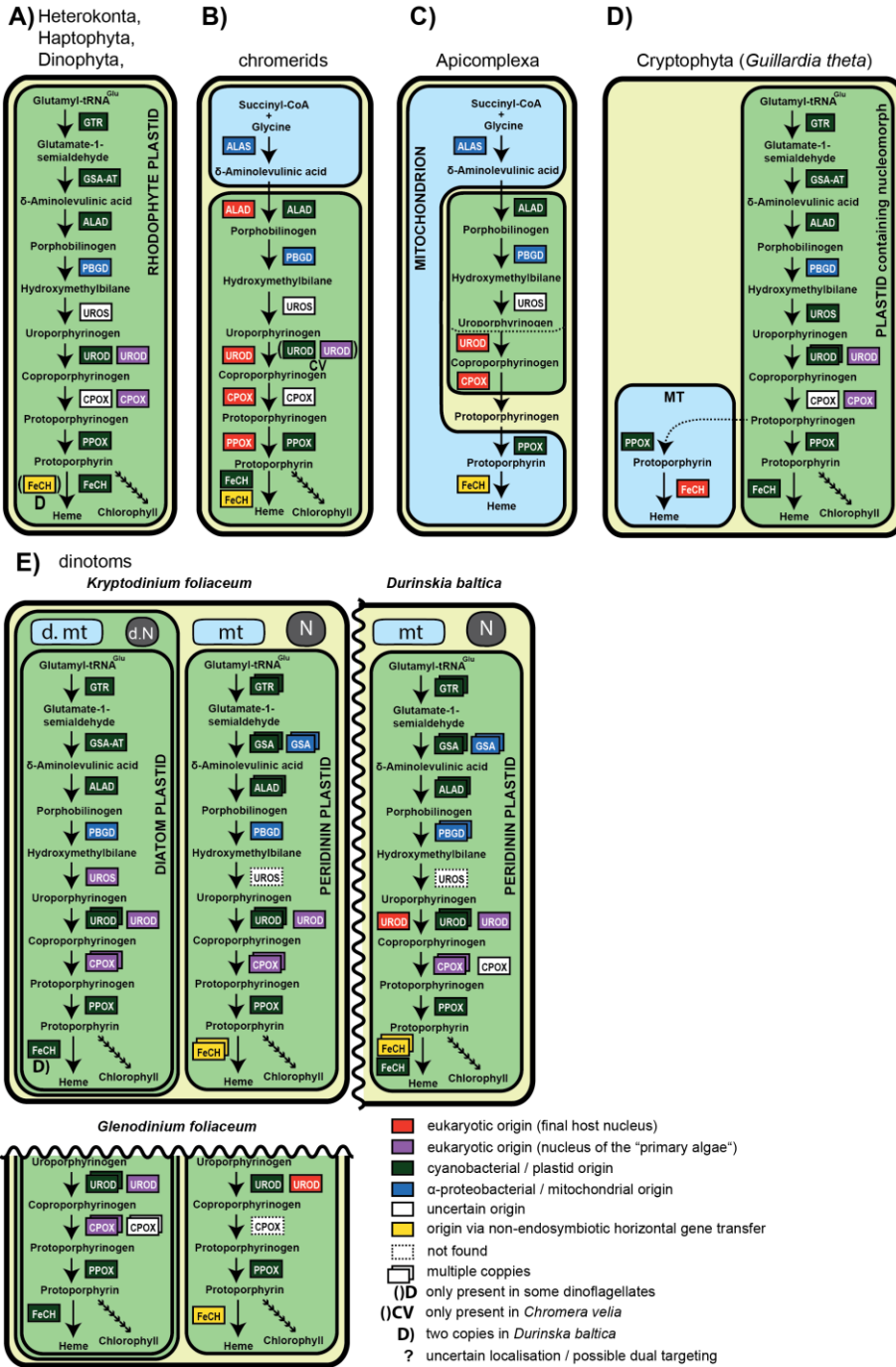


Fig. 4. Origins and subcellular locations of the enzymes of heme biosynthetic pathway in secondary/tertiary phototrophs with the red-derived plastids. Enzymes abbreviations are explained in the text. Colored rectangles indicate the respective origins of genes. The scheme of apicomplexan heme pathway is based on Kořený and Oborník (2011). Origins of chromerid, cryptophyte, and dinotom enzymes were revealed by phylogenetic analyses (studies I and II in this thesis). The localization schemes are based on *in silico* predictions.

Most of the apicomplexan parasites also carry a secondary highly reduced non-photosynthetic red-derived plastid (apicoplast), pointing to the photosynthetic history of their non-parasitic ancestor (McFadden et al., 1996; Kohler et al., 1997; Maréchal and Cesbron-Delauw, 2001; Wilson, 2002; Janouškovec et al., 2010; Füßy and Oborník, 2017¹). The mechanism of apicomplexan heme biosynthesis was largely influenced by the evolutionary transition from the photosynthetic to the parasitic lifestyle (van Dooren et al., 2006). Apicomplexan heme pathway localizes to three different cellular compartments and it is a collection of enzymes of heterotrophic and photosynthetic origins, reflecting only a partial replacement of the host pathway genes by those of the endosymbiont. (Sato et al., 2004; Wu, 2006; Kořený et al., 2013) (Fig. 4C). The metabolic flow between the mitochondrion and the apicoplast necessary for the effectivity of heme pathway is probably facilitated by an intimate association between these organelles (Hopkins et al., 1999; Waller and McFadden, 2005; Botté et al., 2013). The first committed step of the pathway takes place in the mitochondrion where ALA is synthesized by ALAS just like in primary heterotrophs (Sato et al., 2004; Wu, 2006). Next four steps (ALAD, PBGD, UROS and UROD) take place in the apicoplast, with the exception of *Toxoplasma gondii* where the UROD localizes in the cytosol (Wu, 2006). CPOX is located exclusively in the cytosol of all apicomplexans investigated so far (Wu, 2006; Nagaraj et al., 2010) and last two steps (PPOX and FeCH) are again placed in the mitochondrion (van Dooren et al., 2006; Wu, 2006; Nagaraj et al., 2010). Despite the origin of the apicoplast in red-derived plastids, only a portion of enzymes (ALAD, PBGD, UROS, and PPOX) are derived from the red algal endosymbiont. ALAS, UROD and CPOX represent the enzymes derived from the secondary host pathway, and FeCH was probably acquired via non-endosymbiotic gene transfer from proteobacteria (Sato and Wilson, 2003; Oborník and Green, 2005; Kořený et al., 2011), as mentioned above.

Homological pathways with the apicomplexan heme synthesis (in term of origin) were discovered in *Chromera velia* and *Vitrella brassicaformis*, the free-living or symbiotic algae with close phylogenetic relationship to apicomplexan parasites (Moore et al., 2008; Oborník et al., 2009; Janouškovec et al., 2010; Oborník et al., 2011^{1, 2}; Oborník and Lukeš, 2013; Cumbo et al., 2013; Weatherby and Carter, 2013;

Janouškovec et al., 2015; Woo et al., 2015; Füssy and Oborník, 2017²). It has been shown that the heme pathway of *C. velia* consists of enzymes present in both the apicoplasts and the secondary red-derived plastids, which is particularly evident in the example of UROD and CPOX (Fig. 4B). Both enzymes are present in multiple copies, one displaying the secondary host origin in *C. velia* enzymes like in apicomplexans, and the other copies showing origin in the secondary endosymbiont (Kořený et al., 2011). Later it was discovered that heme pathway in *V. brassicaformis* is homologous to that found in *C. velia* with just a small difference in the gene repertoire. UROD of *V. brassicaformis*, just like in apicomplexans, is single-copy and originates from the secondary host nucleus (Woo et al., 2015). However, both algae share one principal and unique feature of heme biosynthesis. Unlike other phototrophs, *C. velia* and *V. brassicaformis* lack the enzymes of the C5 pathway. ALA for both heme and chlorophyll synthesis is formed via the C4 pathway in the mitochondrion (Kořený et al., 2011; Oborník and Lukeš 2013; Oborník and Lukeš 2015; Woo et al., 2015), similarly to primary eukaryotic heterotrophs and apicomplexans (Fig. 4B). We recently found that both algae presumably lack the gene for ferredoxin-dependent glutamate synthase, an enzyme synthesizing the starting substrate for the C5 pathway, that is predicted to be plastid-localized in green algae/plants, red algae, diatoms, and dinoflagellates (Cihlář et al., unpublished). Apparently, *Chromera* and *Vitrella* possess only the cytosolic NADH-dependent glutamate synthase as nothing but three, respectively two, sequences coding for this enzyme were found in their genomes. These sequences do not contain any N-terminal targeting pre-sequences and thus the enzymes are supposed to be cytosolic. The lack of plastid-targeted ferredoxin-dependent glutamate synthase probably left the plastid-localized C5 pathway fully dependent on the import of the cytosolic glutamate and later could have resulted into re-allocation of the first step of the heme biosynthetic pathway into mitochondrion where the ALA is synthesized via the C4 pathway. Obviously, such metabolic bottleneck probably appeared already in apicomplexan, colpodellid and chromerid common ancestor and led to the subsequent loss of the glutamate-dependent C5 pathway in these lineages (Cihlář et al., unpublished). According to phylogenetic studies, the remaining heme pathway enzymes of chromerids have evolutionary origins mostly in

the plastid, and most of them possess a well predicted bipartite targeting sequences (BTS), suggesting their putative location in the complex plastid (Kořený et al., 2011). Although the spatial separation of the beginning and the end of the heme pathway was not found in any other organism, it seems that such an arrangement corresponds to the high demand of tetrapyrroles in the photosynthetic organelles of chromerids. This, on the other hand, necessitates a high demand on ALA transporters to cover the import of ALA required by the plastid tetrapyrrole synthesis. Also, the feedback regulation of the pathway as known from other models (Vavilin and Vermaas, 2002; Masuda and Fujita, 2008; Czarnecki and Grimm, 2012) is probably not possible. One can easily argue that there might be something hidden behind the heme biosynthesis in chromerid algae. In order to confirm this intriguing arrangement of the tetrapyrrole synthesis in *C. velia in vivo*, Jitka Kručinská (a colleague from Miroslav Oborník research group) xenotransfected several genes of the *C. velia* heme pathway (ALAS, both ALADs, UROS, and both FeCHs) into *Toxoplasma gondii* and *Phaeodactylum tricorutum* cells. These experiments, however, brought ambiguous results, especially when compared with *in silico* predictions. In fact, only the ALAS showed a suspected mitochondrial location in both transfection systems. In *T. gondii* both ALADs displayed cytosolic localizations, UROS was targeted in the mitochondrion, and only the two ferrochelatases were compatible with the apicoplast import machinery in *T. gondii* (Kručinská et al., unpublished). In *P. tricorutum*, all the remaining enzymes, including ALADs that do not possess a signal peptide, were either localized to the so-called “blob-like structures” (BLS) indicating periplastid localization (Kilian & Kroth, 2005) or the signals co-localized with ER-tracker indicating the presence of the enzyme in the endoplasmic reticulum, obviously due to the absence of the diatom SP cleavage site motif ASA↓FAP (Gruber et al., 2007) in *C. velia* proteins (Kručinská et al., unpublished). In short, all the examined enzymes (except ALAS) were recognized by the plastid outer membrane import machinery of *P. tricorutum*. Additionally, both ferrochelatases were targeted to the apicoplast of *T. gondii*. However, the observed cytosolic location of both ALADs and mitochondrial location of UROS in *T. gondii* may imply for the possible dual targeting of heme pathway enzymes. There is still need to clarify these results, particularly by performing immunofluorescence assay (IFA) based on antibodies

designed against each particular enzyme of heme pathway, to disclose the true placement and the organization of the heme biosynthesis in both chromerid algae.

2.3.2.2 Biosynthesis of heme in algae with secondary plastids derived from the green lineage

The dinoflagellate *Lepidodinium chlorophorum* (Elbrachter and Schnepf, 1996; Takishita et al., 2008) possesses a chlorophyte-derived secondary plastid that supposedly replaced the original peridinin-pigmented plastid. Nearly all genes of the plastid heme enzymes were found in our transcriptomic data of this dinoflagellate. This led to an interesting finding based on phylogenetic analyses of particular enzymes (Cihlář et al., 2016) (Fig. 5A). Surprisingly, some enzymes (GTR, PPOX, and FeCH) cluster with red algae and secondary algae with red-derived plastids. Other enzymes (ALAD, GSA-AT, and PPOX) group in clades with green algae and plants on the root, but such clade always contains also dinoflagellates appearing in a sister position to the *L. chlorophorum* sequences (Cihlář et al., 2016). The close relationship of *L. chlorophorum* ALAD, GSA-AT, and PPOX enzymes with their orthologs from dinoflagellates with red-derived plastids suggests a possible red origin of these enzymes. Apparently, the origin of the pathway is homologous to that of Peridinin-pigmented dinoflagellates. The original rhodophyte-derived heme pathway appears to be highly conserved in spite of the presence of a chlorophyte-derived plastid in *L. chlorophorum*. It is therefore very likely that the *L. chlorophorum* took advantage of red genes to function in the newly acquired plastid, so the original heme pathway, introduced with the peridinin plastid, remained functionally conserved. This is also in agreement with a mosaic evolution of plastid proteomes according to the “shopping bag” or plastid promiscuity hypotheses (Larkum et al., 2007). This conserved rhodophyte origins of the heme pathway enzymes in *L. chlorophorum* likely represent a set of enzymes originally present in the peridinin-pigmented plastids of dinoflagellates, which protein targeting machinery was compatible with the newly acquired plastid. Thus the entire heme pathway could have been relocated into the new plastid (Cihlář et al., 2016). Here it is worth mentioning the hypothesis that algal lineages currently possessing secondary plastids of red algal origin acquired green algal derived plastids

earlier in their evolutionary history, that were later replaced by the modern red-derived plastids we can see nowadays (Moustafa et al., 2009; Woehle et al., 2011). In this case, the close relationship of *L. chlorophorum* enzymes with their orthologs from secondary algae bearing red-derived plastids could be explained as a result of the previous possession of green genes in these lineages that were transferred to the host nucleus and later retargeted to the newly acquired plastids. Thus it could be possible that genes from chlorophyte-derived secondary plastid of *L. chlorophorum* resemble those genes from peridinin-pigmented plastids, which we now consider red. However, this explanation does not seem to be likely, since the “green endosymbiont first” hypothesis has been repeatedly questioned, particularly, on the basis of credibility of presented phylogenetic analyses (Burki et al., 2012; Deschamps and Moreira, 2012; Moreira and Deschamps, 2014).

Photosynthetic euglenids and chlorarachniophytes seem to have acquired their secondary green plastids more recently. Especially it is evident from the example of photosynthetic euglenids that constitute an advanced monophyletic group and the presence of plastids in their phagotrophic and/or osmotrophic relatives has never been confirmed (Hrdá et al., 2012; Yamaguchi et al., 2012). *Euglena gracilis* synthesizes ALA by both the heterotrophic C4 and the phototrophic C5 pathway, and quite early it was suggested that *E. gracilis* possesses two independent heme pathways for the production of heme and chlorophyll (Weinstein and Beale, 1983), which was later confirmed using sequence data (Kořený and Oborník, 2011). Similarly to *E. gracilis*, *Bigelowiella natans* possesses two nearly complete heme pathways (Cihlář et al., 2016). One is similar to heme synthesis in primary heterotrophs, and likely represents the original pathway of the host (exosymbiont), while the second pathway originates from the algal endosymbiont (Fig. 5B, C). Such an organization of tetrapyrrole synthesis probably resulted from slow reduction of the mitochondrial-cytosolic pathway, which is not immediately replaced by the plastid one, and denotes that both pathways may coexist for quite some evolutionary time within a single cell (Kořený and Oborník, 2011).

SECONDARY PHOTOTROPHS WITH GREEN DERIVED PLASTIDS

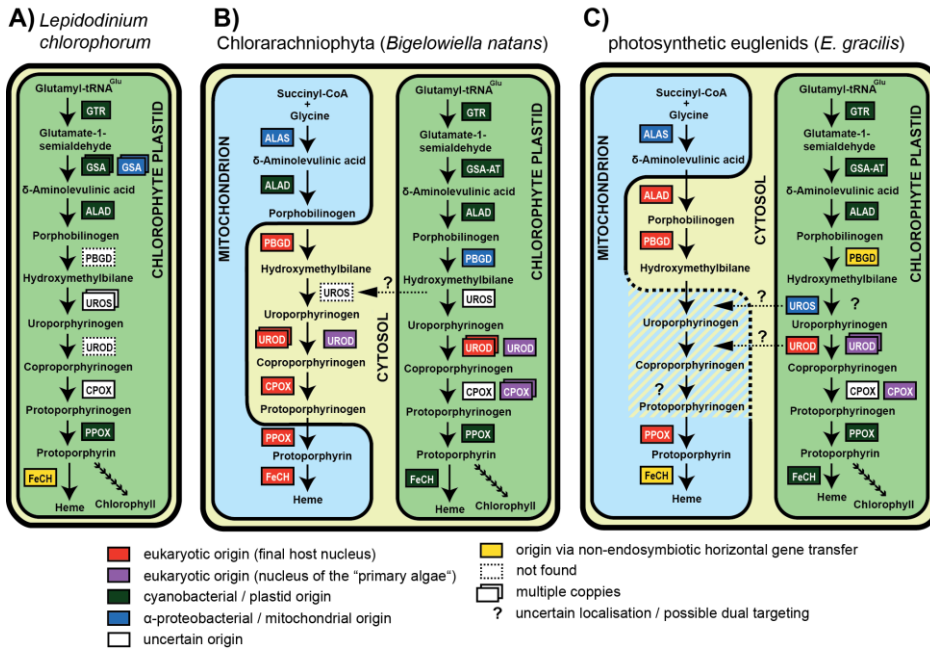


Fig. 5. Origins and subcellular locations of the enzymes of heme biosynthetic pathway in secondary phototrophs with the green-derived plastids. Enzymes abbreviations are explained in the text. Colored rectangles indicate the respective origins of genes. Origins of *L. chlorophorum* and chlorarachniophyte enzymes were revealed by phylogenetic analyses (study II in this thesis), the localization schemes are based on *in silico* predictions. The scheme of *E. gracilis* heme pathways was adopted from Kořený and Oborník (2011).

Regardless of the independent origins of chlorarachniophytes (Rhizaria) and euglenophytes (Excavata) (Rogers et al., 2007), they exhibit similar trends of the evolution of tetrapyrrole synthesis. The reduction of the redundant mitochondrial-cytosolic pathway has already begun in both *E. gracilis* and *B. natans*. In *Euglena*, two enzymes, UROS and UROD, of the original heterotrophic pathway have been functionally replaced by enzyme counterparts of plastid origin, likely via dual targeting (Kořený and Oborník, 2011). A similar pattern was found in *Bigelowiella*, where the plastid-derived ALAD and UROS enzymes functionally replaced their heterotrophic counterparts (Cihlář et al., 2016). On the contrary, paralogs of the heterotrophic UROD copy seem to be retargeted to the plastid in both algae. Interestingly, the above-mentioned retargeted ALAD of *B. natans* is now a mitochondrial enzyme, as it contains a mitochondrial targeting presequence. Apart from these rather unique features and arrangements of tetrapyrrole synthesis present in both algae, the rest of the pathways correspond to what was already found in other heterotrophs and phototrophs.

In contrast to *E. gracilis*, only the plastid FeCH shows an expected chlorophyte origin in *B. natans* and, interestingly, most of the plastid heme pathway genes exhibit affinity to red algae. It has already been reported that *B. natans* possesses a large number of red-related genes, including photosynthesis-related genes, suggesting multiple horizontal gene transfers from the red lineage (Archibald et al., 2003; Yang et al., 2014). At any rate, the protein replacement in an established essential metabolic pathway via endosymbiotic or horizontal gene transfer could have interfered with its function and therefore such replacements are not very likely. At the same time, the similarities in origins of heme pathway genes between *L. chlorophorum* and *B. natans* are quite conspicuous. Hence it is possible that the rhodophyte origin of the heme pathway in *B. natans* was established similarly in *L. chlorophorum*, suggesting a previous presence of a hypothetical red-derived plastid in the ancestor of chlorarachniophytes (Cihlář et al., 2016).

The presence of multiple pathways for tetrapyrrole synthesis in photosynthetic euglenids and chlorarachniophytes is unique and probably represents a significant milestone in the evolutionary history of each endosymbiotic event. It is likely that every extant eukaryotic phototroph passed through a similar stage of metabolic redundancy during the process of transformation from a heterotroph. In early stages of endosymbiosis when plastids were not essential for host survival and they could have been lost without the hassle just as it may happened in early-branching lineages to the recent phototrophic crown groups (e.g. ciliates and oomycetes) (Kořený et al., 2011), that were suggested for the ancient presence of a red algal plastid (Janouškovec et al., 2010). In subsequent stages of the endosymbiosis, the plastid took over some essential metabolic processes of the cytosol/mitochondrion, such as tetrapyrrole synthesis, and became indispensable for the survival of the host. This is most likely a reason why plastids, even those that remained non-photosynthetic, are still present within cells of some heterotrophic or parasitic lineages i.e. apicomplexans and the parasitic algae from the group of Archaeplastida, i.e. *Polytomella* and *Helicosporidium*, still using the plastid-located heme pathway for the synthesis of heme (Atteia et al., 2005; de Koning and Keeling, 2004).

3 Conclusion

The heme pathway in phototrophic eukaryotes plays a crucial role in metabolism. With plastid endosymbiosis in progress, an organism suddenly faces metabolic redundancy that is usually resolved by plastid takeover of the cellular supply of tetrapyrroles. This, in most eukaryotic phototrophs, leads to the exclusive localization of the heme pathway and the synthesis of tetrapyrroles to the plastid compartment. However, some eukaryotic phototrophs possess multiple or hybrid pathways for tetrapyrrole synthesis. Dinotoms synthesize tetrapyrroles in parallel in both the tertiary diatom plastids and the peridinin plastids; *Euglena gracilis* and *Bigelowiella natans* possess both the mitochondrial/cytosolic and the plastid heme pathways, thus probably represent an intermediate state in endosymbiosis; and chromerids use a hybrid pathway initiated in the mitochondrion via C4 pathway, which is predicted to continue in the plastid. From the phylogenetic point of view, heme pathway of all eukaryotes is a true mosaic, sort of an imaginary shopping bag of enzymes originating in proteobacteria, cyanobacteria, and eukaryotes that were collected and combined throughout the history of mitochondrion and plastid endosymbioses. At the same time, the pathways appears to be evolutionarily well conserved even following serial endosymbioses, especially those seen in the dinoflagellate *Lepidodinium chlorophorum* and the chlorarachniophyte *Bigelowiella natans*.

References:

- Anzenbacher P, Anzenbacherová E. (2001) Cytochromes P450 and metabolism of xenobiotics. *Cell Mol Life Sci.* 58:737-47.
- Archibald JM, Rogers MB, Toop M, Ishida K, Keeling PJ. (2003) Lateral gene transfer and the evolution of plastid-targeted proteins in the secondary plastid-containing alga *Bigeloviella natans*. *Proc Natl Acad Sci U S A.* 100:7678-83.
- Archibald JM. (2009) The puzzle of plastid evolution. *Curr Biol.* 19:R81-8.
- Archibald JM. (2015) Genomic perspectives on the birth and spread of plastids. *Proc Natl Acad Sci U S A.* 112:10147-53.
- Atteia A, van Lis R, Beale SI. (2005) Enzymes of the heme biosynthetic pathway in the nonphotosynthetic alga *Polytomella* sp. *Eukaryot Cell.* 4:2087-97.
- Beale SI, Weinstein JD (1991) Biochemistry and regulation of photosynthetic pigment formation in plants and algae. In: P.M. Jordan (Ed.) *Biosynthesis of Terrapyrroles* Elsevier Science Publishers B.V. pp. 155-235.
- Beale SI. (1999) Enzymes of chlorophyll biosynthesis. *Photosynth Res.* 60:43-73.
- Bhattacharya D, Yoon HS, Hackett JD. (2004) Photosynthetic eukaryotes unite: endosymbiosis connects the dots. *Bioessays.* 26:50-60.
- Bodył A, Stiller JW, Mackiewicz P. (2009) Chromalveolate plastids: direct descent or multiple endosymbioses? *Trends Ecol Evol.* 24:119-21.
- Bonifacio A, Martins MO, Ribeiro CW, Fontenele AV, Carvalho FE, Margis-Pinheiro M, Silveira JA. (2011) Role of peroxidases in the compensation of cytosolic ascorbate peroxidase knockdown in rice plants under abiotic stress. *Plant Cell Environ.* 34:1705-22.
- Botté CY, Yamaro-Botte Y, Rupasinghe TWT, Mullin KA, MacRae JI, Spurck TP, Kalanon M, Shears MJ, Coppel RL, Crellin PK, Marechal E, McConville MJ, McFadden GI. (2013) Atypical lipid composition in the purified relict plastid (apicoplast) of malaria parasites. *Proc Natl Acad Sci U S A.* 110: 7506-7511.
- Burki F, Flegontov P, Oborník M, Cihlář J, Pain A, Lukes J, Keeling PJ. (2012) Re-evaluating the green versus red signal in eukaryotes with secondary plastid of red algal origin. *Genome Biol Evol.* 4:626-35.
- Burki F, Kaplan M, Tikhonenkov DV, Zlatogursky V, Minh BQ, Radaykina LV, Smirnov A, Mylnikov AP, Keeling PJ. (2016) Untangling the early diversification of eukaryotes: a

phylogenomic study of the evolutionary origins of Centrohelida, Haptophyta and Cryptista. *Proc Biol Sci.* 283(1823).

Butterfield ER, Howe CJ, Nisbet RE. (2013) An analysis of dinoflagellate metabolism using EST data. *Protist.* 164:218-36.

Cavalier-Smith T. (1999) Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. *J Eukaryot Microbiol.* 46:347-66.

Chelikani P, Fita I, Loewen PC. (2004) Diversity of structures and properties among catalases. *Cell Mol Life Sci.* 61:192-208.

Chen ZX, Silva H, Klessig DF. (1993) Active oxygen species in the induction of plant systemic acquired-resistance by salicylic-acid. *Science* 262: 1883-1886.

Chen W, Dailey HA, Paw BH. (2010) Ferrochelatase forms an oligomeric complex with mitoferrin-1 and Abcb10 for erythroid heme biosynthesis. *Blood* 116:628-30.

Chesnick HM, Morden CW, Schmieg AM. (1996) Identity of the endosymbiont of *Peridinium foliaceum* (Pyrrophyta): Analysis of the *rbclC* operon. *J Phycol.* 32:850-857.

Chiabrando D, Mercurio S, Tolosano E. (2014) Heme and erythropoiesis: more than a structural role. *Haematologica* 99: 973-983.

Chow KS, Singh DP, Roper JM, Smith AG. (1997) A single precursor protein for ferrochelatase-I from *Arabidopsis* is imported *in vitro* into both chloroplasts and mitochondria. *J Biol Chem.* 272:27565-71.

Chow KS, Singh DP, Walker AR, Smith AG. (1998) Two different genes encode ferrochelatase in *Arabidopsis*: mapping, expression and subcellular targeting of the precursor proteins. *Plant J.* 15:531-41.

Cihlář J, Füssy Z, Horák A, Oborník M. (2016) Evolution of the tetrapyrrole biosynthetic pathway in secondary algae: Conservation, redundancy and replacement. *PLoS ONE* 11: e0166338.

Creusot F, Gaisne M, Verdière J, Slonimski PP. (1989) A novel tRNA(Ala) gene and its adjacent sigma element downstream from the CYP1 (HAP1) gene in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* 17:1865-6.

Cumbo VR, Baird AH, Moore RB, Negri AP, Neilan BA, Salih A, van Oppen MJ, Wang Y, Marquis CP. (2013) *Chromera velia* is endosymbiotic in larvae of the reef corals *Acropora digitifera* and *A. tenuis*. *Protist.* 164:237-44.

Curtis BA, Tanifuji G, Burki F, Gruber A, Irimia M, Maruyama S, Arias MC, Ball SG, Gile GH, Hirakawa Y, Hopkins JF, Kuo A, Rensing SA, Schmutz J, Symeonidi A, Elias M, Eveleigh RJ, Herman EK, Klute MJ, Nakayama T, Oborník M, Reyes-Prieto A, Armbrust EV, Aves SJ, Beiko RG, Coutinho P, Dacks JB, Durnford DG, Fast NM, Green BR, Grisdale CJ, Hempel F, Henrissat B, Höppner MP, Ishida K, Kim E, Kořený L, Kroth PG, Liu Y, Malik SB, Maier UG, McRose D, Mock T, Neilson JA, Onodera NT, Poole AM, Pritham EJ, Richards TA, Rocap G, Roy SW, Sarai C, Schaack S, Shirato S, Slamovits CH, Spencer DF, Suzuki S, Worden AZ, Zauner S, Barry K, Bell C, Bharti AK, Crow JA, Grimwood J, Kramer R, Lindquist E, Lucas S, Salamov A, McFadden GI, Lane CE, Keeling PJ, Gray MW, Grigoriev IV, Archibald JM. (2012) Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. *Nature*. 492:59-65.

Czarnecki O, Grimm B. (2012) Post-translational control of tetrapyrrole biosynthesis in plants, algae, and cyanobacteria. *J Exp Bot*. 63: 1675-1687.

Dailey TA, Woodruff JH, Dailey HA. (2005) Examination of mitochondrial protein targeting of haem synthetic enzymes: *in vivo* identification of three functional haem-responsive motifs in 5-aminolaevulinate synthase. *Biochem J*. 386:381-6.

de Koning AP, Keeling PJ. (2004) Nucleus-encoded genes for plastid-targeted proteins in *Helicosporidium*: functional diversity of a cryptic plastid in a parasitic alga. *Eukaryot Cell*. 3:1198-205.

Decker K, Jungermann K, Thauer RK. (1970) Energy production in anaerobic organisms. *Angew Chem Int Ed Engl*. 9:138–158.

Delwiche CF. (1999) Tracing the Thread of Plastid Diversity through the Tapestry of Life. *Am Nat*. 154:S164-S177.

Deschamps P, Moreira D. (2012) Reevaluating the green contribution to diatom genomes. *Genome Biol Evol*. 4:683-8.

Duncan R, Faggart MA, Roger AJ, Cornell NW. (1999) Phylogenetic analysis of the 5-aminolevulinate synthase gene. *Mol Biol Evol*. 16:383-96.

Elbrachter M, Schnepf E. (1996) *Gymnodinium chlorophorum*, a new, green, bloom-forming dinoflagellate (Gymnodiniales, Dinophyceae) with a vestigial prasinophyte endosymbiont. *Phycologia*. 35:381-393.

Falkowski PG, Katz ME, Knoll AH, Quigg A, Raven JA, Schofield O, Taylor FJ. (2004) The evolution of modern eukaryotic phytoplankton. *Science*. 305:354-60.

- Ferreira GC, Andrew TL, Karr SW, Dailey HA. (1988) Organization of the terminal two enzymes of the heme biosynthetic pathway. Orientation of protoporphyrinogen oxidase and evidence for a membrane complex. *J Biol Chem.* 263:3835-9.
- Ferreira GC, Gong J. (1995) 5-Aminolevulinate synthase and the first step of heme biosynthesis. *J Bioenerg Biomembr.* 27:151-9.
- Frankenberg N, Moser J, Jahn D (2003) Bacterial heme biosynthesis and its biotechnological application. *Appl Microbiol Biotechnol.* 63:115-27.
- Füßy Z, Oborník M. (2017) Reductive Evolution of Apicomplexan Parasites from Phototrophic Ancestors. *Pontarotti P. Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts* (pp. 217-236). Springer International Publishing AG.
- Füßy Z, Oborník M. (2017) Chromerids and Their Plastids. In *Advances in Botanical Research* (pp. 187-218). Elsevier Ltd.
- Gould SB, Maier UG, Martin WF. (2015) Protein import and the origin of red complex plastids. *Curr Biol.* 15;25:R515-21
- Green J, Crack JC, Thomson AJ, LeBrun NE. (2009) Bacterial sensors of oxygen. *Curr Opin Microbiol.* 12:145-51.
- Gruber A, Vugrinec S, Hempel F, Gould SB, Maier UG, Kroth PG. (2007) Protein targeting into complex diatom plastids: functional characterisation of a specific targeting motif. *Plant Mol Biol.* 64:519-30.
- Guernsey DL, Jiang H, Campagna DR, Evans SC, Ferguson M, Kellogg MD, Lachance M, Matsuoka M, Nightingale M, Rideout A, Saint-Amant L, Schmidt PJ, Orr A, Bottomley SS, Fleming MD, Ludman M, Dyack S, Fernandez CV, Samuels ME. (2009) Mutations in mitochondrial carrier family gene SLC25A38 cause nonsyndromic autosomal recessive congenital sideroblastic anemia. *Nat Genet.* 41: 651-653.
- Hamza I, Dailey HA. (2012). One ring to rule them all: Trafficking of heme and heme synthesis intermediates in the metazoans. *Biochim Biophys Acta.* 1823: 1617-1632.
- Heazlewood JL, Tonti-Filippini JS, Gout AM, Day DA, Whelan J, Millar AH. (2004) Experimental analysis of the *Arabidopsis* mitochondrial proteome highlights signaling and regulatory components, provides assessment of targeting prediction programs, and indicates plant-specific mitochondrial proteins. *Plant Cell.* 16:241-56.
- Hedges SB, Blair JE, Venturi ML, Shoe JL. (2004) A molecular timescale of eukaryote evolution and the rise of complex multicellular life. *BMC Evol Biol.* 2004 4:2.

- Hehenberger E, Imanian B, Burki F, Keeling PJ. (2014) Evidence for the retention of two evolutionary distinct plastids in dinoflagellates with diatom endosymbionts. *Genome Biol Evol.* 6:2321-34.
- Heinemann IU, Jahn M, Jahn D. (2008) The biochemistry of heme biosynthesis. *Arch Biochem Biophys.* 474:238-51.
- Hey D, Ortega-Rodes P, Fan T, Schnurrer F, Brings L, Hedtke B, Grimm B. (2016) Transgenic Tobacco Lines Expressing Sense or Antisense FERROCHELATASE 1 RNA Show Modified Ferrochelatase Activity in Roots and Provide Experimental Evidence for Dual Localization of Ferrochelatase 1. *Plant Cell Physiol.* 57:2576-2585.
- Hopkins J, Fowler R, Krishna S, Wilson I, Mitchell G, Bannister L. (1999) The plastid in *Plasmodium falciparum* asexual blood stages: a three-dimensional ultrastructural analysis. *Protist.* 150: 283-295.
- Horrigan FT, Heinemann SH, Hoshi T. (2005) Heme regulates allosteric activation of the Slo1 BK channel. *J Gen Physiol.* 126:7-21.
- Hou S, Reynolds MF, Horrigan FT, Heinemann SH, Hoshi T. (2006) Reversible binding of heme to proteins in cellular signal transduction. *Acc Chem Res.* 39:918-24.
- Hrdá Š, Fousek J, Szabová J, Hampl V, Vlček Č. (2012) The plastid genome of *Eutreptiella* provides a window into the process of secondary endosymbiosis of plastid in euglenids. *PLoS One.* 7:e33746.
- Huang S, Taylor NL, Narsai R, Eubel H, Whelan J, Millar AH. (2009) Experimental analysis of the rice mitochondrial proteome, its biogenesis, and heterogeneity. *Plant Physiol.* 149:719-34.
- Imanian B, Keeling PJ. (2007) The dinoflagellates *Durinskia baltica* and *Kryptoperidinium foliaceum* retain functionally overlapping mitochondria from two evolutionarily distinct lineages. *BMC Evol Biol.* 7:172.
- Inagaki Y, Dacks JB, Doolittle WF, Watanabe KI, Ohama T. (2000) Evolutionary relationship between dinoflagellates bearing obligate diatom endosymbionts: insight into tertiary endosymbiosis. *Int J Syst Evol Microbiol.* 50:2075-81.
- Ishida K, Green BR. (2002) Second- and third-hand chloroplasts in dinoflagellates: phylogeny of oxygen-evolving enhancer 1 (PsbO) protein reveals replacement of a nuclear-encoded plastid gene by that of a haptophyte tertiary endosymbiont. *Proc Natl Acad Sci U S A.* 99:9294-9.
- Jain R, Chan MK. (2003) Mechanisms of ligand discrimination by heme proteins. *J Biol Inorg Chem.* 8:1-11.

- Janouškovec J, Horák A, Oborník M, Lukes J, Keeling PJ. (2010) A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids. *Proc Natl Acad Sci U S A*. 107:10949-54.
- Janouškovec J, Tikhonenkov DV, Burki F, Howe AT, Kolísko M, Mylnikov AP, Keeling PJ. (2015) Factors mediating plastid dependency and the origins of parasitism in apicomplexans and their close relatives. *Proc Natl Acad Sci U S A*. 112:10200-7.
- Jiroutová K, Kořený L, Bowler C, Oborník M. (2010) A gene in the process of endosymbiotic transfer. *PLoS One*. 5:e13234.
- Jordan PM, Shemin D. (1972) d-Aminolevulinic acid synthetase. In: Boyer, P.D. (Ed.), *The Enzymes*, vol. 7. Academic Press, New York and London, pp. 339-356.
- Keeling PJ. (2010) The endosymbiotic origin, diversification and fate of plastids. *Philos Trans R Soc Lond B Biol Sci*. 365:729-48.
- Keeling PJ. (2013) The number, speed, and impact of plastid endosymbioses in eukaryotic evolution. *Annu Rev Plant Biol*. 64:583-607.
- Keeling PJ, McCutcheon JP, Doolittle WF. (2015) Symbiosis becoming permanent: Survival of the luckiest. *Proc Natl Acad Sci U S A*. 112:10101-3.
- Kilian O, Kroth PG. (2005) Identification and characterization of a new conserved motif within the presequence of proteins targeted into complex diatom plastids. *Plant J*. 41:175-83.
- Köhler S, Delwiche CF, Denny PW, Tilney LG, Webster P, Wilson RJ, Palmer JD, Roos DS. (1997) A plastid of probable green algal origin in Apicomplexan parasites. *Science*. 275:1485-9.
- Kořený L, Oborník M. (2011) Sequence Evidence for the Presence of Two Tetrapyrrole Pathways in *Euglena gracilis*. *Genome Biol Evol*. 3: 359-364.
- Kořený L, Sobotka R, Janouškovec J, Keeling PJ, Oborník M. 2011. Tetrapyrrole Synthesis of Photosynthetic Chromerids Is Likely Homologous to the Unusual Pathway of Apicomplexan Parasites. *Plant Cell* 23: 3454-3462.
- Kořený L, Sobotka R, Kovářová J, Gnypová A, Flegontov P, Horvath A, Oborník M, Ayala FJ, Lukeš J. (2012) Aerobic kinetoplastid flagellate *Phytomonas* does not require heme for viability. *Proc Natl Acad Sci U S A*. 109: 3808-3813.
- Kořený L, Oborník M, Lukeš J. (2013) Make It, Take It, or Leave It: Heme Metabolism of Parasites. *Plos Pathogens* 9: e1003088.

- Kranz RG, Richard-Fogal C, Taylor JS, Frawley ER. (2009) Cytochrome c biogenesis: mechanisms for covalent modifications and trafficking of heme and for heme-iron redox control. *Microbiol Mol Biol Rev.* 73:510-28
- Lane CE, Archibald JM. (2008) The eukaryotic tree of life: endosymbiosis takes its TOL. *Trends Ecol Evol.* 23:268-75.
- Larkum AW, Lockhart PJ, Howe CJ. (2007) Shopping for plastids. *Trends Plant Sci.* 12:189-95.
- Layer G, Reichelt J, Jahn D, Heinz DW. (2010) Structure and function of enzymes in heme biosynthesis. *Protein Sci.* 19:1137-61.
- Lermontova I, Kruse E, Mock HP, Grimm B. (1997) Cloning and characterization of a plastidal and a mitochondrial isoform of tobacco protoporphyrinogen IX oxidase. *Proc Natl Acad Sci U S A.* 94:8895-900.
- Lister R, Chew O, Rudhe C, Lee MN, Whelan J. (2001) Arabidopsis thaliana ferrochelatase-I and -II are not imported into Arabidopsis mitochondria. *FEBS Lett.* 506:291-5.
- Maréchal E, Cesbron-Delauw MF. (2001) The apicoplast: a new member of the plastid family. *Trends Plant Sci.* 6:200-5.
- Martin W, Stoebe B, Goremykin V, Hapsmann S, Hasegawa M, Kowallik KV. (1998) Gene transfer to the nucleus and the evolution of chloroplasts. *Nature.* 393:162-5.
- Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, Leister D, Stoebe B, Hasegawa M, Penny D. (2002) Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc Natl Acad Sci U S A.* 99:12246-51.
- Martin B, Nowack EC, Melkonian M. (2005) A plastid in the making: evidence for a second primary endosymbiosis. *Protist.* 156:425-32.
- Masuda T, Suzuki T, Shimada H, Ohta H, Takamiya K. (2003) Subcellular localization of two types of ferrochelatase in cucumber. *Planta.* 217:602-9.
- Masuda T, Fujita Y. (2008) Regulation and evolution of chlorophyll metabolism. *Photochem Photobiol Sci.* 7: 1131-1149.
- McFadden GI, Reith ME, Munholland J, Lang-Unnasch N. (1996) Plastid in human parasites. *Nature.* 381:482.

- Mochizuki N, Tanaka R, Grimm B, Masuda T, Moulin M, Smith AG, Tanaka A, Terry MJ. (2010) The cell biology of tetrapyrroles: a life and death struggle. *Trends Plant Sci.* 15: 488-498.
- Moore RB, Oborník M, Janouskovec J, Chrudimský T, Vancová M, Green DH, Wright SW, Davies NW, Bolch CJ, Heimann K, Slapeta J, Hoegh-Guldberg O, Logsdon JM, Carter DA. (2008) A photosynthetic alveolate closely related to apicomplexan parasites. *Nature.* 451:959-63.
- Moreira D, Deschamps P. (2014) What was the real contribution of endosymbionts to the eukaryotic nucleus? Insights from photosynthetic eukaryotes. *Cold Spring Harb Perspect Biol.* 6:a016014.
- Moustafa A, Beszteri B, Maier UG, Bowler C, Valentin K, Bhattacharya D. (2009) Genomic footprints of a cryptic plastid endosymbiosis in diatoms. *Science.* 324:1724-6.
- Murphy MP. (2009) How mitochondria produce reactive oxygen species. *Biochem J.* 417: 1–13.
- Plant Physiol.* 2007 Jun;144(2):1039-51. Epub 2007 Apr 6.
- Nagai S, Koide M, Takahashi S, Kikuta A, Aono M, Sasaki-Sekimoto Y, Ohta H, Takamiya K, Masuda T. (2007) Induction of isoforms of tetrapyrrole biosynthetic enzymes, AtHEMA2 and AtFC1, under stress conditions and their physiological functions in *Arabidopsis*. *Plant Physiol.* 144:1039-51.
- Nagaraj VA, Prasad D, Arumugam R, Rangarajan PN, Padmanaban G. (2009) Characterization of coproporphyrinogen III oxidase in *Plasmodium falciparum* cytosol. *Parasitol Int.* 59:121-7.
- Nagaraj VA, Arumugam R, Prasad D, Rangarajan PN, Padmanaban G. (2010) Protoporphyrinogen IX oxidase from *Plasmodium falciparum* is anaerobic and is localized to the mitochondrion. *Mol Biochem Parasitol.* 174:44-52.
- Narita S, Tanaka R, Ito T, Okada K, Taketani S, Inokuchi H. (1996) Molecular cloning and characterization of a cDNA that encodes protoporphyrinogen oxidase of *Arabidopsis thaliana*. *Gene.* 182:169-75.
- Nowack EC, Melkonian M, Glöckner G. (2008) Chromatophore genome sequence of *Paulinella* sheds light on acquisition of photosynthesis by eukaryotes. *Curr Biol.* 18:410-8.
- Nowack EC, Price DC, Bhattacharya D, Singer A, Melkonian M, Grossman AR. (2016) Gene transfers from diverse bacteria compensate for reductive genome evolution in the

chromatophore of *Paulinella chromatophora*. Proc Natl Acad Sci U S A. 113(43):12214-12219.

Oborník M, Green BR. (2005) Mosaic origin of the heme biosynthesis pathway in photosynthetic eukaryotes. Mol Biol Evol. 22:2343-53.

Oborník M, Janouskovec J, Chrudimský T, Lukes J. (2009) Evolution of the apicoplast and its hosts: from heterotrophy to autotrophy and back again. Int J Parasitol. 39:1-12.

Oborník M, Vancová M, Lai DH, Janouškovec J, Keeling PJ, Lukeš J. (2011) Morphology and ultrastructure of multiple life cycle stages of the photosynthetic relative of apicomplexa, *Chromera velia*. Protist. 162:115-30.

Oborník M, Modrý D, Lukeš M, Černotíková-Stříbrná E, Cihlář J, Tesařová M, Kotabová E, Vancová M, Prášil O, Lukeš J. (2012) Morphology, ultrastructure and life cycle of *Vitrella brassicaformis* n. sp., n. gen., a novel chromerid from the Great Barrier Reef. Protist. 163:306-23.

Oborník M, Lukeš J. (2013) Cell biology of chromerids: autotrophic relatives to apicomplexan parasites. Int Rev Cell Mol Biol. 306:333-69.

Oborník M, Lukeš J. (2015) The Organellar Genomes of *Chromera* and *Vitrella*, the Phototrophic Relatives of Apicomplexan Parasites. Annu Rev Microbiol. 69:129-44.

Panek H, O'Brian MR. (2002) A whole genome view of prokaryotic haem biosynthesis. Microbiology 148: 2273-82.

Papenbrock J, Mock HP, Kruse E, Grimm B. (1999) Expression studies in tetrapyrrole biosynthesis: inverse maxima of magnesium chelatase and ferrochelatase activity during cyclic photoperiods. Planta 208:264-273.

Pereira IAC, Teixeira M, Xavier AV. (1998) Heme proteins in Anaerobes. Struct Bond. 91:65-89.

Petersen J, Ludewig AK, Michael V, Bunk B, Jarek M, Baurain D, Brinkmann H. (2014) *Chromera velia*, endosymbioses and the rhodoplex hypothesis--plastid evolution in cryptophytes, alveolates, stramenopiles, and haptophytes (CASH lineages). Genome Biol Evol. 6:666-84.

Ponce-Toledo RI, Deschamps P, López-García P, Zivanovic Y, Benzerara K, Moreira D. (2017) An Early-Branching Freshwater Cyanobacterium at the Origin of Plastids. Biol. 27:386-391.

Ponka P. (1999) Cell biology of heme. Am J Med Sci. 318:241-56.

- Poole RK, Hughes MN. (2000) New functions for ancient globin family: bacterial response to nitric oxide and nitrosative stress. *Mol Microbiol.* 36:775-83.
- Rodríguez-Ezpeleta N, Brinkmann H, Burey SC, Roure B, Burger G, Löffelhardt W, Bohnert HJ, Philippe H, Lang BF. (2005) Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes. *Curr Biol.* 2005 15:1325-30.
- Rogers MB, Gilson PR, Su V, McFadden GI, Keeling PJ. (2007) The complete chloroplast genome of the chlorarachniophyte *Bigeloniella natans*: evidence for independent origins of chlorarachniophyte and euglenid secondary endosymbionts. *Mol Biol Evol.* 24:54-62.
- Roth JR, Lawrence JG, Bobik TA. (1996) Cobalamin (coenzyme B12): synthesis and biological significance. *Annu Rev Microbiol.* 50:137–181.
- Saldarriaga JF, Taylor FJ, Keeling PJ, Cavalier-Smith T. (2001) Dinoflagellate nuclear SSU rRNA phylogeny suggests multiple plastid losses and replacements. *J Mol Evol.* 53:204-13.
- Sanchez-Puerta MV, Delwiche CF. (2008) A HYPOTHESIS FOR PLASTID EVOLUTION IN CHROMALVEOLATES(1). *J Phycol.* 44:1097-107.
- Santana MA, Tan FC, Smith AG. (2002) Molecular characterisation of coproporphyrinogen oxidase from *Glycine max* and *Arabidopsis thaliana*. *Plant Physiol Biochem.* 40:289-98.
- Sassa S, Nagai T. (1996) The role of heme in gene expression. *Int J Hematol.* 63:167-78.
- Sato S, Wilson RJ. (2003) Proteobacteria-like ferrochelatase in the malaria parasite. *Curr Genet.* 42:292-300.
- Sato S, Clough B, Coates L, Wilson RJ. (2004) Enzymes for heme biosynthesis are found in both the mitochondrion and plastid of the malaria parasite *Plasmodium falciparum*. *Protist.* 155:117-25.
- Schenkman JB, Jansson I. (2003) The many roles of cytochrome b5. *Pharmacol Ther.* 97:139-52.
- Scott AI, Roessner CA (2002) Biosynthesis of cobalamin (vitamin B(12)). *Scott Biochem Soc Trans.* 30:613-20.
- Shih PM, Matzke NJ. (2013) Primary endosymbiosis events date to the later Proterozoic with cross-calibrated phylogenetic dating of duplicated ATPase proteins. *Proc Natl Acad Sci U S A.* 110:12355-60.

Sinclair J, Hamza I. (2015) Lessons from bloodless worms: heme homeostasis in *C. elegans*. *Biometals*. 28:481-9.

Singh DP, Cornah JE, Hadingham S, Smith AG. (2002) Expression analysis of the two ferrochelatase genes in *Arabidopsis* in different tissues and under stress conditions reveals their different roles in haem biosynthesis. *Plant Mol Biol*. 50:773-88.

Smith AG, Marsh O, Elder GH. (1993) Investigation of the subcellular location of the tetrapyrrole-biosynthesis enzyme coproporphyrinogen oxidase in higher plants. *Biochem J*. 292:503-8.

Suzuki T, Masuda T, Singh DP, Tan FC, Tsuchiya T, Shimada H, Ohta H, Smith AG, Takamiya K. (2002) Two types of ferrochelatase in photosynthetic and nonphotosynthetic tissues of cucumber: their difference in phylogeny, gene expression, and localization. *J Biol Chem*. 277:4731-7.

Swenson S, Cannon A, Harris NJ, Taylor NG, Fox JL, Khalimonchuk O. (2016) Analysis of Oligomerization Properties of Heme a Synthase Provides Insights into Its Function in Eukaryotes. *J Biol Chem*. 291:10411-25.

Takishita K, Kawachi M, Noël MH, Matsumoto T, Kakizoe N, Watanabe MM, Inouye I, Ishida K, Hashimoto T, Inagaki Y. (2008) Origins of plastids and glyceraldehyde-3-phosphate dehydrogenase genes in the green-colored dinoflagellate *Lepidodinium chlorophorum*. *Gene*. 410:26-36.

Tang XD, Xu R, Reynolds MF, Garcia ML, Heinemann SH, Hoshi T. (2003) Haem can bind to and inhibit mammalian calcium-dependent Slo1 BK channels. *Nature*. 425:531-5.

Timmis JN, Ayliffe MA, Huang CY, Martin W. (2004) Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat Rev Genet*. 5:123-35.

Turrens JF. (2003) Mitochondrial formation of reactive oxygen species. *J Physiol*. 552: 335–344.

van Dooren GG, Stimmler LM, McFadden GI. (2006) Metabolic maps and functions of the Plasmodium mitochondrion. *FEMS Microbiol Rev*. 30:596-630.

Vavilin DV, Vermaas WFJ. (2002) Regulation of the tetrapyrrole biosynthetic pathway leading to heme and chlorophyll in plants and cyanobacteria. *Physiologia Plantarum* 115: 9-24.

Vothknecht UC, Soll J. (2007) The endosymbiotic origin of organelles: an ancient process still very much in fashion. *Biol Chem*. 388:877.

Waller RF, McFadden GI. (2005) The apicoplast: a review of the derived plastid of apicomplexan parasites. *Curr Issues Mol Biol.* 7:57-79. Weatherby K, Carter D. (2013) *Chromera velia*: The Missing Link in the Evolution of Parasitism. *Adv Appl Microbiol.* 85:119-44.

Watanabe N, Che FS, Iwano M, Takayama S, Yoshida S, Isogai A. (2001) Dual targeting of spinach protoporphyrinogen oxidase II to mitochondria and chloroplasts by alternative use of two in-frame initiation codons. *J Biol Chem.* 276:20474-81.

Weatherby K, Carter D. (2013) *Chromera velia*: The Missing Link in the Evolution of Parasitism. *Adv Appl Microbiol.* 85:119-44.

Weinstein JD, Beale SI. (1983) Separate physiological roles and subcellular compartments for two tetrapyrrole biosynthetic pathways in *Euglena gracilis*. *J Biol Chem.* 258:6799-807.

Williams P, Hardeman K, Fowler J, Rivin C. (2006) Divergence of duplicated genes in maize: evolution of contrasting targeting information for enzymes in the porphyrin pathway. *Plant J.* 45:727-39.

Wilson RJ. (2002) Progress with parasite plastids. *J Mol Biol.* 319:257-74.

Woehle C, Dagan T, Martin WF, Gould SB. (2011) Red and problematic green phylogenetic signals among thousands of nuclear genes from the photosynthetic and apicomplexa-related *Chromera velia*. *Genome Biol Evol.* 3:1220-30.

Woo YH, Ansari H, Otto TD, Klinger CM, Kolisko M, Michálek J, Saxena A, Shanmugam D, Tayyrov A, Veluchamy A, Ali S, Bernal A, del Campo J, Cihlář J, Flegontov P, Gornik SG, Hajdušková E, Horák A, Janouškovec J, Katris NJ, Mast FD, Miranda-Saavedra D, Mourier T, Naeem R, Nair M, Panigrahi AK, Rawlings ND, Padron-Regalado E, Ramaprasad A, Samad N, Tomčala A, Wilkes J, Neafsey DE, Doerig C, Bowler C, Keeling PJ, Roos DS, Dacks JB, Templeton TJ, Waller RF, Lukeš J, Oborník M, Pain A. (2015) Chromerid genomes reveal the evolutionary path from photosynthetic algae to obligate intracellular parasites. *Elife.* 4:e06974.

Wu B. (2006) Heme biosynthetic pathway in apicomplexan parasites. Dissertation available from ProQuest. Paper AAI3246266.

Yamaguchi A, Yubuki N, Leander BS. (2012) Morphostasis in a novel eukaryote illuminates the evolutionary transition from phagotrophy to phototrophy: description of *Rapaza viridis* n. gen. et sp. (Euglenozoa, Euglenida). *BMC Evol Biol.* 12:29.

Yang Y, Matsuzaki M, Takahashi F, Qu L, Nozaki H. (2014) Phylogenomic analysis of "red" genes from two divergent species of the "green" secondary phototrophs, the

chlorarachniophytes, suggests multiple horizontal gene transfers from the red lineage before the divergence of extant chlorarachniophytes. PLoS One. 9:e101158.

Yin L, Bauer CE. (2013) Controlling the delicate balance of tetrapyrrole biosynthesis. Philos Trans R Soc Lond B Biol Sci. 368:20120262.

Yoon HS, Hackett JD, Ciniglia C, Pinto G, Bhattacharya D. (2004) A molecular timeline for the origin of photosynthetic eukaryotes. Mol Biol Evol. 21:809-18.

Yoon HS, Nakayama T, Reyes-Prieto A, Andersen RA, Boo SM, Ishida K, Bhattacharya D. (2009) A single origin of the photosynthetic organelle in different *Paulinella* lineages. BMC Evol Biol. 9:98.

Zhang L, Hach A, Wang C. (1998) Molecular mechanism governing heme signaling in yeast: a higher-order complex mediates heme regulation of the transcriptional activator HAP1. Mol Cell Biol. 18:3819-28.

Zhang L, Hach A. (1999) Molecular mechanism of heme signaling in yeast: the transcriptional activator Hap1 serves as the key mediator. Cell Mol Life Sci. 56:415-26.

Paper I

EVOLUTION OF THE TETRAPYRROLE BIOSYNTHETIC PATHWAY IN SECONDARY ALGAE: CONSERVATION, REDUNDANCY AND REPLACEMENT

Cihlář J, Füssy Z, Horák A, Oborník M.

PLOS One 11(11):e0166338 (2016)

RESEARCH ARTICLE

Evolution of the Tetrapyrrole Biosynthetic Pathway in Secondary Algae: Conservation, Redundancy and Replacement

Jaromír Cihlár^{1,2}*, Zoltán Füßy¹, Aleš Horák^{1,2}, Miroslav Oborník^{1,2,3*}

1 Biology Centre, Czech Academy of Sciences, Institute of Parasitology, České Budějovice, Czech Republic, **2** University of South Bohemia, Faculty of Science, České Budějovice, Czech Republic, **3** Institute of Microbiology, Czech Academy of Sciences, Třeboň, Czech Republic

* These authors contributed equally to this work.

* obornik@paru.cas.cz



OPEN ACCESS

Citation: Cihlár J, Füßy Z, Horák A, Oborník M (2016) Evolution of the Tetrapyrrole Biosynthetic Pathway in Secondary Algae: Conservation, Redundancy and Replacement. PLoS ONE 11(11): e0166338. doi:10.1371/journal.pone.0166338

Editor: Claude Prigent, Institut de Genetique et Developpement de Rennes, FRANCE

Received: September 1, 2016

Accepted: October 26, 2016

Published: November 18, 2016

Copyright: © 2016 Cihlár et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files. Novel sequences of interest were deposited in GenBank under the accession no. KX344033-47.

Funding: The Czech Science Foundation (GAP506/12/1522) and the Czech Academy of Sciences provided funding to MO and ZF. Computation resources were provided by CERIT-SC and MetaCentrum, Brno, Czech Republic. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Tetrapyrroles such as chlorophyll and heme are indispensable for life because they are involved in energy fixation and consumption, i.e. photosynthesis and oxidative phosphorylation. In eukaryotes, the tetrapyrrole biosynthetic pathway is shaped by past endosymbioses. We investigated the origins and predicted locations of the enzymes of the heme pathway in the chlorarachniophyte *Bigelowiella natans*, the cryptophyte *Guillardia theta*, the “green” dinoflagellate *Lepidodinium chlorophorum*, and three dinoflagellates with diatom endosymbionts (“dinotoms”): *Durinskia baltica*, *Glenodinium foliaceum* and *Kryptoperidinium foliaceum*. *Bigelowiella natans* appears to contain two separate heme pathways analogous to those found in *Euglena gracilis*; one is predicted to be mitochondrial-cytosolic, while the second is predicted to be plastid-located. In the remaining algae, only plastid-type tetrapyrrole synthesis is present, with a single remnant of the mitochondrial-cytosolic pathway, a ferrochelatase of *G. theta* putatively located in the mitochondrion. The green dinoflagellate contains a single pathway composed of mostly rhodophyte-origin enzymes, and the dinotoms hold two heme pathways of apparently plastidal origin. We suggest that heme pathway enzymes in *B. natans* and *L. chlorophorum* share a predominantly rhodophytic origin. This implies the ancient presence of a rhodophyte-derived plastid in the chlorarachniophyte alga, analogous to the green dinoflagellate, or an exceptionally massive horizontal gene transfer.

Introduction

Plastid acquisitions are rare evolutionary events that give host cells the metabolic capacities of their new photosynthetic organelles. While there are only two documented primary plastid acquisitions [1,2], involving a eukaryote as host and cyanobacterium as the endosymbiont, the history of higher order eukaryote-to-eukaryote acquisitions is intensely debated [3–8]. Based on available data, it is believed that the red plastids of cryptophytes, alveolates, stramenopiles, and haptophytes (the “CASH taxa”) originate from a single ancient event with a rhodophyte alga as the endosymbiont [9–14]. However, phylogenies of the host organisms are often

Competing Interests: The authors have declared that no competing interests exist.

Abbreviations: ALA, δ -aminolevulinic acid; ALAD, ALA dehydratase; ALAS, ALA synthase; CASH, group of complex algae with red-derived plastid of putatively common origin, comprising cryptophytes, alveolates, stramenopiles and haptophytes; CPOX, coproporphyrinogen oxidase; FeCH, ferrochelatase; GSA-AT, glutamate-1-semialdehyde 2,1-aminotransferase; GTR, glutamate-tRNA reductase; HGT, horizontal gene transfer; PBGD, porphobilinogen deaminase; PPOX, protoporphyrinogen oxidase; UROD, uroporphyrinogen decarboxylase; UROS, uroporphyrinogen-III synthase.

incompatible with this scenario [6,15–18], suggesting that the plastid was transferred horizontally in at least some of these lineages. Furthermore, higher order endosymbioses and horizontal gene transfer (HGT) may be blurring our vision of eukaryotic evolution [7]. For instance, some dinoflagellates replaced their ancestral peridinin-pigmented plastids with plastids originating from serial secondary or tertiary endosymbioses [7,19–23]. The original red plastid was replaced by a secondary green plastid in *Lepidodinium chlorophorum* [24,25]; while in so-called dinotoms (*Glennodinium foliaceum*, *Durinskia baltica*), the newly obtained tertiary endosymbiont is an engulfed diatom [26,27]. Serial plastid endosymbioses are sometimes discernible by phylogenetic signal (e.g. if a green plastid replaces a red one). However, ancient events can still be difficult to pinpoint, which might account for the contradictory and peculiar phylogenetic signals observed—for example, the number of green genes in the CASH taxa [28–33] and the proposed independent origin of plastid genes in two main branches of alveolates, the dinoflagellates and apicomplexans [34,35]. In contrast, chlorarachniophytes and phototrophic euglenids acquired green plastids and their extant relatives are heterotrophic, allowing for straightforward evolutionary interpretations of gene origins, based on phylogenetic clustering with their heterotrophic kin or with the chlorophyte plastid donors [31,36,37].

The process of endosymbiosis involves endosymbiont genome reduction via gene transfer to the host nucleus [38,39], allowing for enhanced host control over the organelle [40,41] and reduced functional redundancy of cellular biochemistry [42]. However, the level of reduction differs among algae possessing complex photosynthetic organelles. Most of them, such as diatoms, dinoflagellates or phototrophic euglenids, have a highly reduced algal endosymbiont with multiple membranes surrounding the plastid as the only apparent morphological footprints revealing past complex endosymbioses. Organelle reduction tends to be higher in cases of older symbioses, but also depends on other factors including plastid number and evolutionary constraints [43,44]. For instance, cryptophyte and chlorarachniophyte plastids seem to be evolutionarily frozen and retain a remnant nucleus (nucleomorph) that provides genetic material required, e.g. for the maintenance of protein import mechanisms [31,45,46]. In dinotoms, the diatom endosymbiont still contains a plastid, a mitochondrion and a nucleus and is thought to represent an almost entirely independent cellular compartment [47,48]. Furthermore, it appears that the host dinotom cell holds a metabolically active remnant of the original peridinin-pigmented plastid, presumably the eyespot [47,49].

One of the essential biochemical pathways carried out in plastids is tetrapyrrole synthesis. Tetrapyrroles are cyclic porphyrins coordinated by iron (heme) or magnesium (chlorophyll). They are essential for life, since heme is a substantial component of the respiratory chain and chlorophyll is an indispensable compound in the conversion of light energy to chemical bonds in carbohydrates through photosynthesis. Although the kinetoplastid flagellate *Phytomonas serpens* has been shown to be able to live in the absence of heme, it is an extremely rare metabolic deviation [50]. In phototrophic eukaryotes, tetrapyrroles are required in three cellular compartments: the cytosol, the mitochondrion, and the plastid. Most phototrophs synthesize tetrapyrrole compounds exclusively in the plastid and transport them to other compartments (overview in Fig 1). The biosynthetic process, however, can be more complex; in the excavate alga *Euglena gracilis* two independent tetrapyrrole biosynthesis pathways are present [51–55], likely because of the recent acquisition of the secondary green plastid [36,37]. These parallel tetrapyrrole pathways differ in both evolutionary origin and starting substrate. One pathway derives from the heterotrophic (secondary) host and uses condensation of succinyl-CoA and glycine (via the C4 pathway) to synthesize δ -aminolevulinic acid (ALA), the first common precursor, in the mitochondrion. The heterotrophic-type synthesis takes place partly in the mitochondrion and partly in the cytosol as it does in eukaryotic primary heterotrophs [55,56]. The other pathway is located entirely in the plastid and generates ALA from glutamate (via the C5

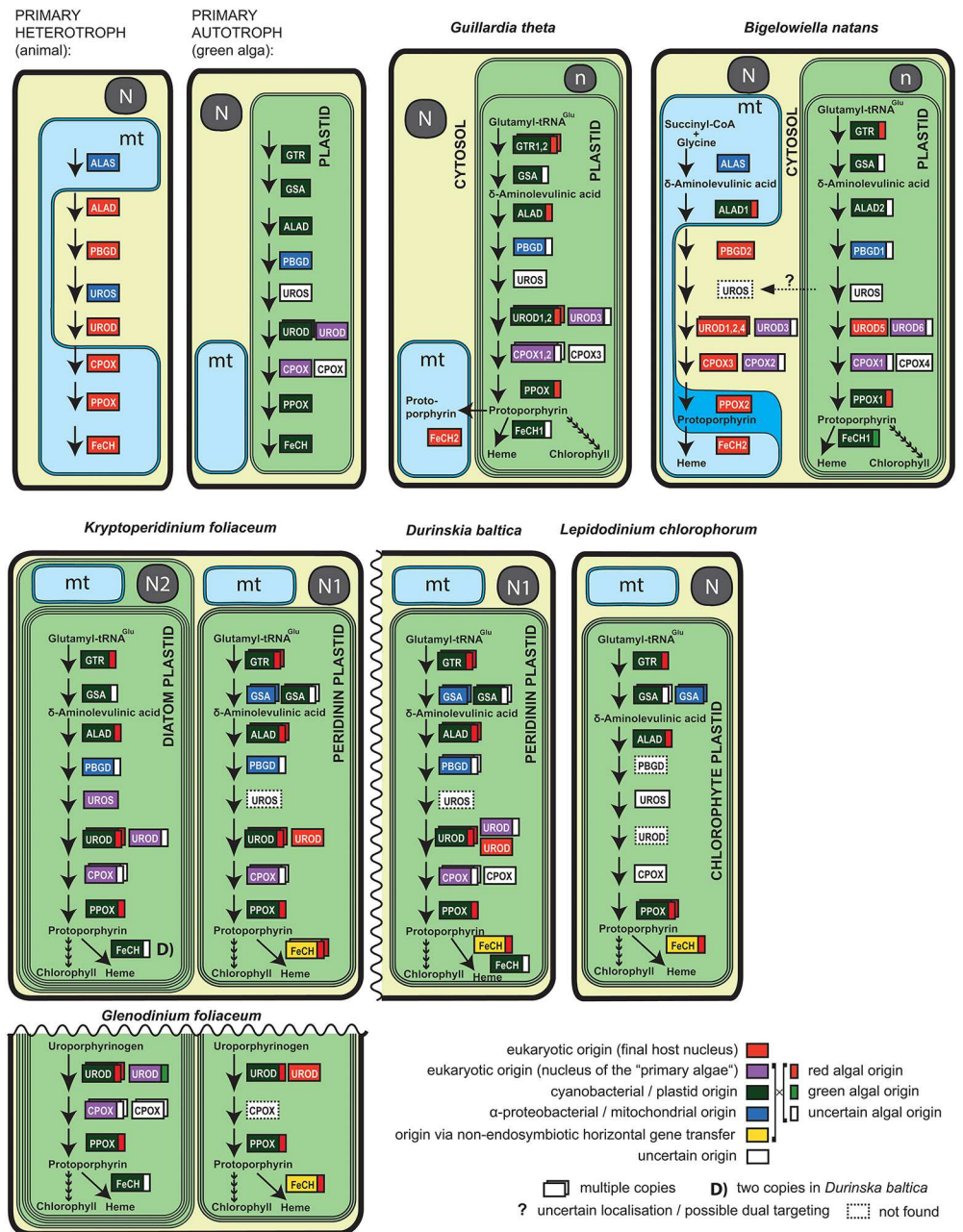


Fig 1. Arrangement of the heme biosynthetic pathway in algae with complex plastids. Inferred origins of enzymes are represented by colored boxes with flags where applicable. Localizations of *B. natans* and *G. theta* enzymes were predicted by SignalP and TargetP (see [Material and Methods](#)). Dashed arrows indicate a possible dual localization of UROS in both the cytosol and the plastid of *B. natans*. Only key metabolites are shown for clarity, for example the starting substrates for mitochondrial-cytosolic C4 (succinyl-CoA and glycine) and plastid C5 pathways (glutamyl-tRNA^{Glu}). Parts of the pathway identical to *K. foliaceum* are not depicted in the *D. baltica* and *G. foliaceum* scheme. Schematic representation of organelles: N, N1—nucleus of the host; N2—nucleus of the endosymbiont diatom; n—nucleomorph of the endosymbiont; mt—mitochondrion. Enzymes: ALAS—delta-aminolevulinic acid synthase; GTR—glutamate-tRNA reductase; GSA—glutamate-1-semialdehyde 2,1-aminotransferase; ALAD—aminolevulinic acid dehydratase; PBGD—porphobilinogen deaminase; UROS—uroporphyrinogen-III synthase; UROD—uroporphyrinogen decarboxylase; CPOX—coproporphyrinogen oxidase; PPOX—protoporphyrinogen oxidase; FeCH—ferrochelatase. A typical pathway in a primary heterotroph and a primary autotroph are shown for comparison (Košný and Oborník 2011).

doi:10.1371/journal.pone.0166338.g001

pathway). Evolutionarily, the plastid pathway originates from the green algal endosymbiont and, ultimately, from the cyanobacterium engulfed during primary endosymbiosis. In *E. gracilis*, the two tetrapyrrole pathways do not overlap, and thus produce tetrapyrroles separately for the cytosol and mitochondrion and for the plastid [51,54]. The presence of two redundant tetrapyrrole pathways is interpreted as an intermediate state in endosymbiosis [55] and would allow for the loss of one of the pathways in order to streamline cellular biochemistry. Usually, the mitochondrial-cytosolic pathway is lost in the course of evolution in eukaryotic phototrophs [55,57] and it is rare to see the plastid pathway disappear in exchange for the cytosolic counterpart, but the parasite *Perkinsus marinus* [58], for example, retained the heterotrophic heme synthesis pathway despite the continued presence of a relict plastid (reviewed in [59]). The alveolate alga *Chromera velia*, on the other hand, employs a hybrid tetrapyrrole pathway. Synthesis is initiated in the mitochondrion via the C4 pathway and is predicted to continue in the plastid. This hybrid synthesis qualifies *Chromera* as the only known phototroph able to synthesize tetrapyrroles from glycine [57], similar to heme biosynthetic processes in apicomplexan parasites that still possess a remnant, non-photosynthetic plastid [60].

The organization of heme synthesis is currently uncharacterized in most algae with complex plastids. In order to map the level of pathway conservation, reduction or replacement in further phototrophic lineages, we investigated phylogenetic relationships and predicted the cellular locations of enzymes involved in tetrapyrrole biosynthesis in the cryptophyte *Guillardia theta*, the chlorarachniophyte *Bigeloviella natans*, the green dinoflagellate *Lepidodinium chlorophorum* and the dinotoms *Durinskia baltica*, *Glenodinium foliaceum* and *Kryptoperidinium foliaceum*.

Results and Discussion

Bigeloviella natans possesses two heme pathways

In *B. natans*, we identified thirteen and nine sequences, respectively, of enzymes belonging to the algal endosymbiont heme pathway (autotrophic pathway) and to the heterotrophic (mitochondrial-cytosolic) pathway. The chlorarachniophyte host pathway is typical for eukaryotic heterotrophs, with ALA being synthesized by the mitochondrial C4 pathway; the enzymes involved are predicted to localize to the mitochondrion (aminolevulinic acid synthase, or BnALAS; protoporphyrinogen oxidase, BnPPOX2; and ferrochelatase, BnFeCH2) and cytosol (porphobilinogen deaminase, BnPBGD2; uroporphyrinogen decarboxylase BnUROD1-4, and coproporphyrinogen oxidase, BnCPOX2, -3) (see Fig 1, Material and Methods and S1 Table for details). An N-terminal mitochondrial transit peptide was not predicted in BnALAS, but mitochondrial transit peptides have not been detected in any eukaryotic ALAS [57] examined so far, in spite of the fact that an N-terminal extension is apparent when the eukaryotic protein is aligned to bacterial homologs (not shown) and its amino acid composition resembles that of mitochondrial transit peptides. Moreover, ALAS has never been experimentally found outside of the mitochondrion, likely because it uses succinyl-CoA, a product of the mitochondrial citrate cycle, as the initial substrate [57,61]. The host gene for aminolevulinic acid dehydratase (ALAD) was lost and was likely replaced by a cyanobacterial (plastid) homolog (Fig 2); a predicted mitochondrial transit peptide at its N-terminus further supports a mitochondrial location (see S1 Table for details). BnPBGD2, responsible for the next step of the mitochondrial-cytosolic pathway, forms an unsupported but stable clade with other eukaryotic sequences, branching as sister to *E. gracilis* and oomycetes (non-photosynthetic stramenopiles). This clade, composed of animal, fungal, (phototrophic) excavate, heterotrophic stramenopile and chlorarachniophyte sequences, very likely represents the only PBGD enzymes originating in the eukaryotic nucleus because all phototrophic eukaryotes utilize PBGD of α -proteobacterial origin (S1 Fig) [50,56,57].

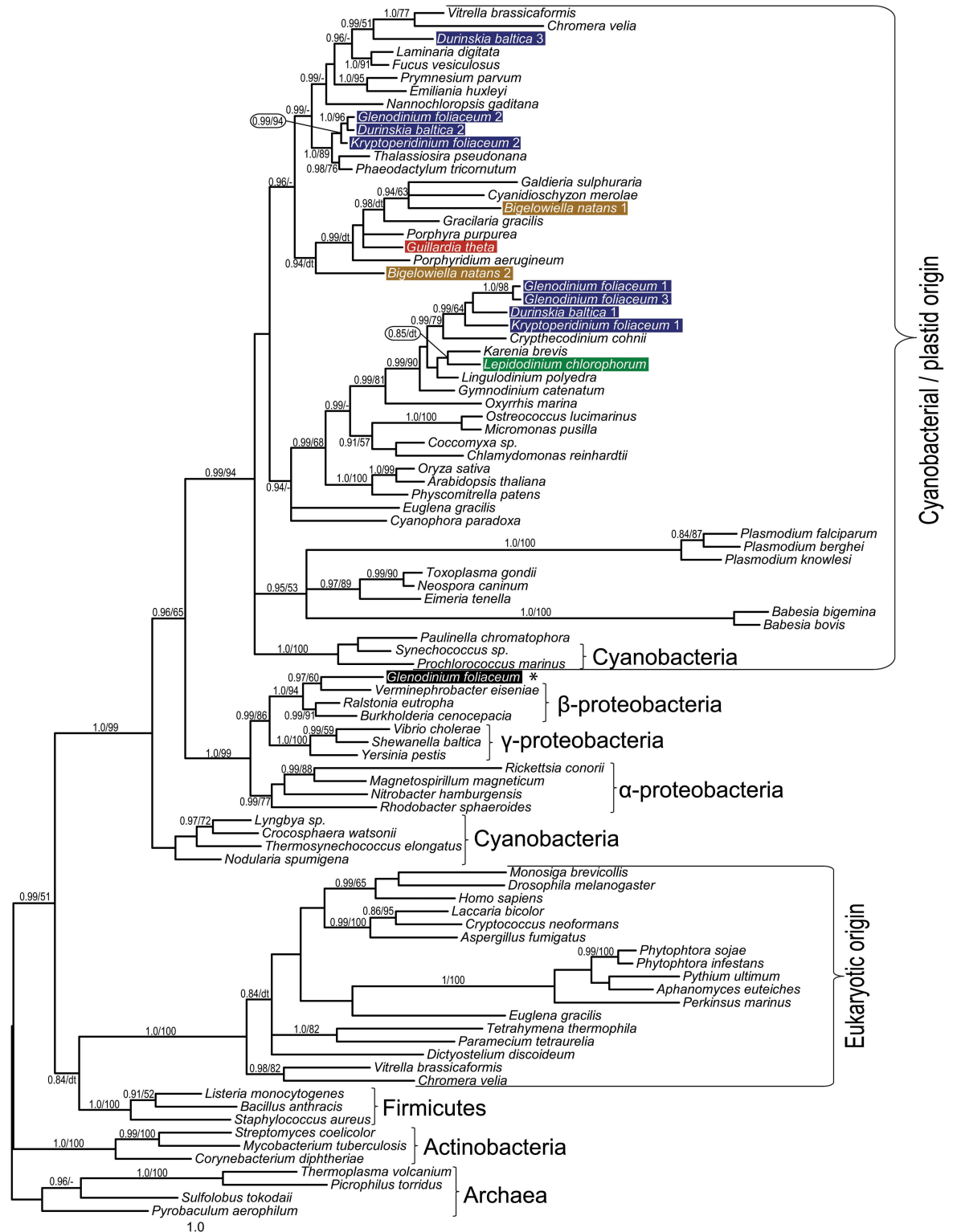


Fig 2. Bayesian phylogenetic tree as inferred from ALAD amino acid sequences. Taxa of interest in this study are highlighted by colored bars: blue for dinotoms, green for *Lepidodinium chlorophorum*, ochre for *Bigelowiella natans* and red for *Guillardia theta*. The tree shows red algal origin for *B. natans* and *G. theta* enzymes. For *L. chlorophorum*, we suggest a gene duplication / loss of paralogs scenario (see text); despite branching as sister to green algae, other dinoflagellates contained in the same clade do not possess a green algal plastid. Numbers near branches indicate Bayesian posterior probabilities followed

by the bootstrap of respective clades from the likelihood analysis. Only support values greater than 0.85 (Bayesian) and 50 (likelihood) are shown. dt—different topology in the ML tree, see [S2 Fig](#); a dash indicates an unsupported topology. An asterisk marks inferred bacterial contamination in *G. foliaceum* data.

doi:10.1371/journal.pone.0166338.g002

In spite of the presence of two tetrapyrrole pathways in the chlorarachniophyte, only a single gene coding for a putatively plastid-targeted uroporphyrinogen-III synthase (UROS) was found in the genome database. The enzyme possesses a bipartite targeting sequence at the N-terminus, necessary for delivering the protein into the secondary plastid ([Fig 1](#), [S1 Table](#)). Notably, all UROS genes in eukaryotic phototrophs form a single compact cluster, although the origin of this phototrophic clade is not clear ([S1 Fig](#)). Since we failed to find the cytosolic (heterotrophic) UROS in the genome, we can speculate that it was either not covered by the sequencing or annotation methods used or the recovered *B. natans* protein is dual-targeted to the cytosol and the plastid.

Most of the enzymes involved in the mitochondrial-cytosolic pathway are encoded by a single gene in *B. natans*, with the exception of UROD. All six genes coding for UROD display eukaryotic origin in *B. natans*; however, four of the UROD sequences (BnUROD1, -2, -4 and -5) are obvious multiple paralogs likely originating from the secondary host nucleus, while two paralogs originate from the endosymbiont (algal) nucleus (BnUROD3 and -6) ([Fig 1](#) and [S1 Fig](#)). According to predictions, three UROD enzymes (BnUROD1, -2, -3) are cytosolic, while three other URODs appear to be plastid-located, and heterotrophic enzyme BnUROD5 may have functionally replaced the cyanobacterial counterparts in the autotrophic pathway ([S1 Table](#)).

There are two genes coding for CPOX (BnCPOX2, -3) predicted to function within the heterotrophic pathway in *B. natans*; the former appeared within the clade composed of red-derived secondary algae, and the latter has a nuclear origin with an unsupported sister position to ciliate sequences ([S1 Fig](#)). A eukaryotic origin is also suggested for the putatively mitochondrion-located BnPPOX2 in *B. natans*. The eukaryotic clade is supported by Bayesian analysis in this case but its internal structure is not resolved, forming numerous polytomies ([S1 Fig](#)), with *B. natans* PPOX2 appearing as the earliest eukaryotic branch. The mitochondrial ferrochelatase 2 is derived from the chlorarachniophyte host and is phylogenetically affiliated with heterotrophic stramenopiles (oomycetes), the only representatives of the SAR group in this particular clade ([Fig 3](#)).

The autotrophic tetrapyrrole pathway displays mosaic origins in *B. natans*, similar to that in other eukaryotic phototrophs. It is mostly composed of cyanobacterial-derived enzymes (glutamate-tRNA reductase BnGTR, glutamate-1-semialdehyde 2,1-aminotransferase BnGSA-AT, BnALAD1, -2, BnPPOX1 and BnFeCH1), but also of enzymes likely originating from the endosymbiont (primary host) nucleus (BnUROD6, BnCPOX1), one enzyme displaying an α -proteobacterial (likely mitochondrial) origin (BnPBGD1), and an additional CPOX (BnCPOX4) of uncertain origin (see [Fig 1](#) for summary). Importantly, many of the aforementioned enzymes show unexpected phylogenetic affiliations: GTR, GSA-AT, ALAD1, CPOX1, -2 and PPOX1 in *B. natans* cluster with rhodophytes or algae with red secondary plastids (see [Figs 1, 2, 4](#) and [S1](#) for details), in spite of the chlorophyte origin of the current chlorarachniophyte plastid [[62](#)]. Only a single *B. natans* enzyme, one of the two ferrochelatases, displays the expected and supported chlorophyte origin (BnFeCH1, [Fig 3](#)). Although bootstrap values for the ML analyses are moderately supported, the hypothetical scenarios for red origin of many of these proteins are highly supported by Bayesian inference, which is more robust in analyses of data with high variability across sites. Furthermore, considering the good support for plastid genome relationship with green algae [[62](#)], one would expect the clear and well-supported association of *B. natans* sequences with either chlorophytes or heterotrophic eukaryotes; this is not observed.

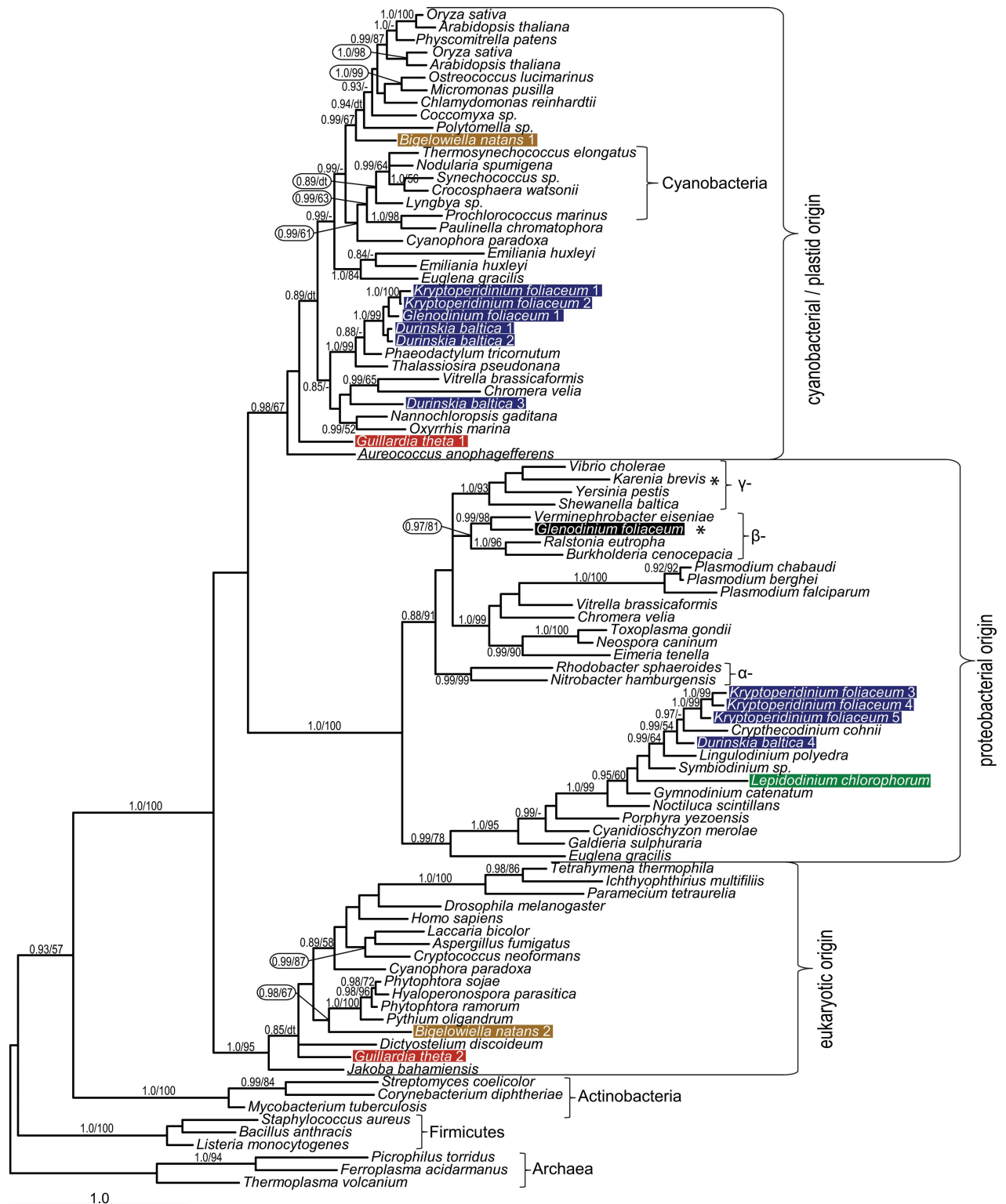


Fig 3. Bayesian phylogenetic tree as inferred from FeCH amino acid sequences. Taxa of interest in this study are highlighted by colored bars: blue for dinotoms, green for *Lepidodinium chlorophorum*, ochre for *Bigelowiella natans* and red for *Guillardia theta*. We document two orthologs, one of unresolved cyanobacterial origin and the other of eukaryotic origin, for *B. natans* and *G. theta* enzymes. The *L. chlorophorum* sequence branches together with other dinoflagellates, suggesting its origin lies in the peridinin plastid repertoire. Numbers near branches indicate Bayesian posterior probabilities followed by the bootstrap of respective clades from the likelihood

analysis. Only support values greater than 0.85 (Bayesian) and 50 (likelihood) are shown. dt—different topology in the ML tree, see [S2 Fig](#); a dash indicates unsupported topology. An asterisk marks inferred bacterial contamination in *G. foliaceum* and *Karenia brevis* data.

doi:10.1371/journal.pone.0166338.g003

Several enzymes of the autotrophic pathway are present in multiple copies in *B. natans*, namely ALAD ([Fig 2](#)), UROD and CPOX ([S1 Fig](#)). The genome of *B. natans* contains two, likely paralagous, genes encoding ALAD, at least one of them affiliated with the red lineage ([Fig 2](#)). Three genes coding for autotrophic CPOX have been found in the *B. natans* genome, and all three of them appear to be plastid-targeted ([S1 Table](#)): one gene (BnCPOX4) is recovered with sequences from other phototrophs with no support; the origin of this clade is unclear. BnCPOX1 appeared within the clade composed of red-derived secondary algae and appeared to be related neither to chlorophytes nor to rhodophytes ([S2 Fig](#)). We propose that these genes might originate from the endosymbiont (primary host) nucleus ([Figs 1 and S1](#)).

The putative cellular locations corresponding to all enzymes involved in tetrapyrrole biosynthesis in *B. natans* are shown in [S1 Table](#), as inferred using SignalP [63] and TargetP [64]. The entire autotrophic tetrapyrrole pathway is likely located in the plastid stroma of the chlorarachniophytes; no enzyme seems to be targeted to the periplastidal space with the possible exception of BnPBGD2, which fulfills some of the criteria described by Curtis et al. [31], namely high number of introns, D/K amino acids at the C terminus, and transit peptide net charge -1. The biological function of an isolated enzyme in this compartment, however, would be unclear and this targeting is unlikely. The plastid-origin BnALAD1 is putatively retargeted to the chlorarachniophyte mitochondrion where it is involved in the mitochondrial-cytosolic pathway. Conversely, paralogs of heterotrophic UROD4 and -5 appear to be retargeted to the plastid compartment ([Fig 1 and S1 Table](#)).

B. natans shares the analogous biparallel architecture of tetrapyrrole biosynthesis with photosynthetic euglenids. Plastid acquisition occurred relatively recently in euglenids, since phototrophic euglenids constitute a monophyletic group [36] and the previous presence of plastids was never confirmed in phagotrophic euglenids or their osmotrophic relatives [37]. In spite of the independent origins of chlorarachniophytes (Rhizaria) and euglenophytes (Excavata) [62], they display a similar pattern of tetrapyrrole synthesis [55]. Both algae possess two nearly complete pathways, one originating from the primary heterotrophic host, the other from the engulfed algal endosymbiont ([Fig 1](#)). In both algae the reduction of the redundant pathway has already begun and the mitochondrial-cytosolic pathway is partially reduced. In *Euglena*, two enzymes of plastid origin functionally replaced the original heterotrophic pathway genes for UROS and UROD, either via dual targeting or by sharing the products of the reactions they catalyze between compartments [55]. Similarly, in *Bigeloviella*, the original ALAD and UROS from the heterotrophic pathway were likely replaced via dual localization of the plastid-derived enzymes or through the exchange of pathway intermediates ([Fig 1](#)). Furthermore, one of the PBGD enzymes (possibly BnPBGD2) must be dually targeted for the heterotrophic pathway to function. As in other phototrophs, both *B. natans* and *E. gracilis* contain multiple copies (orthologs) of UROD and CPOX. In summary, the level of metabolic reduction is comparable in *Bigeloviella* and *Euglena*, which suggests that they acquired their green plastids at approximately the same time (assuming similar rates of evolution) or a constraint imposed on cellular metabolism that prevents a loss of redundancy in chlorarachniophytes.

The ultimate step of the cryptophyte pathway is bifurcated

With the exception of the ferredoxin predicted to localize in the mitochondrion (see below), only the set of enzymes originating from the algal endosymbiont and putatively targeted to the

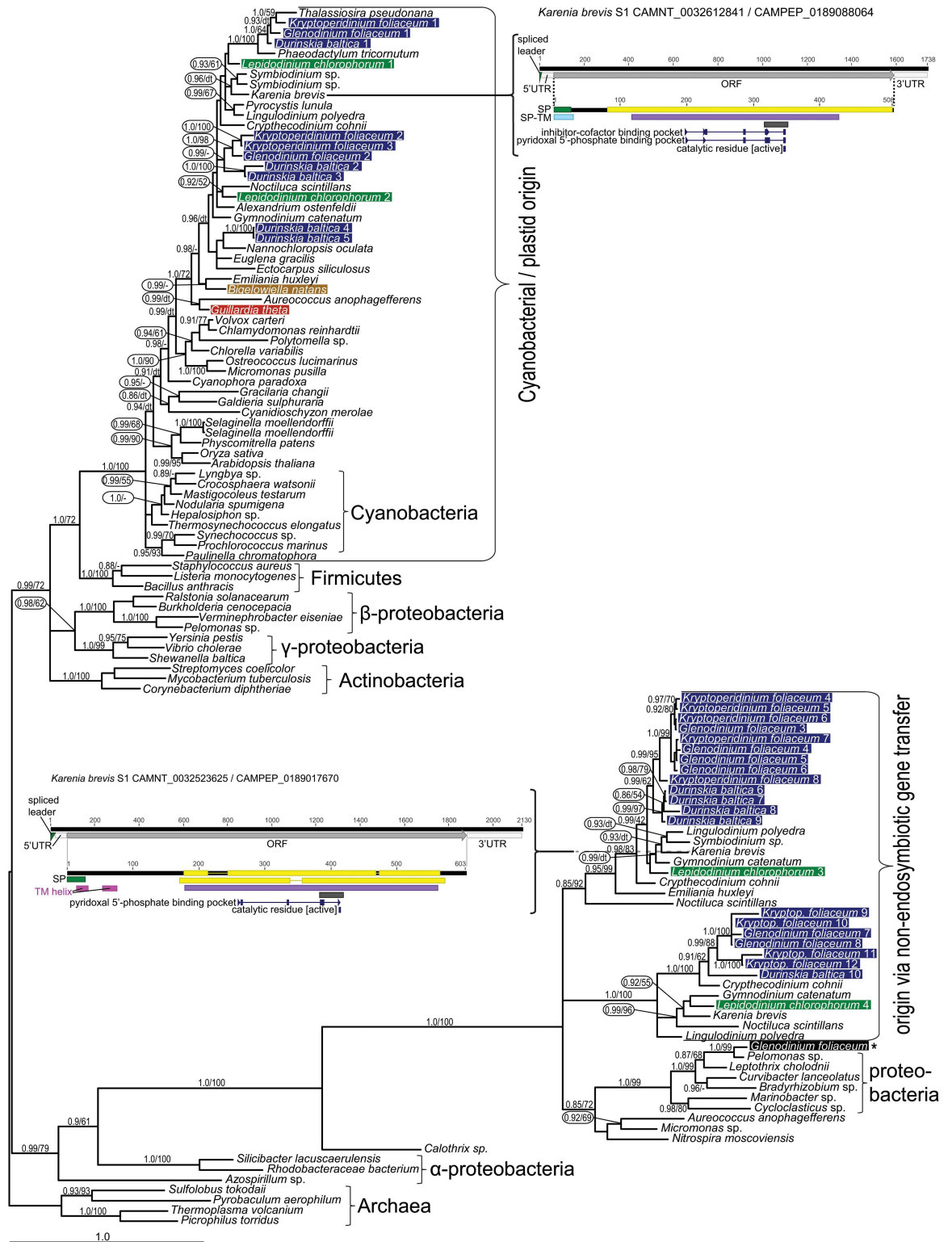


Fig 4. Bayesian phylogenetic tree as inferred from GSA-AT amino acid sequences. Taxa of interest in this study are highlighted by colored bars, blue for dinotoms, green for *Lepidodinium chlorophorum*, ochre for *Bigelowiella natans* and red for *Guillardia theta*. Numbers near branches indicate Bayesian posterior probabilities followed by the bootstrap of respective clades from the maximum likelihood (ML) analysis. Only support values greater than 0.85 (Bayesian) and 50 (ML) are shown. dt—different topology in the ML tree, see S2 Fig; a dash indicates unsupported topology. The tree demonstrates the

cyanobacterial origin of canonical GSA-AT, while the non-canonical GSA-AT originates in proteobacteria. Schematics of *Karenia brevis* transcripts and respective proteins are shown for complete representatives of canonical and non-canonical GSA-AT. The presence of a spliced-leader sequence at the 5' end suggests nuclear encoding and transcription of these genes. An N-terminal presequence of the resulting protein putatively targets both enzymes into the plastid. The canonical and non-canonical enzymes share motifs for pyridoxal 5'-phosphate binding and a catalytic residue. UTR—untranslated region; ORF—open reading frame; TM—transmembrane domain; SP—signal peptide; SP-TM—signal peptide predicted by the SignalP-TM networks; yellow bar—Panther Class III aminotransferase / glutamate-1-semialdehyde 2,1-aminomutase hit; violet bar—Pfam Class III aminotransferase hit; grey bar—PROSITE Class III aminotransferase hit; numbers represent scale in nt or aa.

doi:10.1371/journal.pone.0166338.g004

plastid compartment was found in *Guillardia theta* (Fig 1). We found three copies of UROD (one duplicated cyanobacterial gene and one gene originating from the endosymbiont nucleus) and CPOX (one duplicated gene from the endosymbiont nucleus and one gene of uncertain origin, GtCPOX2; see Figs 1 and S1 and S1 Table); both multiplied sets of enzymes are consequently combinations of orthologs originating from the primary endosymbiotic event (see S1 Fig for details), with one of them duplicated a second time. Gene duplication and functional specialization are also seen in other algae and plants [65]. While the duplication of the cyanobacterial UROD seems to be deeply branching, paralogs of CPOX arose more recently. Most of the enzymes involved originated from a cyanobacterial plastid ancestor related to rhodophytes and the CASH lineage (GtGTR1 and -2, GtGSA-AT, GtALAD, GtUROD1 and -2, GtPPOX; see Figs 2–4 and S1), supplemented by enzymes originating from the endosymbiont nucleus (UROD3, CPOX1 and -3), α -proteobacteria (mitochondria; GtPBGD) and enzymes encoded by algae-affiliated genes of unknown origins (GtUROS, GtCPOX2) (see overview in Fig 1). This is in line with the presence of the rhodophyte-derived secondary plastid in cryptophytes [66,67]. Clustering of GtUROD3 with homologs from green algae, *E. gracilis* and sequences from the CASH group (S1 Fig) may be the result of gene duplication followed by lineage-specific gene loss (discussed below).

The mitochondrion-located ferrochelatase GtFeCH2 is eukaryotic in origin, as are its homologs in heterotrophic eukaryotes and the glaucophyte *Cyanophora paradoxa* (Figs 1 and 4), and putatively targeted to the mitochondrion (S1 Table). The retention of mitochondrial ferrochelatase may be the result of slower rates of evolution in the cryptophyte plastid when compared to other red-derived secondary plastids [55–57,68], an evolutionary constraint placed on its role in the biology of the organism, or an independent and more recent plastid acquisition in cryptophytes. The latter view is consistent with the presence of a nucleomorph in cryptophytes and with the growing body of evidence showing that phototrophic cryptophytes emerged through an independent endosymbiosis event [18,69,70].

Novel type of GSA-AT and a proteobacterial FeCH in dinotoms

We analyzed transcriptomes from the dinotoms (dinoflagellates with tertiary diatom endosymbionts) *Glenodinium foliaceum* CCAP 1116/3, *Kryptoperidinium foliaceum* CCMP 1326 and *Durinskia baltica* (available via the MMETSP initiative, see S3 Table). Two redundant heme pathways are present in dinotoms: one is located in the diatom endosymbiont, while the second remains in the putative remnant of the original peridinin plastid, the eye spot [47]. These separate pathways seem to supply tetrapyrroles to the two independent symbiotic partners. Our inferred trees display similar topologies and evolutionary relationships as those published previously (see Fig 1 for summary; [47]); the endosymbiont pathway enzymes are related to sequences from free-living diatoms, while the host enzymes cluster together with other dinoflagellates. Furthermore, we identified a new biochemically

uncharacterized subfamily of putative GSA-AT, which is found in other dinoflagellates as well as in *Aureococcus anophagefferens*, *Emiliania huxleyi* and the chlorophyte *Micromonas*. Since the non-canonical putative GSA-AT clade contains a smaller clade composed of proteobacteria, the possibility of bacterial contamination has to be taken into account. However, spliced-leader sequences in the GSA-AT transcripts from *Karenia brevis* (CAMNT_0032523625 and CAMNT_0032609079, see Fig 4 for details) indicate that the genes are located in the dinoflagellate nucleus. The respective translated sequences contain N-terminal presequences putatively targeting the protein product to the peridinin plastid [71] (Figs 1 and 4). There is also a single sequence from *G. foliaceum* closely related to *Pelomonas* sp. in the bacterial cluster; this particular gene might be a bacterial contaminant (Fig 4). The origin of the novel GSA-AT clade remains unclear.

We also analyzed the last enzyme of the pathway, ferrochelatase (FeCH), not included in previously published analyses [47]. Ferrochelatase in particular shows complex origins in eukaryotic phototrophs, which includes non-endosymbiotic gene transfer from proteobacteria to the ancestor of chromerids and apicomplexans (Fig 3). Two clades in the ferrochelatase tree contain dinotoms: the genes of cyanobacterial origin came from the diatom endosymbiont; this clade also contains the heterotrophic dinoflagellate *Oxyrrhis marina*, numerous cyanobacteria, a glaucophyte, the rhizarian *Paulinella chromatophora*, chlorophytes, plants, heterokonts, eustigmatophytes, a haptophyte, *Euglena gracilis* and chromerids. The other clade is sister to apicomplexans, chromerids, and *E. gracilis*, and contains rhodophytes, peridinin-pigmented dinoflagellates, and a heterotrophic dinoflagellate (*Cryptheconidium cohni*). Its origin is unclear but might be proteobacterial (Fig 3). The tree topology is consistent with the presence of two ferrochelatases in chromerids [57], phototrophic euglenids [55], and dinotoms. While in apicomplexan parasites the complex origin of ferrochelatase is a result of non-endosymbiotic gene transfer from proteobacteria to Apicomplexa, the ferrochelatases in dinotoms arose from endosymbiotic association with the ancestor of the peridinin plastid and later tertiary endosymbiosis with the diatom endosymbiont.

The heme pathway is redundant in dinotoms; however, there are putatively necessary genes missing from their transcriptomes (Fig 1). The most striking absence is that of dinoflagellate-like UROS. We are unable to unambiguously discriminate between endosymbiont and host nuclear-encoded enzymes based solely on their sequences as spliced leaders are often missing; still, it appears that the diatom-like enzymes are exclusively used by the diatom endosymbiont, mainly due to retained characteristics required for diatom-like protein transport (the ASAFAP motif [47,72]) as well as difficulties in the hypothetical transport of proteins from the host cytoplasm over 6 membranes (endomembrane system of the host + putative plasma membrane of endosymbiont + four membranes of the diatom plastid). Conversely, any transport of UROS from inside the diatom endosymbiont compartment to the eyespot (remnant of the original dinoflagellate peridinin plastid) is hard to imagine. The absence of the original dinoflagellate UROS could be explained by transport of pre-uroporphyrinogen (hydroxymethylbilane), however, pre-uroporphyrinogen is highly unstable [73]. Furthermore, all the antecedent enzymes in the eyespot pathway (GTR, GSA-AT, ALAD, PBGD) would become redundant and therefore should have been lost from the genome. Insufficient sequencing and high divergence of UROS may explain the total absence of transcripts of the dinoflagellate-like UROS. On the other hand, the absence of *Glenodinium* orthologs of KfUROD1, KfUROD4 and the KfCPOX3+4+5 cluster and the *Kryptoperidinium* ortholog of the GfCPOX3+4 cluster (S1 Fig) may also be a result of gene loss, as other functional copies remained. In several cases, sequencing and assembly errors may interfere with determining the exact number of closely related paralogs, GSA-AT being an example of high gene copy number in dinoflagellates (Fig 4).

The rhodophyte pathway is conserved in algae with green plastids

Using our transcriptomic data (S2 Table), we mapped the tetrapyrrole pathway in the green dinoflagellate *Lepidodinium chlorophorum* [24,74]. We found most of the genes of the plastid tetrapyrrole pathway and no traces of the mitochondrial-cytosolic pathway (see Fig 1 for summaries), consistent with the autotrophic history of the species. Some enzymes (LcGTR, LcPPOX2 and LcFeCH) cluster with the red lineage (S1 Fig). Other sequences branch with green algae and plants but always also branch together with other dinoflagellate sequences (LcALAD, LcGSA-AT1, -2, LcPPOX1). For example, sequences of canonical GSA-AT from *L. chlorophorum* form a cluster with sequences from the red lineage (Fig 4); however streptophytes, chlorophytes and rhodophytes are unresolved and placed in the ancestral position (see Figs 4 and S1 for details).

Interpreting most phylogenetic trees is very complicated, mainly due to the existence of multiple genes originating in the host nucleus, the endosymbiont nucleus, cyanobacteria (plastid), and proteobacteria. Together these factors make the evolution of heme pathway enzymes difficult to follow, particularly in eukaryotic phototrophs. However, the phylogenetic placement of *L. chlorophorum* GSA-AT and other enzymes together with their orthologs from dinoflagellates with rhodophyte-derived plastids suggests a red origin. In LcALAD and LcPPOX, with affinities to green algae, we presume the topology could be the result of the duplication of cyanobacteria-derived genes and subsequent lineage-specific gene loss (Fig 2), similar to GtUROD3. In other words, the hypothetical ancestor of the plastid contained two paralogs, for clarity here denoted as red and green, and one these paralogs was later lost in each lineage (the green one in rhodophytes and stramenopiles, the red one in the green lineage and dinoflagellates), masking the true origin of the gene. A clear example is seen in PPOX: the gene passed through a duplication event (S1 Fig) and *L. chlorophorum* genes are present in both cyanobacterial PPOX clades, with either rhodophytes or chlorophytes at the root. Again, they group together with other algae possessing rhodophyte-derived plastids. This suggests the green paralog was inherited vertically, not via endosymbiotic gene transfer from the green endosymbiont. It is noteworthy that the paralogs of ALAD and PPOX retained in rhodophytes also remained in stramenopiles, haptophytes and dinotoms, while the genes found in green algae and plants are present only in dinoflagellates. Consequently, most of the “red-related” genes in dinotoms apparently originate from, and reside in, the diatom endosymbiont (Figs 1 and 2, S1) and the ancestral dinoflagellate genes appear “green-related”. An alternative hypothesis would imply horizontal (eukaryote-to-eukaryote) gene transfer of ALAD and PPOX from green algal prey or from a putative green plastid to the ancestor of dinoflagellates, to the exclusion of chromerids and apicomplexans that possess a red-related gene. This putative green plastid would appear cryptic from today’s perspective, as it must have been later functionally replaced and partially genetically masked by the current red-derived peridinin plastid.

Therefore, it appears that most of the enzymes considered here originate from a rhodophyte source, in spite of the dinoflagellate’s chlorophyte-derived plastid. The chlorophyte-derived plastid is thought to have replaced the original peridinin pigmented dinoflagellate plastid through serial secondary endosymbiosis in this species [23]. Regardless, we observed that the original rhodophyte-derived pathway introduced with the peridinin plastid is highly conserved in this dinoflagellate, in agreement with the “shopping bag” or plastid promiscuity hypothesis, resulting in a mosaic evolution of the plastid proteome [75]. This functional conservation of tetrapyrrole biosynthesis genes might result from a predisposition of red genes to be targeted to the new plastid (as they were already successfully targeted to the old one)—an advantage the newcomer green genes presumably lacked.

The predominantly rhodophyte origins of the heme pathway in *L. chlorophorum* likely represent a set from the previously acquired peridinin plastid, but the “purchased” old genes were put into a newer, better shopping bag. The presence of red-derived heme pathway enzymes in *B. natans* plastids could be a result of non-endosymbiotic HGT from (red) algal prey according to the “you are what you eat” hypothesis [76]. Indeed, a rich fraction of red-related genes in *B. natans*, including photosynthesis-related genes, has already been reported [77,78]. Additional photosynthesis-related proteins were recruited from bacteria, indicating that extensive HGT does take place in chlorarachniophytes [77]. One explanation for the ease of HGT in this case comes from the origin of heme pathway genes; *B. natans* sequences often cluster with the CASH taxa. If these genes were horizontally transferred from one of the CASH lineages, they already coded for some plastid targeting signal as heme synthesis is presumably plastid-located in these lineages. *Bigelowiella natans* does not use the same protein import complex (SELMA) as algae from the CASH taxa [79], but plastid proteins in these groups have some structural similarities (e.g. the presence of a bipartite N terminal extension comprising a signal peptide plus a transit peptide) suggesting similarities also in the protein transport mechanisms. On the other hand, the abovementioned HGT events from the red lineage must have taken place after the green plastid acquisition but before the divergence of two basal chlorarachniophyte lineages comprising *Lotharella amoebiformis* and *B. natans* [78]. Altogether, it seems less probable that massive gene replacement via non-endosymbiotic HGT would occur in enzymes forming an essential and compartmentalized metabolic pathway. Taking into account similarities with *L. chlorophorum*, we can speculate that the green dinoflagellate “heritage” scenario also applies to *B. natans*. The rhodophyte origin of the tetrapyrrole pathway in *B. natans* may therefore reflect the previous presence of a hypothetical red-derived plastid in the ancestor of chlorarachniophytes (Fig 5). Rhizarians exhibit predatory heterotrophic or parasitic lifestyles, and a cryptic plastid could be held initially as a kleptoplast [80]. Similarly in dinoflagellates, the origins of “green” ALAD and PPOX may trace back to a cryptic endosymbiosis or gene transfer from the green lineage in the common ancestor of extant dinoflagellates including *Oxyrrhis* (Fig 2), mirroring the gene flow from the red lineage observed in *B. natans* [78]. However, the number of genes significantly related to the green lineage is extremely limited in studied dinoflagellates and *Chromera velia* [33,81,82] and does not necessarily imply the cryptic introduction of a green plastid. Indeed, the observed topologies might be artifacts showing false phylogenetic affinities. However, considering balanced sampling of each higher taxon (CASH taxa, green algae plus plants, red algae), this is less likely to happen in all cases.

In order to determine how endosymbiotic events occurred, a robust reconstruction of the gene repertoire of photosynthetic algae is needed. Not all genes diverged at the same time and gene multiplication and lineage-specific losses may hinder phylogenetic signal resolution. Indeed, genes with conserved or ancient evolutionary histories display different topologies than those acquired more recently via HGT [34] or those possessing less conserved functions [83]. Curtis et al. [31] reported a high number of algal-related genes in *G. theta* that acquired new functions and putatively also cellular localizations through endosymbiotic gene transfer to the host nucleus, regardless of their evolutionary origin. This is in conflict with the conserved origins implicated in this study; we suggest that enzymes of essential plastid pathways, such as tetrapyrrole biosynthesis, resist functional replacement due to conserved localization to a specialized compartment. The protein transport mechanism (SELMA), present in all investigated CASH taxa, is strong molecular evidence for the monophyletic origin of the CASH plastid [5,6,14]. During the course of evolution, proteins transported into the plastid via this mechanism have acquired an N-terminal transport signal. This potentially enables their lateral movement to new eukaryote hosts and allows them to maintain their original functions inside the organelle, provided the same transport mechanism is employed in the new host.

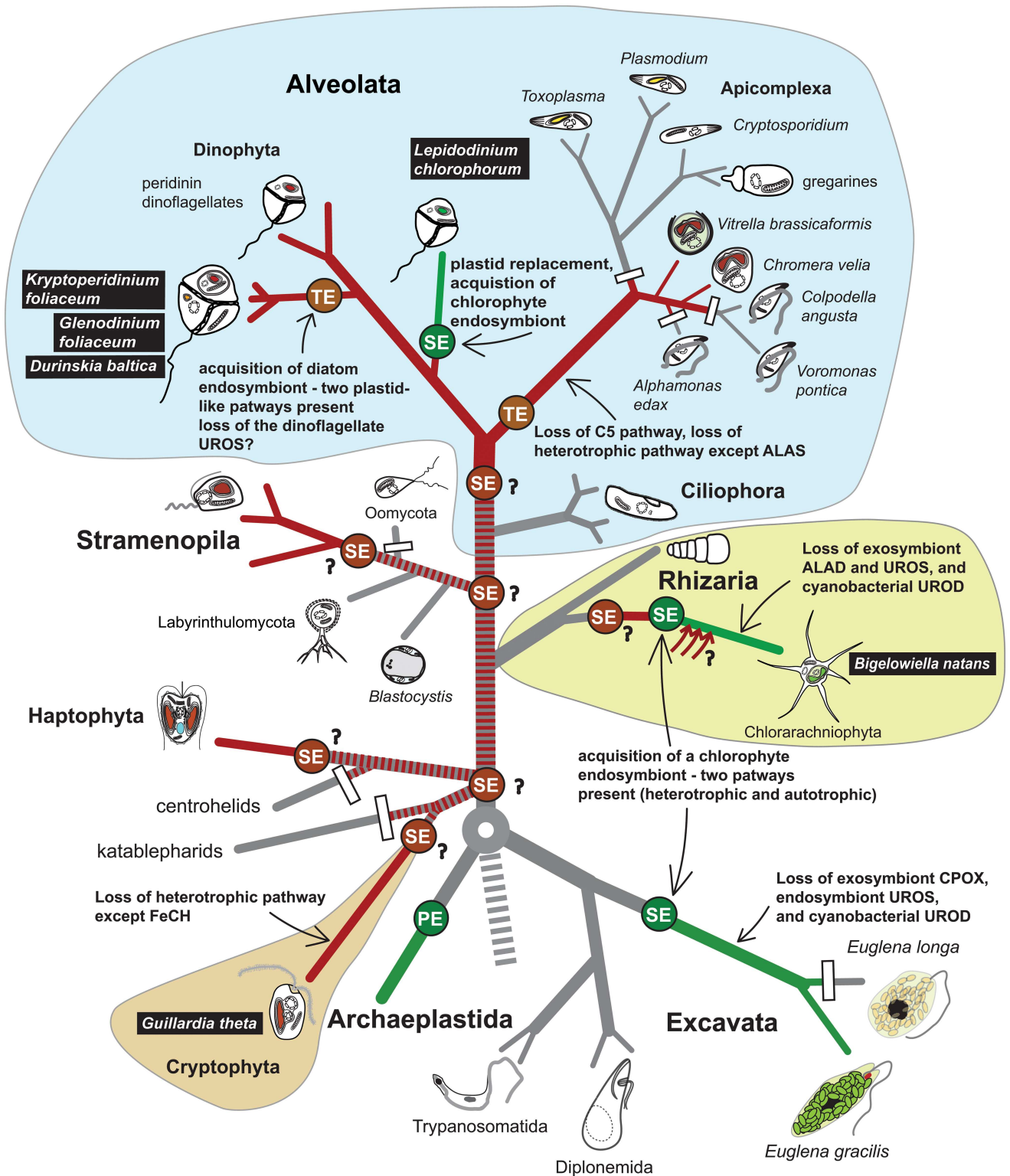


Fig 5. A simplified scheme of the evolution of the tetrapyrrole biosynthesis pathway with an emphasis on the models from this study (black bars). Primary endosymbiosis (PE) gave rise to the Archaeplastida comprising red algae, green algae and glaucophytes. Following the divergence of main eukaryotic lineages, secondary (SE) or tertiary endosymbiosis (TE) events equipped the ancestors of CASH taxa (cryptophytes, alveolates, stramenopiles and haptophytes) with photosynthetic capabilities. Contradictory evidence has been debated over the last years as for the history of CASH plastid acquisitions (e.g. [5,8,34]). A plastid-early scenario (the chromalveolate hypothesis) posits that all

CASH taxa are monophyletic and the CASH plastid was vertically transferred (dashed red line) and lost in extant non-photosynthetic descendants (such as ciliates and most rhizarians). Plastid-late scenarios require multiple lateral acquisitions of the CASH plastid (question marks) but better reflect some current phylogenomic analyses of the plastid recipients (e.g. [18,69]). Loss of photosynthesis/plastids have been documented in many sister lineages, such as oomycetes or apicomplexans, however these are in line with plastid-late scenarios as well. A cryptic SE with a CASH alga or numerous HGT (red arrows) events are inferred before the divergence of extant chlorarachniophytes (this work, [78]), which was masked by the acquisition of the current green algal endosymbiont. A similar situation in *L. chlorophorum* led to the peridinin plastid replacement with a green plastid, however the majority of red-related heme pathway enzymes remained functionally conserved in the successor plastid. The loss of the heterotrophic pathway possibly occurred several times independently in the stramenopile and dinoflagellate lineages, as *Perkinsus marinus*, sister to dinoflagellates, still contains a functional mitochondrial-cytosolic pathway [8].

doi:10.1371/journal.pone.0166338.g005

This is consistent with theories of lateral CASH plastid transfer into unrelated branches of the eukaryote tree, giving rise to the CASH taxa [34]. Conversely in cases of serial plastid replacement, a potential new plastid would encounter a pre-existing, functional set of proteins ensuring the function of the original organelle. The new plastid might also be inhabited by the original plastid's protein compendium including the protein transport machinery, rather than continuing to use its own proteome, which would be incompatible with SELMA. With this in mind, we presume that proteins having the ability to be transported to the CASH plastid could also be transported into a successor plastid, enabling the detection of serial endosymbiotic histories with higher confidence than cytoplasmic proteins.

Conclusions

The tetrapyrrole biosynthesis pathway in phototrophic eukaryotes is an evolutionary mosaic originating in proteobacteria, cyanobacteria and eukaryotes. It represents a shopping bag of enzymes collected during the history of plastid endosymbiosis retained, due to its essential role in metabolism, even after photosynthetic capabilities have been lost. Here we confirm that the tertiary plastids of dinotoms represent largely independent compartments with tetrapyrrole biosynthesis occurring parallel to biosynthesis in the peridinin plastid. The enzymes putatively localized to the former plastid branch sister to dinoflagellate enzymes, while the tertiary plastid contains enzymes branching sister to those of diatoms, mirroring the origin of the respective organelles. In *G. theta*, the pathway is located almost entirely in the plastid, with the exception of a eukaryotic ferrochelatase apparently localized to the mitochondrion, indicating either a slow evolutionary rate or an evolutionary constraint. Furthermore, we observed that the majority of the pathway is evolutionarily conserved and related to the red lineage even in organisms that currently possess a plastid of green algal provenance, i.e. the dinoflagellate *Lepidodinium chlorophorum* and the chlorarachniophyte *B. natans*. Hence, if the protein targeting machinery is compatible with the new plastid compartment, the tetrapyrrole synthesis pathway can be relocated "as is", which is illustrated in the case of *L. chlorophorum*. Intriguingly, such a scenario may imply the existence of a cryptic red-derived plastid earlier in the history of chlorarachniophytes. While the evolution of eukaryotes is becoming clearer with increasing data from deeper lineages, the history of plastid acquisitions resists revealing an unequivocal scenario due to massive gene transfer and phylogenetic bias. We suggest that a targeted approach directed at conserved processes could result in new, relevant hypotheses even in the genomic era.

Material and Methods

The complete genomic sequences of the cryptophyte alga *Guillardia theta* (<http://genome.jgi.doe.gov/Guith1/Guith1.home.html>) and the chlorarachniophyte *Bigeloviella natans* (<http://genome.jgi.doe.gov/Big1a1/Big1a1.home.html>) were searched using BLAST [84] for genes encoding enzymes involved in the synthesis of tetrapyrroles (ALAS, GTR, GSA-AT, ALAD,

PBGD, UROS, UROD, CPOX, PPOX, and FeCH). Homologous sequences were those used in Kořený *et al.* [57]; newly added sequences are listed in S3 Table. All alignments were made using MUSCLE [85] and ambiguous regions were removed in SeaView [86]. Phylogenetic trees were constructed using Maximum Likelihood (RAxML v8.2.4; [87]), Bayesian inference (PHYLOBAYES v3.3b; [88]) and a method designed to deal with amino acid saturation (AsaturA v18.10.2002; [89]). ML trees were computed using the LG model with gamma distribution in 4 categories and 1000 replicates. Bayesian inferences were calculated with the following parameters: 2 chains, 15,000 generations under the C20 model with Poisson exchange rate, sampling every 100 generations, and a maximum divergence of 0.1.

Sequences from *G. theta* and *B. natans* were inspected for the presence of N-terminal leader sequences by SignalP 3.0 [63] and TargetP [64], predicting localization to either the mitochondrion (mitochondrial transit peptide) or the plastid (bipartite leader composed of ER signal peptide followed by a transit peptide). GSA-AT of *Karenia brevis* (Fig 4) were automatically annotated using the InterProScan feature of Geneious 8.1 [90].

The transcriptome library of *Lepidodinium chlorophorum* (Roscoff collection no. RCC1488) was generated using the NEBNext Ultra Directional RNA Library kit (New England Biolabs, Ipswich, MA, USA) according to the manufacturer's instructions. Quality assessment and sequencing were performed in a specialized facility, using the Illumina MiSeq (2×250 bp) platform. The generated reads were quality-trimmed using the FASTQ Toolkit (v1.0) of the Illumina BaseSpace platform and then assembled using Trinity v2.1.1 [91] and SOAPdenovo2 v2.0 r240 [92] and clustered using CAP3 [93]. Gene assembly completion was assessed with BUSCO software using the complete eukaryotic gene profile [94] on protein models generated by the TransDecoder script of the Trinity package. Some characteristics of the obtained transcriptome are listed in S2 Table. Novel sequences of interest were deposited in GenBank under the accession no. KX344033-47.

Supporting Information

S1 Fig. Bayesian phylogenetic trees as inferred from the amino acid sequences. A) ALAS, B) GTR, C) PBGD, D) UROS, E) UROD, F) CPOX and G) PPOX. Taxa of interest of this study are highlighted by colored bars: blue for dinotoms, green for *Lepidodinium chlorophorum*, ochre for *Bigelowiella natans* and red for *Guillardia theta*. The tree demonstrates the mitochondrial origin of ALAS. Numbers near branches indicate Bayesian posterior probabilities followed by the bootstrap of respective clades from the likelihood analysis. Only support values greater than 0.85 (Bayesian) and 50 (likelihood) are shown. dt—different topology in the likelihood tree, see S2 Fig; a dash indicates unsupported topology. Asterisks mark possible contaminations. LcPPOXa, -b, -c; LcUROSa, -b = non-overlapping protein models, putatively fragments of LcPPOX1 and LcUROS.
(PDF)

S2 Fig. Maximum likelihood trees as inferred from amino acid sequences. Numbers near branches indicate bootstrap values; only support values greater than 50 are shown. A, ALAS—delta-aminolevulinic acid synthase; B, GTR—glutamate-tRNA reductase; C, GSA—glutamate-1-semialdehyde aminotransferase; D, ALAD—aminolevulinic acid dehydratase; E, PBGD—porphobilinogen deaminase; F, UROS—uroporphyrinogen synthase; G, UROD—uroporphyrinogen decarboxylase; H, CPOX—coproporphyrinogen oxidase; I, PPOX—protoporphyrinogen oxidase; J, FeCH—ferrochelatase. LcPPOXa, -b, -c; LcUROSa, -b = non-overlapping protein models, putatively fragments of LcPPOX1 and LcUROS.
(PDF)

S1 Table. Targeting presequences in *Bigeloviella natans* and *Guillardia theta*. Targeting probabilities were determined using SignalP and TargetP as described in Materials and Methods. Respective targeting peptide sequences are listed. The presence of a signal peptide followed by a chloroplast targeting peptide (cTP) implies localization to the plastid; mitochondrial enzymes encode mitochondrial targeting peptide presequences, while cytoplasmic enzymes lack presequences. If available, models with longer N-termini (e.g. Pasa, Fgenesh) were included in pre-sequence analysis and are listed in the table.
(PDF)

S2 Table. Characteristics of the obtained transcriptome libraries of *Lepidodinium chlorophorum*. Number of reads and bases of two libraries are listed after quality trimming (two reads per library, see [Material and Methods](#)). The resulting number of contigs and coding sequences were analyzed using the BUSCO pipeline with a set of 429 BUSCO groups of orthologs. Ortholog counts: C = complete; D = duplicated; F = fragments; M = missing.
(PDF)

S3 Table. List of sequences added to the original datasets of Kořený *et al.* [26, 28]. Gene copy designation for species of interest in this study is shown in brackets according to their designation in respective trees (an asterisk marks putative contaminant sequences). Databases: CGP = Cyanophora Genome Project; CryptoDB = Cryptosporidium Genomic Resource; gb = GenBank; Gruber *et al.* 2015 Plant J, 10.1111/tpj.12734; jgi = DOE Joint Genome Institute; MMETSP = Marine Microbial Eukaryote Transcriptome Sequencing Project; Nori = NoriBLAST, Porphyra Genome Project; psb = bioinformatics.psb.ugent.be; VH = Courtesy of Vladimír Hampl, unpublished; Cmb = combined samples.
(PDF)

Acknowledgments

We acknowledge computation resources provided by CERIT-SC and MetaCentrum, Brno, Czech Republic. We thank Heather J. Esson for language correction, Luděk Kořený for helpful discussions and John M. Archibald for providing original sequence data before their public release via JGI.

Author Contributions

Conceptualization: MO.

Investigation: JC ZF AH.

Methodology: MO.

Visualization: JC ZF MO.

Writing – original draft: JC ZF MO.

References

1. McFadden GI, Guy L, Saw JH, Ettema TJG, Eme L, Sharpe SC, et al. Origin and evolution of plastids and photosynthesis in eukaryotes. *Cold Spring Harb Perspect Biol.* 2014; 6. doi: [10.1101/cshperspect.a016105](https://doi.org/10.1101/cshperspect.a016105) PMID: [24691960](https://pubmed.ncbi.nlm.nih.gov/24691960/)
2. Nowack ECM, Melkonian M, Glöckner G. Chromatophore genome sequence of *Paulinella* sheds light on acquisition of photosynthesis by eukaryotes. *Curr Biol.* 2008; 18: 410–418. doi: [10.1016/j.cub.2008.02.051](https://doi.org/10.1016/j.cub.2008.02.051) PMID: [18356055](https://pubmed.ncbi.nlm.nih.gov/18356055/)

3. Keeling PJ. The number, speed, and impact of plastid endosymbioses in eukaryotic evolution. *Annu Rev Plant Biol.* 2013; 64: 583–607. doi: [10.1146/annurev-arplant-050312-120144](https://doi.org/10.1146/annurev-arplant-050312-120144) PMID: [23451781](https://pubmed.ncbi.nlm.nih.gov/23451781/)
4. Stiller JW, Schreiber J, Yue J, Guo H, Ding Q, Huang J. The evolution of photosynthesis in chromist algae through serial endosymbioses. *Nat Commun.* 2014; 5: 5764. doi: [10.1038/ncomms6764](https://doi.org/10.1038/ncomms6764) PMID: [25493338](https://pubmed.ncbi.nlm.nih.gov/25493338/)
5. Zimorski V, Ku C, Martin WF, Gould SB. Endosymbiotic theory for organelle origins. *Curr Opin Microbiol.* 2014; 22: 38–48. doi: [10.1016/j.mib.2014.09.008](https://doi.org/10.1016/j.mib.2014.09.008) PMID: [25306530](https://pubmed.ncbi.nlm.nih.gov/25306530/)
6. Gould SB, Maier U-G, Martin WF. Protein import and the origin of red complex plastids. *Curr Biol.* 2015; 25: R515–R521. doi: [10.1016/j.cub.2015.04.033](https://doi.org/10.1016/j.cub.2015.04.033) PMID: [26079086](https://pubmed.ncbi.nlm.nih.gov/26079086/)
7. Archibald JM. Genomic perspectives on the birth and spread of plastids. *Proc Natl Acad Sci U S A.* 2015; 112: 10147–53. doi: [10.1073/pnas.1421374112](https://doi.org/10.1073/pnas.1421374112) PMID: [25902528](https://pubmed.ncbi.nlm.nih.gov/25902528/)
8. Waller RF, Gornik SG, Kořený L, Pain A. Metabolic pathway redundancy within the apicomplexan-dinoflagellate radiation argues against an ancient chromalveolate plastid. *Commun Integr Biol.* 2016; 9. doi: [10.1080/19420889.2015.1116653](https://doi.org/10.1080/19420889.2015.1116653) PMID: [27066182](https://pubmed.ncbi.nlm.nih.gov/27066182/)
9. Cavalier-Smith T. Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. *J Eukaryot Microbiol.* 1999; 46: 347–366. doi: [10.1111/j.1550-7408.1999.tb04614.x](https://doi.org/10.1111/j.1550-7408.1999.tb04614.x) PMID: [18092388](https://pubmed.ncbi.nlm.nih.gov/18092388/)
10. Lane CE, Archibald JM. The eukaryotic tree of life: endosymbiosis takes its TOL. *Trends Ecol Evol.* 2008; 23: 268–275. doi: [10.1016/j.tree.2008.02.004](https://doi.org/10.1016/j.tree.2008.02.004) PMID: [18378040](https://pubmed.ncbi.nlm.nih.gov/18378040/)
11. Sanchez-Puerta MV, Delwiche CF. A hypothesis for plastid evolution in chromalveolates. *J Phycol.* 2008; 44: 1097–1107. doi: [10.1111/j.1529-8817.2008.00559.x](https://doi.org/10.1111/j.1529-8817.2008.00559.x) PMID: [27041706](https://pubmed.ncbi.nlm.nih.gov/27041706/)
12. Bodyl A, Stiller JW, Mackiewicz P. Chromalveolate plastids: direct descent or multiple endosymbioses? *Trends Ecol Evol.* 2009; 24: 119–121. doi: [10.1016/j.tree.2008.11.003](https://doi.org/10.1016/j.tree.2008.11.003) PMID: [19200617](https://pubmed.ncbi.nlm.nih.gov/19200617/)
13. Janouškovec J, Horák A, Oborník M, Lukeš J, Keeling PJ. A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids. *Proc Natl Acad Sci U S A.* 2010; 107: 10949–54. doi: [10.1073/pnas.1003335107](https://doi.org/10.1073/pnas.1003335107) PMID: [20534454](https://pubmed.ncbi.nlm.nih.gov/20534454/)
14. Felsner G, Sommer MS, Gruenheit N, Hempel F, Moog D, Zauner S, et al. ERAD components in organisms with complex red plastids suggest recruitment of a preexisting protein transport pathway for the periplastid membrane. *Genome Biol Evol.* 2011; 3: 140–150. doi: [10.1093/gbe/evq074](https://doi.org/10.1093/gbe/evq074) PMID: [21081314](https://pubmed.ncbi.nlm.nih.gov/21081314/)
15. Hampf V, Hug L, Leigh JW, Dacks JB, Lang BF, Simpson AGB, et al. Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic “supergroups”. *Proc Natl Acad Sci U S A.* 2009; 106: 3859–64. doi: [10.1073/pnas.0807880106](https://doi.org/10.1073/pnas.0807880106) PMID: [19237557](https://pubmed.ncbi.nlm.nih.gov/19237557/)
16. Brown MW, Sharpe SC, Silberman JD, Heiss A a, Lang BF, Simpson AGB, et al. Phylogenomics demonstrates that breviate flagellates are related to opisthokonts and apusomonads. *Proc Biol Sci.* 2013; 280: 20131755. doi: [10.1098/rspb.2013.1755](https://doi.org/10.1098/rspb.2013.1755) PMID: [23986111](https://pubmed.ncbi.nlm.nih.gov/23986111/)
17. Yabuki A, Kamikawa R, Ishikawa S a, Kolisko M, Kim E, Tanabe AS, et al. *Palpitomonas bilix* represents a basal cryptist lineage: insight into the character evolution in Cryptista. *Sci Rep.* 2014; 4: 4641. doi: [10.1038/srep04641](https://doi.org/10.1038/srep04641) PMID: [24717814](https://pubmed.ncbi.nlm.nih.gov/24717814/)
18. Burki F, Kaplan M, Tikhonenkov DV, Zlatogursky V, Minh BQ, Radaykina LV, et al. Untangling the early diversification of eukaryotes: a phylogenomic study of the evolutionary origins of Centrohelida, Haptophyta and Cryptista. *Proc R Soc B Biol Sci.* 2016; 283: 20152802. doi: [10.1098/rspb.2015.2802](https://doi.org/10.1098/rspb.2015.2802) PMID: [26817772](https://pubmed.ncbi.nlm.nih.gov/26817772/)
19. Delwiche CF. Tracing the thread of plastid—Diversity through the tapestry of life. *Am Nat.* 1999; 154: S164–S177. Available: <http://www.jstor.org/stable/10.2307/2463984> doi: [10.1086/303291](https://doi.org/10.1086/303291) PMID: [10527925](https://pubmed.ncbi.nlm.nih.gov/10527925/)
20. Inagaki Y, Dacks JB, Ford Doolittle W, Watanabe KI, Ohama T. Evolutionary relationship between dinoflagellates bearing obligate diatom endosymbionts: Insight into tertiary endosymbiosis. *Int J Syst Evol Microbiol.* 2000; 50: 2075–2081. doi: [10.1099/00207713-50-6-2075](https://doi.org/10.1099/00207713-50-6-2075) PMID: [11155982](https://pubmed.ncbi.nlm.nih.gov/11155982/)
21. Saldarriaga JF, Taylor FJR, Keeling PJ, Cavalier-smith T. Dinoflagellate nuclear SSU rRNA phylogeny suggests multiple plastid losses and replacements. *J Mol Evol.* 2001; 53: 204–213. doi: [10.1007/s002390010210](https://doi.org/10.1007/s002390010210) PMID: [11523007](https://pubmed.ncbi.nlm.nih.gov/11523007/)
22. Ishida K, Green BR. Second- and third-hand chloroplasts in dinoflagellates: Phylogeny of oxygen-evolving enhancer 1 (PsbO) protein reveals replacement of a nuclear-encoded plastid gene by that of a haptophyte tertiary endosymbiont. *Proc Natl Acad Sci U S A.* 2002; 99: 9294–9299. doi: [10.1073/pnas.142091799](https://doi.org/10.1073/pnas.142091799) PMID: [12089328](https://pubmed.ncbi.nlm.nih.gov/12089328/)
23. Keeling PJ. The endosymbiotic origin, diversification and fate of plastids. *Philos Trans R Soc Lond B Biol Sci.* 2010; 365: 729–48. doi: [10.1098/rstb.2009.0103](https://doi.org/10.1098/rstb.2009.0103) PMID: [20124341](https://pubmed.ncbi.nlm.nih.gov/20124341/)

24. Elbrachter M, Schnepf E. *Gymnodinium chlorophorum*, a new, green, bloom-forming dinoflagellate (Gymnodiniales, Dinophyceae) with a vestigial prasinophyte endosymbiont. *Phycologia*. 1996; 35: 381–393. doi: [10.2216/I0031-8884-35-5-381.1](https://doi.org/10.2216/I0031-8884-35-5-381.1)
25. Takishita K, Kawachi M, Noel MH, Matsumoto T, Kakizoe N, Watanabe MM, et al. Origins of plastids and glyceraldehyde-3-phosphate dehydrogenase genes in the green-colored dinoflagellate *Lepidodinium chlorophorum*. *Gene*. 2008; 410: 26–36. doi: [10.1016/j.gene.2007.11.008](https://doi.org/10.1016/j.gene.2007.11.008) PMID: [18191504](https://pubmed.ncbi.nlm.nih.gov/18191504/)
26. Chesnick JM, Morden CW, Schmiegel AM. Identity of the endosymbiont of *Peridinium foliaceum* (Pyrrophyta): Analysis of the rbcLS operon. *J Phycol*. 1996; 32: 850–857. doi: [10.1111/J.0022-3646.1996.00850.X](https://doi.org/10.1111/J.0022-3646.1996.00850.X)
27. Imanian B, Keeling PJ. The dinoflagellates *Durinskia baltica* and *Kryptoperidinium foliaceum* retain functionally overlapping mitochondria from two evolutionarily distinct lineages. *BMC Evol Biol*. 2007; 7: 172. doi: [10.1186/1471-2148-7-172](https://doi.org/10.1186/1471-2148-7-172) PMID: [17892581](https://pubmed.ncbi.nlm.nih.gov/17892581/)
28. Moustafa A, Beszteri BB, Maier UG, Bowler C, Valentin K, Bhattacharya D. Genomic footprints of a cryptic plastid endosymbiosis in diatoms. *Science*. 2009; 324: 1724–6. doi: [10.1126/science.1172983](https://doi.org/10.1126/science.1172983) PMID: [19556510](https://pubmed.ncbi.nlm.nih.gov/19556510/)
29. Dorrell RG, Smith AG. Do red and green make brown?: Perspectives on plastid acquisitions within chromalveolates. *Eukaryot Cell*. 2011; 10: 856–868. doi: [10.1128/EC.00326-10](https://doi.org/10.1128/EC.00326-10) PMID: [21622904](https://pubmed.ncbi.nlm.nih.gov/21622904/)
30. Woehle C, Dagan T, Martin WF, Gould SB. Red and problematic green phylogenetic signals among thousands of nuclear genes from the photosynthetic and apicomplexa-related *Chromera velia*. *Genome Biol Evol*. 2011; 3: 1220–1230. doi: [10.1093/gbe/evr100](https://doi.org/10.1093/gbe/evr100) PMID: [21965651](https://pubmed.ncbi.nlm.nih.gov/21965651/)
31. Curtis BA, Tanifuji G, Burki F, Gruber A, Irimia M, Maruyama S, et al. Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. *Nature*. 2012; 492: 59–65. doi: [10.1038/nature11681](https://doi.org/10.1038/nature11681) PMID: [23201678](https://pubmed.ncbi.nlm.nih.gov/23201678/)
32. Deschamps P, Moreira D. Reevaluating the green contribution to diatom genomes. *Genome Biol Evol*. 2012; 4: 683–688. doi: [10.1093/gbe/evs053](https://doi.org/10.1093/gbe/evs053) PMID: [22684208](https://pubmed.ncbi.nlm.nih.gov/22684208/)
33. Burki F, Flegontov P, Oborník M, Cihlář J, Pain A, Lukeš J, et al. Re-evaluating the green versus red signal in eukaryotes with secondary plastid of red algal origin. *Genome Biol Evol*. 2012; 4: 626–635. doi: [10.1093/gbe/evs049](https://doi.org/10.1093/gbe/evs049) PMID: [22593553](https://pubmed.ncbi.nlm.nih.gov/22593553/)
34. Petersen J, Ludewig AK, Michael V, Bunk B, Jarek M, Baurain D, et al. *Chromera velia*, endosymbioses and the rhodoplex hypothesis—Plastid evolution in cryptophytes, alveolates, stramenopiles, and haptophytes (CASH lineages). *Genome Biol Evol*. 2014; 6: 666–684. doi: [10.1093/gbe/evu043](https://doi.org/10.1093/gbe/evu043) PMID: [24572015](https://pubmed.ncbi.nlm.nih.gov/24572015/)
35. Ševčíková T, Horák A, Klimeš V, Zbránková V, Demir-Hilton E, Sudek S, et al. Updating algal evolutionary relationships through plastid genome sequencing: did alveolate plastids emerge through endosymbiosis of an ochrophyte? *Sci Rep*. 2015; 5: 10134. doi: [10.1038/srep10134](https://doi.org/10.1038/srep10134) PMID: [26017773](https://pubmed.ncbi.nlm.nih.gov/26017773/)
36. Yamaguchi A, Yubuki N, Leander BS. Morphostasis in a novel eukaryote illuminates the evolutionary transition from phagotrophy to phototrophy: description of *Rapaza viridis* n. gen. et sp. (Euglenozoa, Euglenida). *BMC Evol Biol*. BioMed Central Ltd; 2012; 12: 29. doi: [10.1186/1471-2148-12-29](https://doi.org/10.1186/1471-2148-12-29) PMID: [22401606](https://pubmed.ncbi.nlm.nih.gov/22401606/)
37. Hrdá Š, Fousek J, Szabová J, Hampel V, Vlček Č. The plastid genome of *Eutreptiella* provides a window into the process of secondary endosymbiosis of plastid in euglenids. *PLoS One*. 2012; 7: e33746. doi: [10.1371/journal.pone.0033746](https://doi.org/10.1371/journal.pone.0033746) PMID: [22448269](https://pubmed.ncbi.nlm.nih.gov/22448269/)
38. Deusch O, Landan G, Roettger M, Gruenheit N, Kowallik K V., Allen JF, et al. Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor. *Mol Biol Evol*. 2008; 25: 748–761. doi: [10.1093/molbev/msn022](https://doi.org/10.1093/molbev/msn022) PMID: [18222943](https://pubmed.ncbi.nlm.nih.gov/18222943/)
39. Huang S, Shingaki-Wells RN, Taylor NL, Millar AH. The rice mitochondria proteome and its response during development and to the environment. *Front Plant Sci*. 2013; 4: 16. doi: [10.3389/fpls.2013.00016](https://doi.org/10.3389/fpls.2013.00016) PMID: [23403394](https://pubmed.ncbi.nlm.nih.gov/23403394/)
40. Gross J, Bhattacharya D. Mitochondrial and plastid evolution in eukaryotes: an outsiders' perspective. *Nat Rev Genet*. 2009; 10: 495–505. doi: [10.1038/nrg2649](https://doi.org/10.1038/nrg2649) PMID: [19506574](https://pubmed.ncbi.nlm.nih.gov/19506574/)
41. Basak I, Moeller SG. Emerging facets of plastid division regulation. *Planta*. 2013. pp. 389–398. doi: [10.1007/s00425-012-1743-6](https://doi.org/10.1007/s00425-012-1743-6) PMID: [22965912](https://pubmed.ncbi.nlm.nih.gov/22965912/)
42. Gould SB, Waller RF, McFadden GI. Plastid evolution. *Annu Rev Plant Biol*. 2008; 59: 491–517. doi: [10.1146/annurev.arplant.59.032607.092915](https://doi.org/10.1146/annurev.arplant.59.032607.092915) PMID: [18315522](https://pubmed.ncbi.nlm.nih.gov/18315522/)
43. Barbrook AC, Howe CJ, Purton S. Why are plastid genomes retained in non-photosynthetic organisms? *Trends Plant Sci*. 2006; 11: 101–108. doi: [10.1016/j.tplants.2005.12.004](https://doi.org/10.1016/j.tplants.2005.12.004) PMID: [16406301](https://pubmed.ncbi.nlm.nih.gov/16406301/)
44. Smith DR, Crosby K, Lee RW. Correlation between nuclear plastid DNA abundance and plastid number supports the limited transfer window hypothesis. *Genome Biol Evol*. 2011; 3: 365–371. doi: [10.1093/gbe/evr001](https://doi.org/10.1093/gbe/evr001) PMID: [21292629](https://pubmed.ncbi.nlm.nih.gov/21292629/)

45. Gillott MA, Gibbs SP. Cryptomonad nucleomorph: its ultrastructure and evolutionary significance. *J Phycol.* 1980; 16: 558–568. doi: [10.1111/j.1529-8817.1980.tb03074.x](https://doi.org/10.1111/j.1529-8817.1980.tb03074.x)
46. Gilson PR, McFadden GI. Good things in small packages: the tiny genomes of chlorarachniophyte endosymbionts. *BioEssays.* 1997; 19: 167–173. doi: [10.1002/bies.950190212](https://doi.org/10.1002/bies.950190212) PMID: [9046247](https://pubmed.ncbi.nlm.nih.gov/9046247/)
47. Hehenberger E, Imanian B, Burki F, Keeling PJ. Evidence for the retention of two evolutionary distinct plastids in dinoflagellates with diatom endosymbionts. *Genome Biol Evol.* 2014; 6: 2321–2334. doi: [10.1093/gbe/evu182](https://doi.org/10.1093/gbe/evu182) PMID: [25172904](https://pubmed.ncbi.nlm.nih.gov/25172904/)
48. Imanian B, Keeling PJ. Horizontal gene transfer and redundancy of tryptophan biosynthetic enzymes in dinotoms. *Genome Biol Evol.* 2014; 6: 333–343. doi: [10.1093/gbe/evu014](https://doi.org/10.1093/gbe/evu014) PMID: [24448981](https://pubmed.ncbi.nlm.nih.gov/24448981/)
49. Dodge JD. The functional and phylogenetic significance of dinoflagellate eyespots. *BioSystems.* 1983; 16: 259–267. doi: [10.1016/0303-2647\(83\)90009-6](https://doi.org/10.1016/0303-2647(83)90009-6) PMID: [6687045](https://pubmed.ncbi.nlm.nih.gov/6687045/)
50. Kořený L, Sobotka R, Kovářová J, Gnipová A, Flegontov P, Horváth A, et al. Aerobic kinetoplastid flagellate *Phytomonas* does not require heme for viability. *Proc Natl Acad Sci U S A.* 2012; 109: 3808–3813. doi: [10.1073/pnas.1201089109](https://doi.org/10.1073/pnas.1201089109) PMID: [22355128](https://pubmed.ncbi.nlm.nih.gov/22355128/)
51. Weinstein JD, Beale SI. Separate Physiological roles and subcellular compartments for two tetrapyrrole biosynthesis pathway in *Euglena gracilis*. *J Biol Chem.* 1983; 258: 6799–6807. PMID: [6133868](https://pubmed.ncbi.nlm.nih.gov/6133868/)
52. Mayer SM, Beale SI, Weinstein JD. Enzymatic conversion of glutamate to δ -aminolevulinic acid in soluble extract of *Euglena gracilis*. *J Biol Chem.* 1987; 262: 12541–12549. PMID: [2442164](https://pubmed.ncbi.nlm.nih.gov/2442164/)
53. Mayer SM, Beale S. δ -Aminolevulinic acid biosynthesis from glutamate in *Euglena gracilis*. *Plant Physiol.* 1991; 97: 1094–1102. PMID: [16668494](https://pubmed.ncbi.nlm.nih.gov/16668494/)
54. Iida K, Mimura I, Kajiwara M. Evaluation of two biosynthetic pathways to δ -aminolevulinic acid in *Euglena gracilis*. *Eur J Biochem.* 2002; 269: 291–7. Available: <http://www.ncbi.nlm.nih.gov/pubmed/11784323> PMID: [11784323](https://pubmed.ncbi.nlm.nih.gov/11784323/)
55. Kořený L, Oborník M. Sequence evidence for the presence of two tetrapyrrole pathways in *Euglena gracilis*. *Genome Biol Evol.* 2011; 3: 359–364. doi: [10.1093/gbe/evr029](https://doi.org/10.1093/gbe/evr029) PMID: [21444293](https://pubmed.ncbi.nlm.nih.gov/21444293/)
56. Oborník M, Green BR. Mosaic origin of the heme biosynthesis pathway in photosynthetic eukaryotes. *Mol Biol Evol.* 2005; 22: 2343–2353. doi: [10.1093/molbev/msi230](https://doi.org/10.1093/molbev/msi230) PMID: [16093570](https://pubmed.ncbi.nlm.nih.gov/16093570/)
57. Kořený L, Sobotka R, Janouškovec J, Keeling PJ, Oborník M. Tetrapyrrole synthesis of photosynthetic chromerids is likely homologous to the unusual pathway of apicomplexan parasites. *Plant Cell.* 2011; 23: 3454–62. doi: [10.1105/tpc.111.089102](https://doi.org/10.1105/tpc.111.089102) PMID: [21963666](https://pubmed.ncbi.nlm.nih.gov/21963666/)
58. Janouškovec J, Tikhonenkov D V, Burki F, Howe AT, Kolísko M, Mylnikov AP, et al. Factors mediating plastid dependency and the origins of parasitism in apicomplexans and their close relatives. *Proc Natl Acad Sci U S A.* 2015; 112: 10200–7. doi: [10.1073/pnas.1423790112](https://doi.org/10.1073/pnas.1423790112) PMID: [25717057](https://pubmed.ncbi.nlm.nih.gov/25717057/)
59. Fernández Robledo JA, Caler E, Matsuzaki M, Keeling PJ, Shanmugam D, Roos DS, et al. The search for the missing link: A relic plastid in *Perkinsus*? [Internet]. *International Journal for Parasitology.* 2011. pp. 1217–1229. doi: [10.1016/j.ijpara.2011.07.008](https://doi.org/10.1016/j.ijpara.2011.07.008) PMID: [21889509](https://pubmed.ncbi.nlm.nih.gov/21889509/)
60. Ralph SA, van Dooren GG, Waller RF, Crawford MJ, Fraunholz MJ, Foth BJ, et al. Metabolic maps and functions of the *Plasmodium falciparum* apicoplast. *Nat Rev Microbiol.* 2004; 2: 203–216. doi: [10.1038/nrmicro843](https://doi.org/10.1038/nrmicro843) PMID: [15083156](https://pubmed.ncbi.nlm.nih.gov/15083156/)
61. Kořený L, Lukeš J, Oborník M. Evolution of the haem synthetic pathway in kinetoplastid flagellates: An essential pathway that is not essential after all? *Int J Parasitol.* 2010; 40: 149–156. doi: [10.1016/j.ijpara.2009.11.007](https://doi.org/10.1016/j.ijpara.2009.11.007) PMID: [19968994](https://pubmed.ncbi.nlm.nih.gov/19968994/)
62. Rogers MB, Gilson PR, Su V, McFadden GI, Keeling PJ. The complete chloroplast genome of the chlorarachniophyte *Bigeloviella natans*: Evidence for independent origins of chlorarachniophyte and euglenid secondary endosymbionts. *Mol Biol Evol.* 2007; 24: 54–62. doi: [10.1093/molbev/msl129](https://doi.org/10.1093/molbev/msl129) PMID: [16990439](https://pubmed.ncbi.nlm.nih.gov/16990439/)
63. Bendtsen JD, Nielsen H, Von Heijne G, Brunak S. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol.* 2004; 340: 783–795. doi: [10.1016/j.jmb.2004.05.028](https://doi.org/10.1016/j.jmb.2004.05.028) PMID: [15223320](https://pubmed.ncbi.nlm.nih.gov/15223320/)
64. Emanuelsson O, Brunak S, von Heijne G, Nielsen H. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc.* 2007; 2: 953–971. doi: [10.1038/nprot.2007.131](https://doi.org/10.1038/nprot.2007.131) PMID: [17446895](https://pubmed.ncbi.nlm.nih.gov/17446895/)
65. Masuda T, Suzuki T, Shimada H, Ohta H, Takamiya K. Subcellular localization of two types of ferrochelatase in cucumber. *Planta.* 2003; 217: 602–609. doi: [10.1007/s00425-003-1019-2](https://doi.org/10.1007/s00425-003-1019-2) PMID: [12905021](https://pubmed.ncbi.nlm.nih.gov/12905021/)
66. Maier UG, Rensing SA, Igloi GL, Maerz M. Twintrons are not unique to the *Euglena* chloroplast genome: structure and evolution of a plastome cpn60 gene from a cryptomonad. *Mol Gen Genet.* 1995; 246: 128–131. doi: [10.1007/BF00290141](https://doi.org/10.1007/BF00290141) PMID: [7823908](https://pubmed.ncbi.nlm.nih.gov/7823908/)
67. Douglas SE, Penny SL. The plastid genome of the cryptophyte alga, *Guillardia theta*: Complete sequence and conserved syntenic groups confirm its common ancestry with red algae. *J Mol Evol.* 1999; 48: 236–244. doi: [10.1007/PL00006462](https://doi.org/10.1007/PL00006462) PMID: [9929392](https://pubmed.ncbi.nlm.nih.gov/9929392/)

68. Woo YH, Ansari H, Otto TD, Klinger CM, Kolisko M, Michálek J, et al. Chromerid genomes reveal the evolutionary path from photosynthetic algae to obligate intracellular parasites. *Elife*. 2015; 4: 1–41. doi: [10.7554/eLife.06974](https://doi.org/10.7554/eLife.06974) PMID: [26175406](https://pubmed.ncbi.nlm.nih.gov/26175406/)
69. Burki F, Okamoto N, Pombert J-F, Keeling PJ. The evolutionary history of haptophytes and cryptophytes: phylogenomic evidence for separate origins. *Proc Biol Sci*. 2012; 279: 2246–54. doi: [10.1098/rspb.2011.2301](https://doi.org/10.1098/rspb.2011.2301) PMID: [22298847](https://pubmed.ncbi.nlm.nih.gov/22298847/)
70. Parfrey LW, Lahr DJG, Knoll AH, Katz LA. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proc Natl Acad Sci U S A*. 2011; 108: 13624–9. doi: [10.1073/pnas.1110633108](https://doi.org/10.1073/pnas.1110633108) PMID: [21810989](https://pubmed.ncbi.nlm.nih.gov/21810989/)
71. Zhang H, Hou Y, Miranda L, Campbell DA, Sturm NR, Gaasterland T, et al. Spliced leader RNA trans-splicing in dinoflagellates. *Proc Natl Acad Sci U S A*. 2007; 104: 4618–4623. doi: [10.1073/pnas.0700258104](https://doi.org/10.1073/pnas.0700258104) PMID: [17360573](https://pubmed.ncbi.nlm.nih.gov/17360573/)
72. Gruber A, Vugrinec S, Hempel F, Gould SB, Maier UG, Kroth PG. Protein targeting into complex diatom plastids: Functional characterisation of a specific targeting motif. *Plant Mol Biol*. 2007; 64: 519–530. doi: [10.1007/s11103-007-9171-x](https://doi.org/10.1007/s11103-007-9171-x) PMID: [17484021](https://pubmed.ncbi.nlm.nih.gov/17484021/)
73. Jordan PM. The biosynthesis of uroporphyrinogen III: mechanism of action of porphobilinogen deaminase. *Ciba Found Symp*. NETHERLANDS; 1994; 180: 70–96.
74. Hansen G, Botes L, De Salas M. Ultrastructure and large subunit rDNA sequences of *Lepidodinium viride* reveal a close relationship to *Lepidodinium chlorophorum* comb. nov. (= *Gymnodinium chlorophorum*). *Phycol Res*. 2007; 55: 25–41. doi: [10.1111/j.1440-1835.2006.00442.x](https://doi.org/10.1111/j.1440-1835.2006.00442.x)
75. Larkum AWD, Lockhart PJ, Howe CJ. Shopping for plastids. *Trends Plant Sci*. 2007; 12: 189–195. doi: [10.1016/j.tplants.2007.03.011](https://doi.org/10.1016/j.tplants.2007.03.011) PMID: [17416546](https://pubmed.ncbi.nlm.nih.gov/17416546/)
76. Doolittle WF. You are what you eat: A gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends Genet*. 1998; 14: 307–311. doi: [10.1016/S0168-9525\(98\)01494-2](https://doi.org/10.1016/S0168-9525(98)01494-2) PMID: [9724962](https://pubmed.ncbi.nlm.nih.gov/9724962/)
77. Archibald JM, Rogers MB, Toop M, Ishida K-I, Keeling PJ. Lateral gene transfer and the evolution of plastid-targeted proteins in the secondary plastid-containing alga *Bigeloviella natans*. *Proc Natl Acad Sci U S A*. 2003; 100: 7678–7683. doi: [10.1073/pnas.1230951100](https://doi.org/10.1073/pnas.1230951100) PMID: [12777624](https://pubmed.ncbi.nlm.nih.gov/12777624/)
78. Yang Y, Matsuzaki M, Takahashi F, Qu L, Nozaki H. Phylogenomic analysis of “red” genes from two divergent species of the “green” secondary phototrophs, the chlorarachniophytes, suggests multiple horizontal gene transfers from the red lineage before the divergence of extant chlorarachniophytes. *PLoS One*. 2014; 9. doi: [10.1371/journal.pone.0101158](https://doi.org/10.1371/journal.pone.0101158) PMID: [24972019](https://pubmed.ncbi.nlm.nih.gov/24972019/)
79. Hiraoka Y, Burki F, Keeling PJ. Genome-based reconstruction of the protein import machinery in the secondary plastid of a chlorarachniophyte alga. *Eukaryot Cell*. 2012; 11: 324–333. doi: [10.1128/EC.05264-11](https://doi.org/10.1128/EC.05264-11) PMID: [22267775](https://pubmed.ncbi.nlm.nih.gov/22267775/)
80. Stoecker DK, Johnson MD, De Vargas C, Not F. Acquired phototrophy in aquatic protists. *Aquat Microb Ecol*. 2009; 57: 279–310. doi: [10.3354/ame01340](https://doi.org/10.3354/ame01340)
81. Burki F, Imanian B, Hehenberger E, Hiraoka Y, Maruyama S, Keeling PJ. Endosymbiotic gene transfer in tertiary plastid-containing dinoflagellates. *Eukaryot Cell*. 2014; 13: 246–255. doi: [10.1128/EC.00299-13](https://doi.org/10.1128/EC.00299-13) PMID: [24297445](https://pubmed.ncbi.nlm.nih.gov/24297445/)
82. Moreira D, Deschamps P. What was the real contribution of endosymbionts to the eukaryotic nucleus? insights from photosynthetic eukaryotes. *Cold Spring Harb Perspect Biol*. 2014; 6: 1–9. doi: [10.1101/cshperspect.a016014](https://doi.org/10.1101/cshperspect.a016014) PMID: [24984774](https://pubmed.ncbi.nlm.nih.gov/24984774/)
83. Pittis AA, Gabaldón T. Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. *Nature*. 2016; 531: 101–4. doi: [10.1038/nature16941](https://doi.org/10.1038/nature16941) PMID: [26840490](https://pubmed.ncbi.nlm.nih.gov/26840490/)
84. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990; 215: 403–10. doi: [10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2) PMID: [2231712](https://pubmed.ncbi.nlm.nih.gov/2231712/)
85. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004; 32: 1792–1797. doi: [10.1093/nar/gkh340](https://doi.org/10.1093/nar/gkh340) PMID: [15034147](https://pubmed.ncbi.nlm.nih.gov/15034147/)
86. Gouy M, Guindon S, Gascuel O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol*. 2010; 27: 221–4. doi: [10.1093/molbev/msp259](https://doi.org/10.1093/molbev/msp259) PMID: [19854763](https://pubmed.ncbi.nlm.nih.gov/19854763/)
87. Stamatakis A. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*. 2006; 22: 2688–2690. doi: [10.1093/bioinformatics/btl446](https://doi.org/10.1093/bioinformatics/btl446) PMID: [16928733](https://pubmed.ncbi.nlm.nih.gov/16928733/)
88. Lartillot N, Philippe H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol*. 2004; 21: 1095–1109. doi: [10.1093/molbev/msh112](https://doi.org/10.1093/molbev/msh112) PMID: [15014145](https://pubmed.ncbi.nlm.nih.gov/15014145/)

89. Van de Peer Y, Frickey T, Taylor JS, Meyer A. Dealing with saturation at the amino acid level: A case study based on anciently duplicated zebrafish genes. *Gene*. 2002; 295: 205–211. PMID: [12354655](#)
90. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*. 2012; 28: 1647–1649. doi: [10.1093/bioinformatics/bts199](#) PMID: [22543367](#)
91. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*. 2011; 29: 644–52. doi: [10.1038/nbt.1883](#) PMID: [21572440](#)
92. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*. 2012; 1: 18. doi: [10.1186/2047-217X-1-18](#) PMID: [23587118](#)
93. Huang X, Madan a. CAP 3: A DNA sequence assembly program. *Genome Res*. 1999; 9: 868–877. doi: [10.1101/gr.9.9.868](#) PMID: [10508846](#)
94. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 2015; 31: 3210–3212. doi: [10.1093/bioinformatics/btv351](#) PMID: [26059717](#)

Supplementary Material

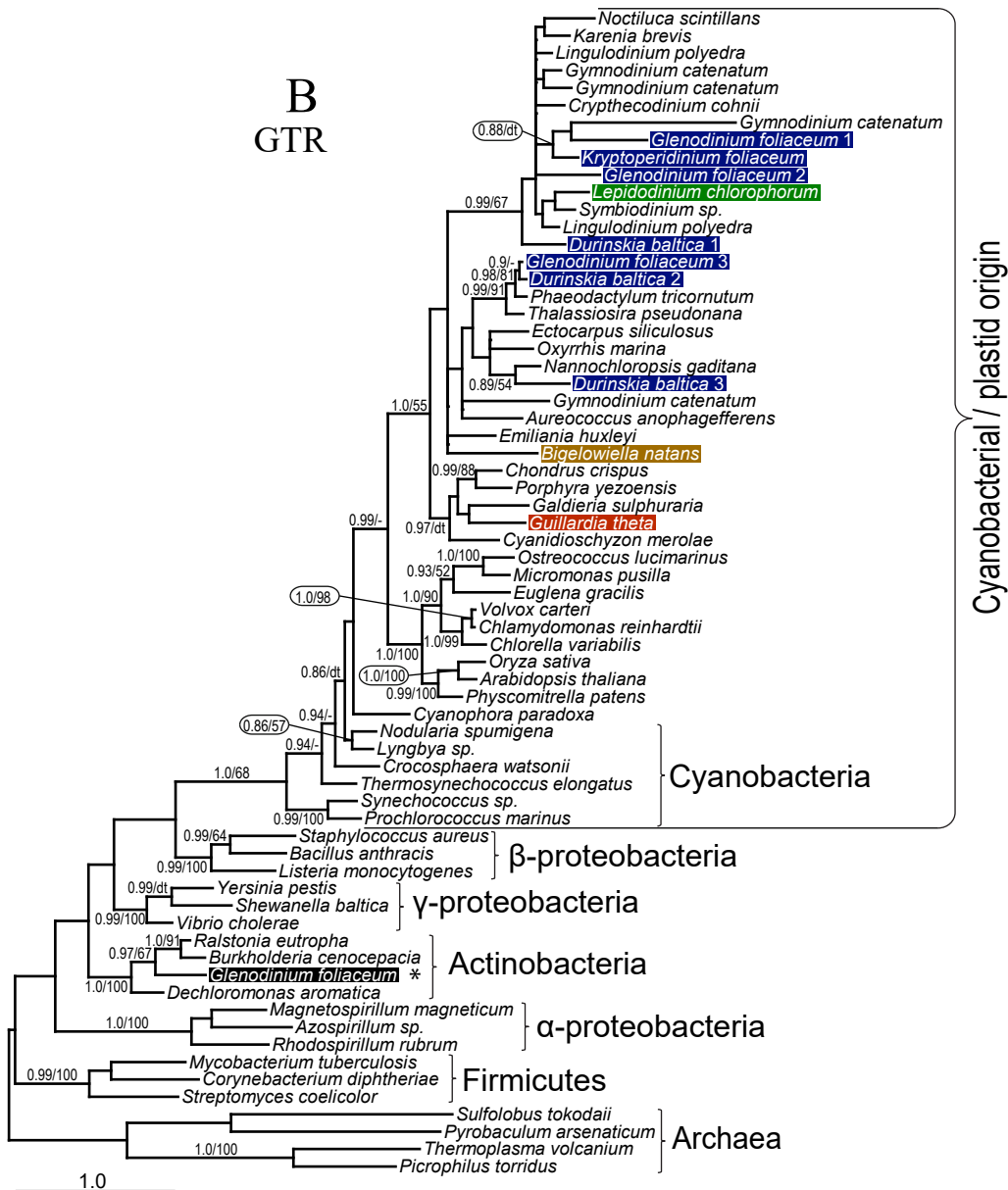


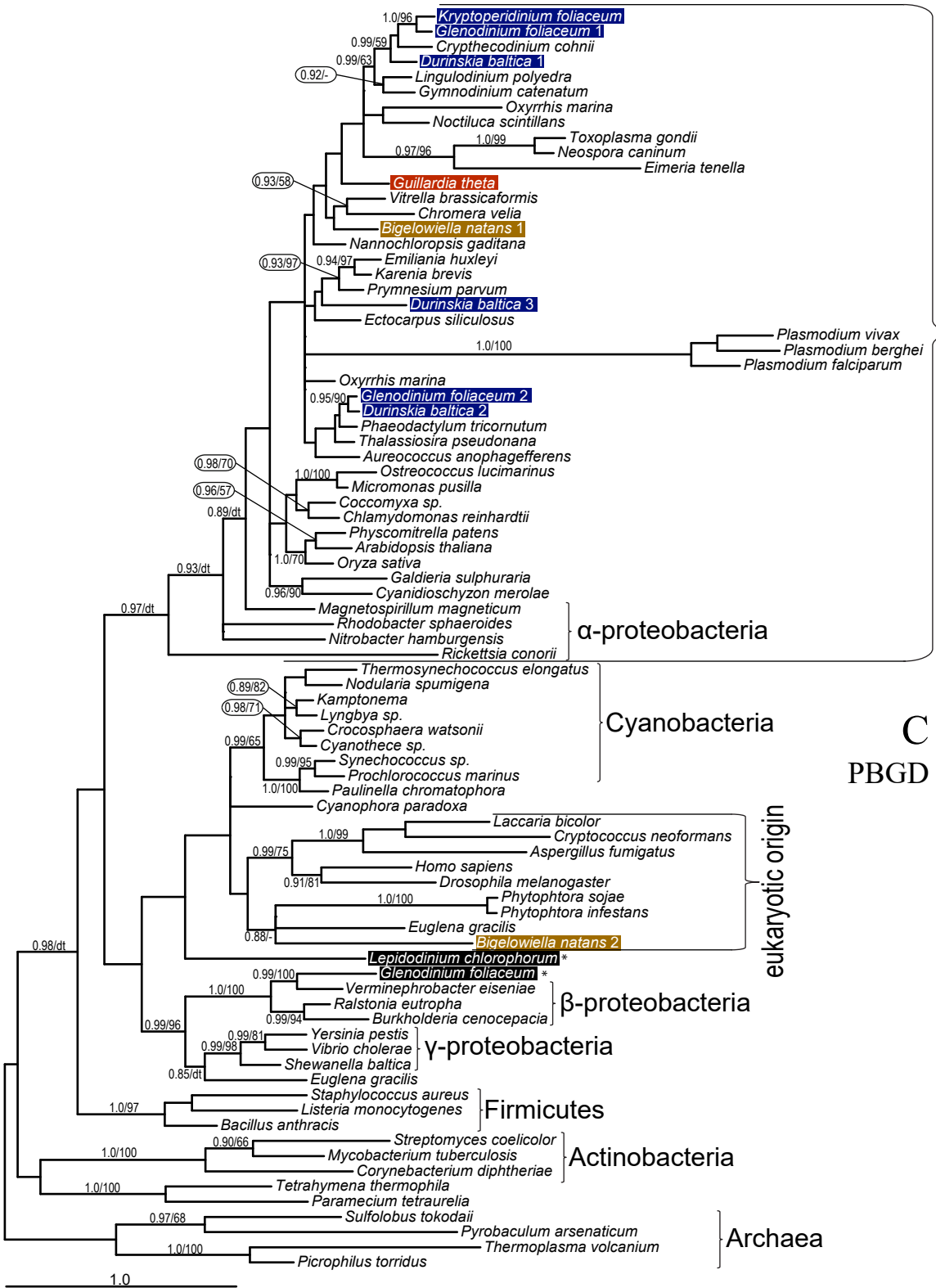
α-proteobacterial/mitochondrial origin

proteobacteria

A
ALAS

B
GTR





α-proteobacterial / mitochondrial origin

C
PBGD

eukaryotic origin

Archaea

α-proteobacteria

Cyanobacteria

β-proteobacteria

γ-proteobacteria

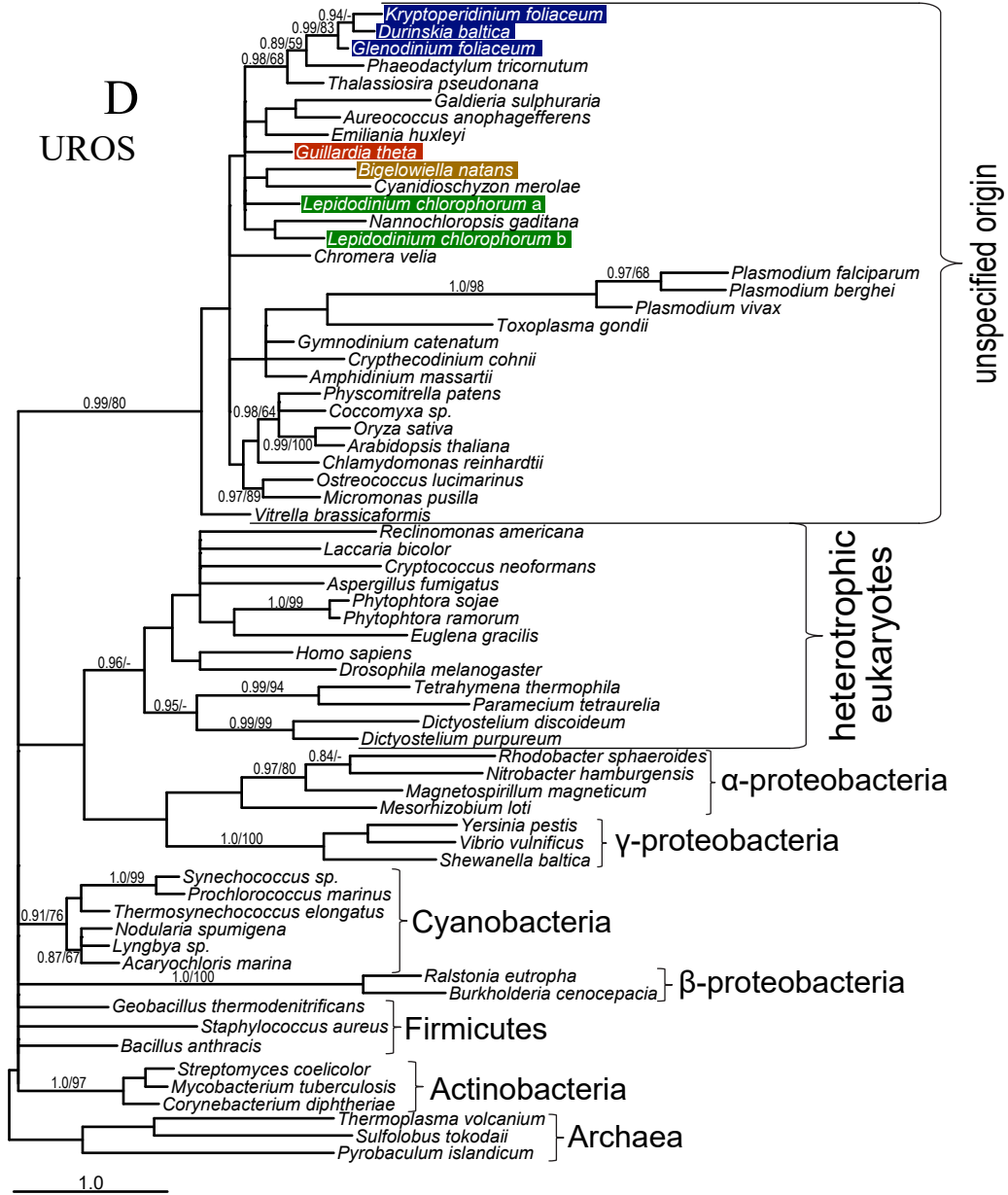
Firmicutes

Actinobacteria

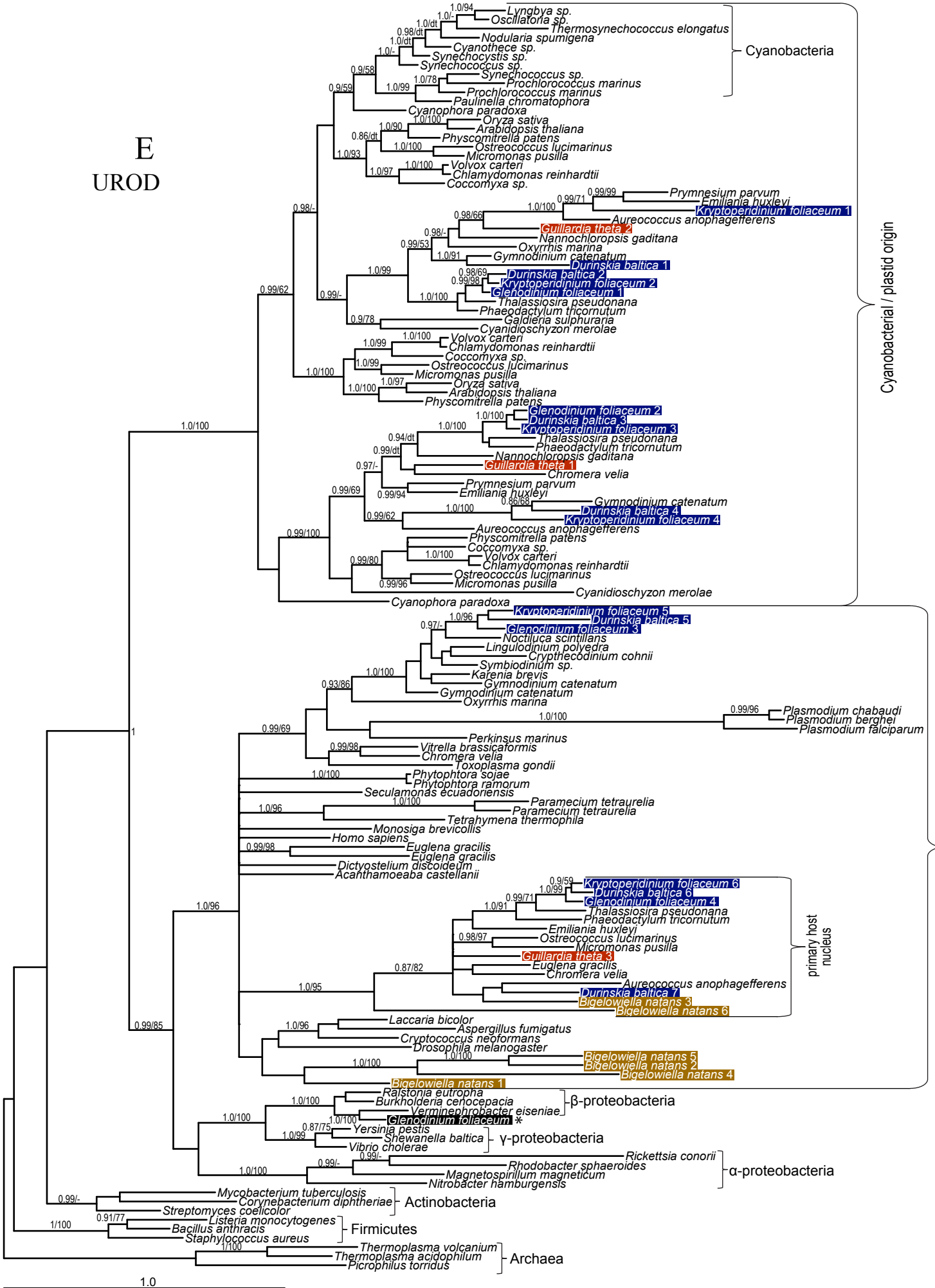
Pyrobaculum arsenaticum
Thermoplasma volcanium

1.0

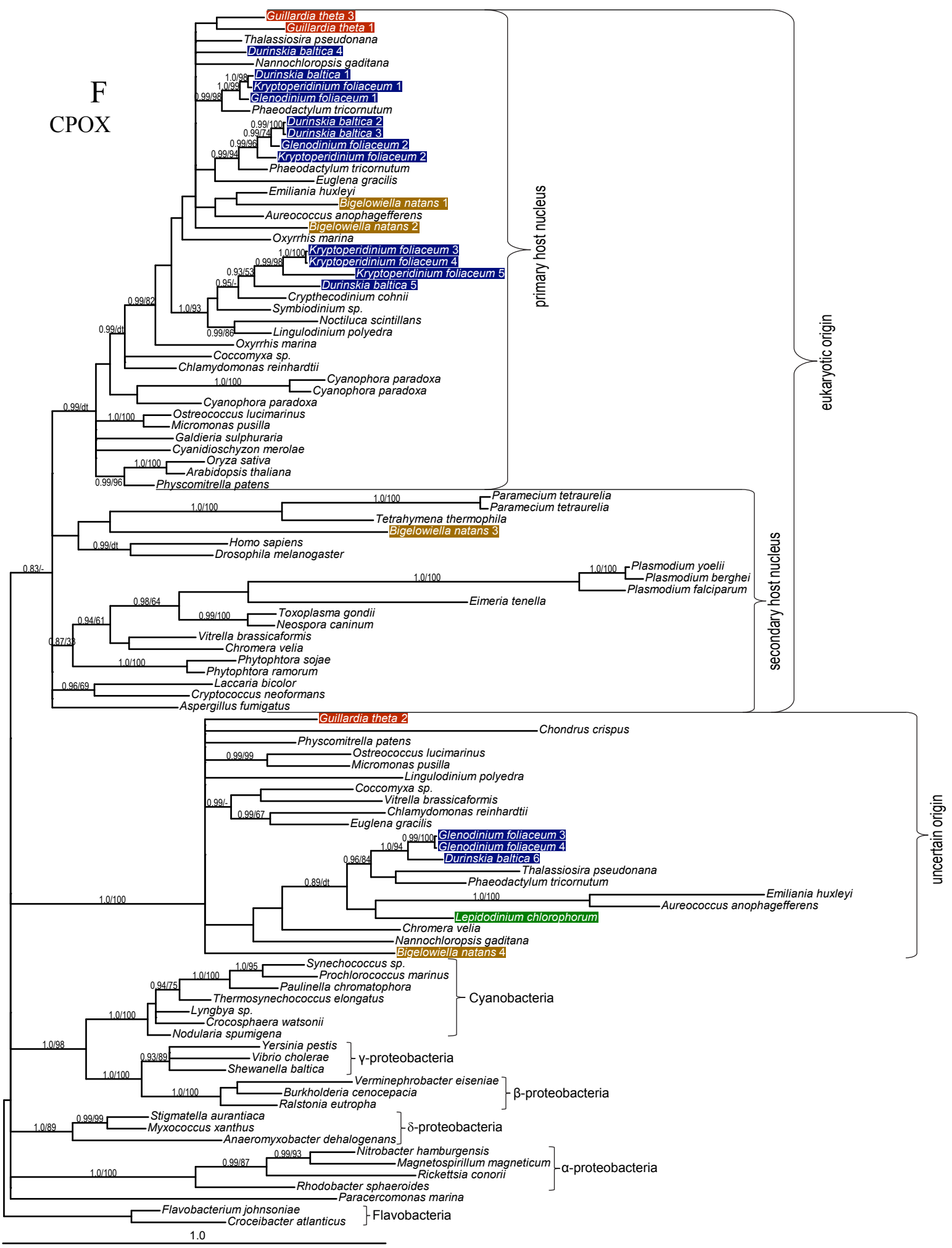
D
UROS



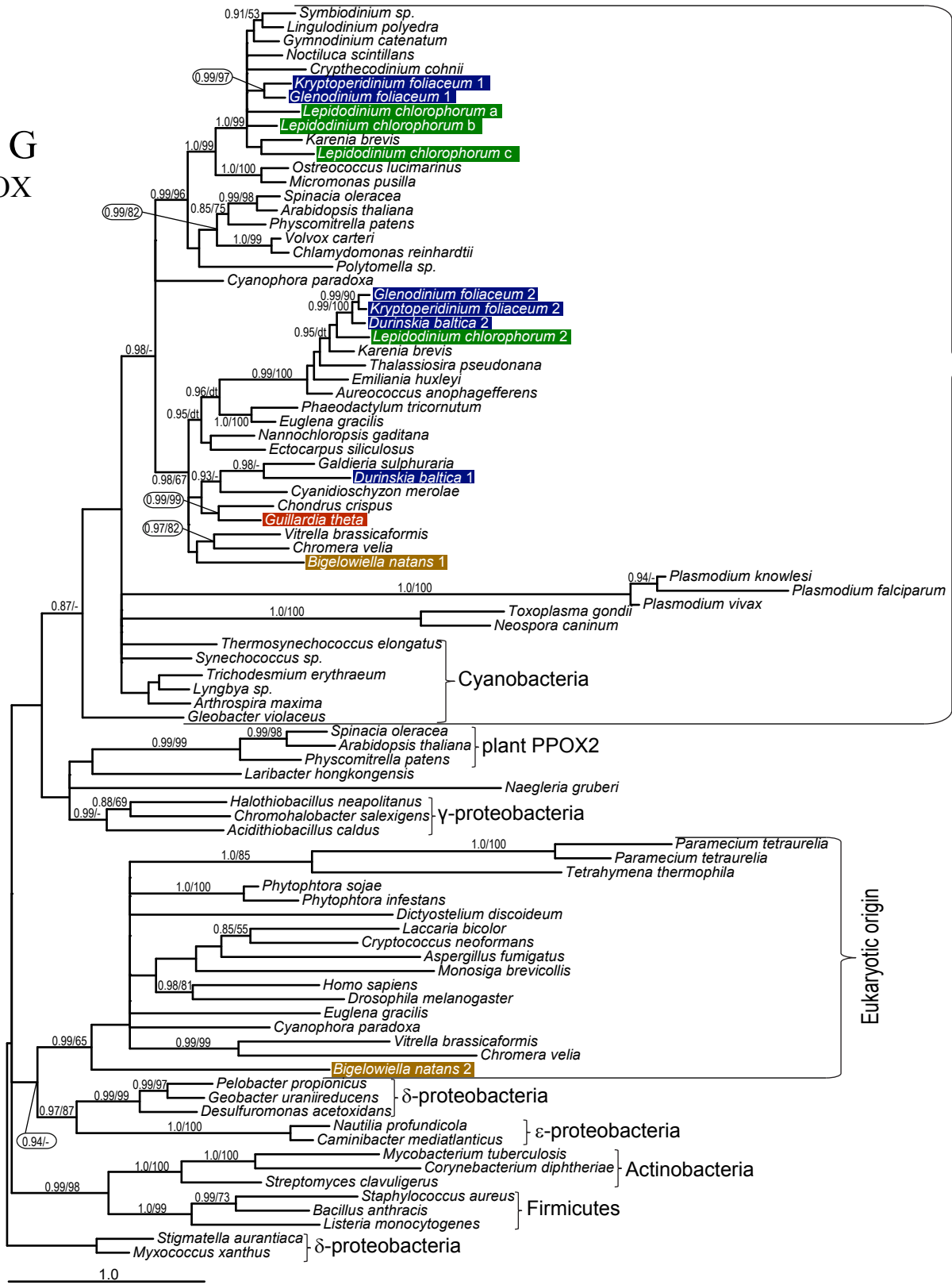
E
UROD



F
CPOX



G
PPOX



Cyanobacterial / plastid origin

Eukaryotic origin

1.0

Fig. S1 Bayesian phylogenetic trees as inferred from the amino acid sequences.

A) ALAS, B) GTR, C) PBGD, D) UROS, E) UROD, F) CPOX and G) PPOX. Taxa of interest of this study are highlighted by colored bars: blue for dinotoms, green for *Lepidodinium chlorophorum*, ochre for *Bigelowiella natans* and red for *Guillardia theta*. The tree demonstrates the mitochondrial origin of ALAS. Numbers near branches indicate Bayesian posterior probabilities followed by the bootstrap of respective clades from the likelihood analysis. Only support values greater than 0.85 (Bayesian) and 50 (likelihood) are shown. dt—different topology in the likelihood tree; a dash indicates unsupported topology. Asterisks mark possible contaminations. LcPPOXa, -b, -c; LcUROSa, -b = non-overlapping protein models, putatively fragments of LcPPOX1 and LcUROS.

Paper II

CHROMERID GENOMES REVEAL THE EVOLUTIONARY PATH FROM PHOTOSYNTHETIC ALGAE TO OBLIGATE INTRACELLULAR PARASITES

Woo YH, Ansari H, Otto TD, Klinger CM, Kolisko M, Michálek J, Saxena A, Shanmugam D, Tayyrov A, Veluchamy A, Ali S, Bernal A, del Campo J, Cihlár J, Flegontov P, Gornik SG, Hajdušková E, Horák A, Janouškovec J, Katris NJ, Mast FD, Miranda-Saavedra D, Mourier T, Naeem R, Nair M, Panigrahi AK, Rawlings ND, Padron-Regalado E, Ramaprasad A, Samad N, Tomčala A, Wilkes J, Neafsey DE, Doerig C, Bowler C, Keeling PJ, Roos DS, Dacks JB, Templeton TJ, Waller RF, Lukeš J, Oborník M, Pain A.

eLife 4:e06974 (2015)

Chromerid genomes reveal the evolutionary path from photosynthetic algae to obligate intracellular parasites

Yong H Woo^{1*}, Hifzur Ansari¹, Thomas D Otto², Christen M Klinger^{3†}, Martin Kolisko^{4†}, Jan Michálek^{5,6†}, Alka Saxena^{1†‡}, Dhanasekaran Shanmugam^{7†}, Annageldi Tayyrov^{1†}, Alaguraj Veluchamy^{8†§}, Shahjahan Ali^{9¶}, Axel Bernal¹⁰, Javier del Campo⁴, Jaromír Cihlář^{5,6}, Pavel Flegontov^{5,11}, Sebastian G Gornik¹², Eva Hajdušková⁵, Aleš Horák^{5,6}, Jan Janoušovec⁴, Nicholas J Katris¹², Fred D Mast¹³, Diego Miranda-Saavedra^{14,15}, Tobias Mourier¹⁶, Raece Naeem¹, Mridul Nair¹, Aswini K Panigrahi⁹, Neil D Rawlings¹⁷, Eriko Padron-Regalado¹, Abhinay Ramaprasad¹, Nadira Samad¹², Aleš Tomčala^{5,6}, Jon Wilkes¹⁸, Daniel E Neafsey¹⁹, Christian Doerig²⁰, Chris Bowler⁸, Patrick J Keeling⁴, David S Roos¹⁰, Joel B Dacks³, Thomas J Templeton^{21,22}, Ross F Waller^{12,23}, Julius Lukeš^{5,6,24}, Miroslav Oborník^{5,6,25}, Arnab Pain^{1*}

*For correspondence: yong.woo@kaust.edu.sa (YHW); arnab.pain@kaust.edu.sa (AP)

†These authors contributed equally to this work

Present address: [†]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Institute, Seattle, United States; [§]Biological and Environmental Sciences and Engineering Division, Center for Desert Agriculture, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia; [¶]The Samuel Roberts Noble Foundation, Ardmore, United States

Competing interests: The authors declare that no competing interests exist.

Funding: See page 17

Received: 16 February 2015

Accepted: 16 June 2015

Published: 15 July 2015

Reviewing editor: Magnus Nordborg, Vienna BioCenter, Austria

© Copyright Woo et al. This article is distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use and redistribution provided that the original author and source are credited.

¹Pathogen Genomics Laboratory, Biological and Environmental Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia; ²Parasite Genomics, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, United Kingdom; ³Department of Cell Biology, University of Alberta, Edmonton, Canada; ⁴Canadian Institute for Advanced Research, Department of Botany, University of British Columbia, Vancouver, Canada; ⁵Institute of Parasitology, Biology Centre, Czech Academy of Sciences, České Budějovice, Czech Republic; ⁶Faculty of Sciences, University of South Bohemia, České Budějovice, Czech Republic; ⁷Biochemical Sciences Division, CSIR National Chemical Laboratory, Pune, India; ⁸Ecology and Evolutionary Biology Section, Institut de Biologie de l'École Normale Supérieure, CNRS UMR8197 INSERM U1024, Paris, France; ⁹Bioscience Core Laboratory, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia; ¹⁰Department of Biology, University of Pennsylvania, Philadelphia, United States; ¹¹Life Science Research Centre, Faculty of Science, University of Ostrava, Ostrava, Czech Republic; ¹²School of Botany, University of Melbourne, Parkville, Australia; ¹³Seattle Biomedical Research Institute, Seattle, United States; ¹⁴Centro de Biología Molecular Severo Ochoa, CSIC/Universidad Autónoma de Madrid, Madrid, Spain; ¹⁵IE Business School, IE University, Madrid, Spain; ¹⁶Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark; ¹⁷European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge, United Kingdom; ¹⁸Wellcome Trust Centre For Molecular Parasitology, Institute of Infection, Immunity and Inflammation, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, United Kingdom; ¹⁹Broad Genome Sequencing and Analysis Program, Broad Institute of MIT and Harvard, Cambridge, United States; ²⁰Department of Microbiology, Monash University, Clayton, Australia; ²¹Department of Microbiology and Immunology, Weill Cornell Medical College, New York, United States; ²²Department of Protozoology, Institute of Tropical Medicine, Nagasaki University, Nagasaki, Japan; ²³Department of Biochemistry, University of Cambridge, Cambridge, United Kingdom; ²⁴Canadian Institute for Advanced Research, Toronto, Canada; ²⁵Institute of Microbiology, Czech Academy of Sciences, České Budějovice, Czech Republic

Abstract The eukaryotic phylum *Apicomplexa* encompasses thousands of obligate intracellular parasites of humans and animals with immense socio-economic and health impacts. We sequenced nuclear genomes of *Chromera velia* and *Vitrella brassicaformis*, free-living non-parasitic photosynthetic algae closely related to apicomplexans. Proteins from key metabolic pathways and from the endomembrane trafficking systems associated with a free-living lifestyle have been progressively and non-randomly lost during adaptation to parasitism. The free-living ancestor contained a broad repertoire of genes many of which were repurposed for parasitic processes, such as extracellular proteins, components of a motility apparatus, and DNA- and RNA-binding protein families. Based on transcriptome analyses across 36 environmental conditions, *Chromera* orthologs of apicomplexan invasion-related motility genes were co-regulated with genes encoding the flagellar apparatus, supporting the functional contribution of flagella to the evolution of invasion machinery. This study provides insights into how obligate parasites with diverse life strategies arose from a once free-living phototrophic marine alga.

DOI: [10.7554/eLife.06974.001](https://doi.org/10.7554/eLife.06974.001)

eLife digest Single-celled parasites cause many severe diseases in humans and animals. The apicomplexans form probably the most successful group of these parasites and include the parasites that cause malaria. Apicomplexans infect a broad range of hosts, including humans, reptiles, birds, and insects, and often have complicated life cycles. For example, the malaria-causing parasites spread by moving from humans to female mosquitoes and then back to humans.

Despite significant differences amongst apicomplexans, these single-celled parasites also share a number of features that are not seen in other living species. How and when these features arose remains unclear. It is known from previous work that apicomplexans are closely related to single-celled algae. But unlike apicomplexans, which depend on a host animal to survive, these algae live freely in their environment, often in close association with corals.

Woo et al. have now sequenced the genomes of two photosynthetic algae that are thought to be close living relatives of the apicomplexans. These genomes were then compared to each other and to the genomes of other algae and apicomplexans. These comparisons reconfirmed that the two algae that were studied were close relatives of the apicomplexans.

Further analyses suggested that thousands of genes were lost as an ancient free-living algae evolved into the apicomplexan ancestor, and further losses occurred as these early parasites evolved into modern species. The lost genes were typically those that are important for free-living organisms, but are either a hindrance to, or not needed in, a parasitic lifestyle. Some of the ancestor's genes, especially those that coded for the building blocks of flagella (structures which free-living algae use to move around), were repurposed in ways that helped the apicomplexans to invade their hosts. Understanding this repurposing process in greater detail will help to identify key molecules in these deadly parasites that could be targeted by drug treatments. It will also offer answers to one of the most fascinating questions in evolutionary biology: how parasites have evolved from free-living organisms.

DOI: [10.7554/eLife.06974.002](https://doi.org/10.7554/eLife.06974.002)

Introduction

The phylum *Apicomplexa* is comprised of eukaryotic, unicellular, obligate intracellular parasites, infecting a diverse range of hosts from marine invertebrates, amphibians, reptiles, birds to mammals including humans. More than 5000 species have been described to date, and over 1 million apicomplexan species are estimated to exist (Adl et al., 2007; Pawlowski et al., 2012). Clinically and economically important apicomplexan pathogens, for example, *Babesia*, *Cryptosporidium*, *Eimeria*, *Neospora*, *Theileria*, *Toxoplasma* (Tenter et al., 2000), and the malaria-causing parasite *Plasmodium* wreak profound negative impacts on animal and human welfare.

Despite their diverse host tropism (Roos, 2005) and life cycle strategies, apicomplexans possess several unifying molecular and cellular features, including the abundance of specific classes of nucleic acid-binding

proteins with regulatory functions in parasitic processes (Campbell et al., 2010; Flueck et al., 2010; Radke et al., 2013; Kafsack et al., 2014; Sinha et al., 2014), extracellular proteins for interactions with the host (Templeton et al., 2004a; Anantharaman et al., 2007), an apical complex comprising a system of cytoskeletal elements and secretory organelles (Hu et al., 2006), an inner membrane complex (IMC) derived from the alveoli (Eisen et al., 2006; Kono et al., 2012; Shoguchi et al., 2013), and a non-photosynthetic secondary plastid, termed the apicoplast (McFadden et al., 1996). How and when these features arose is unclear, owing to the lack of suitable outgroup species for comparative analyses.

Chromerids comprise single-celled photosynthetic colpodellids closely associated (and likely symbiotic) with corals (Cumbo et al., 2013; Janouškovec et al., 2013). Phylogenetic analysis demonstrates that these algae are closely related to Apicomplexa (Janouškovec et al., 2013), confirming the long-standing hypothesis that apicomplexan parasites originated from a free-living, photosynthetic alga (McFadden et al., 1996; Moore et al., 2008). Two known chromerid species, *Chromera velia* and *Vitrella brassicaformis* (Moore et al., 2008; Oborník et al., 2011, 2012), can be cultivated in the laboratory, and their plastid (Janouškovec et al., 2010) and mitochondrial genomes (Flegontov et al., 2015) have been described. We explored whole nuclear genomes of *Chromera* and *Vitrella* to understand how obligate intracellular parasitism has evolved in Apicomplexa.

Results and discussion

Genome assembly and annotation

A shotgun approach was used to sequence and assemble the *Chromera* and *Vitrella* nuclear genome into 5953 and 1064 scaffolds totaling 193.6 and 72.7 million base-pairs (Mb). The disparity in genome size is attributable largely to the presence of transposable elements (TEs) totaling ~30 Mb in *Chromera* vs only 1.5 Mb in *Vitrella*, as the predicted number of protein-coding genes is almost the same at 26,112 and 22,817, respectively. Detailed characterizations of the two genomes and their gene structures are described in Appendix 1 and **Supplementary files 1, 2**.

Ancestral gene content of free-living and parasitic species

We constructed a phylogenetic tree of 26 species, comprising *Chromera*, *Vitrella*, 15 apicomplexans, 2 dinoflagellates, 2 ciliates, 4 stramenopiles, and a green alga. On the phylogenetic tree (**Figure 1A**), *Chromera* and *Vitrella* formed a group closest to the apicomplexan clade, consistent with previous phylogenies (Moore et al., 2008; Janouškovec et al., 2010, 2013, 2015; Oborník et al., 2012). The long branches from their common node are consistent with drastic differences in morphology, life cycle (Oborník et al., 2012), plastid (Janouškovec et al., 2010) and mitochondrial genomes (Flegontov et al., 2015) between the two chromerids (**Figure 1A**). Likewise, despite common origins, apicomplexans show extensively diverse lifestyles, including host tropism and invasion phenotypes (**Figure 1B**).

We reconstructed the parsimonious gene repertoires for the ancestors of the 26 species, at the nodes of the phylogenetic tree (**Figure 2A; Figure 2—figure supplement 1**). We note five key nodes on the evolutionary paths to present-day apicomplexans: the alveolate ancestor; the common ancestor of Apicomplexa and chromerids, termed the proto-apicomplexan ancestor; the apicomplexan ancestor; the ancestor of apicomplexan lineages, for example, coccidia and hematozoa; and extant apicomplexans (**Figure 2A**). Protein-coding genes from the 26 species were clustered by OrthoMCL (Li et al., 2003) into groups of homologous genes, hereafter defined as orthogroups. We note that an orthogroup could have homologous genes from different species (putative orthologs) or from the same species (putative paralogs arising from gene duplications). Gains or losses of orthogroups are displayed as green or red sections of a pie on the phylogenetic tree in **Figure 2A**. Divergence of the proto-apicomplexan ancestor from the alveolate ancestor (Stage I) was accompanied by losses of 1668 and gains of 2197 orthogroups (sum of the two 'pies' in Stage I). Transition of the free-living proto-apicomplexan ancestor to the apicomplexan ancestor (Stage II) is accompanied by many gene losses (3862 orthogroups) but few gains (81 orthogroups) (**Figure 2A**). Divergence of coccidians, for example, *Toxoplasma gondii*, from the apicomplexan ancestor (Stage III) is characterized by modest changes (537 losses; 414 gains), whereas divergence of hematozoans, for example, *Plasmodium* spp., is marked by drastic losses (1384 losses; 77 gains) (**Figure 2A**). Further divergence of apicomplexan taxa beyond Stage III is characterized by modest, lineage-specific gains (**Figure 2A**). Functional composition of gained genes at various stages will be discussed in later sections. Paucity of gained genes (81 orthogroups) during Stage II indicates that the genome of the

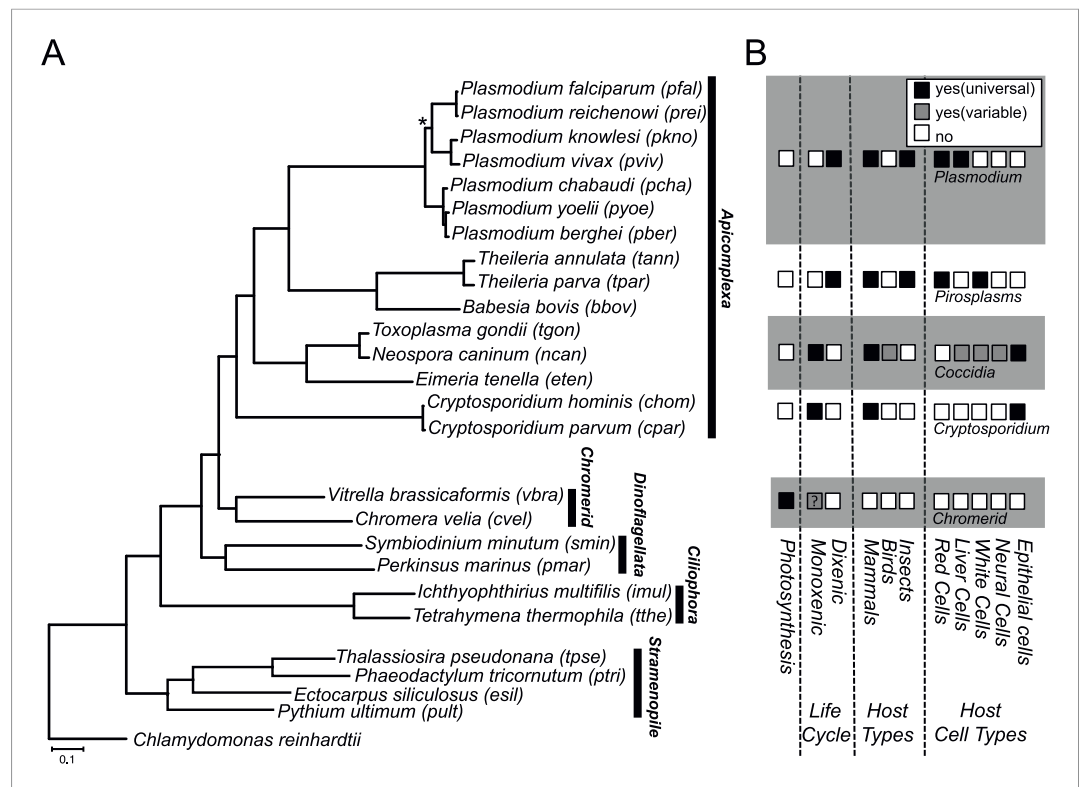


Figure 1. Phylogenetic, parasitological, and genomic context of chromerids. **(A)** Phylogenetic tree of 26 alveolate and outgroup species (see **Figure 1—source data 1** for the list of species). Multiple sequence alignments of 101 genes, which have 1:1 orthologs across all species (**Figure 1—source data 2**) were concatenated to a single matrix of 33,997 aligned amino acids. A maximum likelihood tree was inferred using RAXML with 1000 bootstraps, with *Chlamydomonas reinhardtii* as an outgroup. All clades are supported with bootstrap values of 100% except one node (*) with 99%, and also with 1.00 posterior probability from a bayesian phylogenetic tree based on PhyloBayes (**Lartillot and Philippe, 2004**) (CAT-GTR). **(B)** Lifestyles of the apicomplexan and chromerid species under investigation. '?': uncertainty due to lack of relevant data.

DOI: [10.7554/eLife.06974.003](https://doi.org/10.7554/eLife.06974.003)

The following source data are available for figure 1:

Source data 1. List of 24 species excluding *Chomera* and *Vitrella* used in this study and their data sources.

DOI: [10.7554/eLife.06974.004](https://doi.org/10.7554/eLife.06974.004)

Source data 2. A list of 101 shared orthogroups with a single gene in all of the 26 species, used for the species phylogenetic tree.

DOI: [10.7554/eLife.06974.005](https://doi.org/10.7554/eLife.06974.005)

free-living ancestor possessed most of the genes that were present in the common ancestor of apicomplexans and survived in their present-day descendants.

Progressive, lineage-specific losses during apicomplexan evolution

Parasite evolution has been associated with genome reduction across several branches of the tree of life (**Keeling, 2004; Sakharkar et al., 2004; Morrison et al., 2007**). Examples also exist, however, where parasite genomes are not reduced (**Pombert et al., 2014**) but expanded (**Raffaele and Kamoun, 2012**), underscoring the fact that the genome reduction process during parasite evolution is not completely understood. We sought to characterize in detail the dynamics of gene loss across apicomplexan evolution, particularly for components of molecular processes that are hallmarks of free-living lifestyle. We performed a systematic analysis of the cellular components involved in: (1) cellular metabolic pathways; (2) the endomembrane trafficking systems, regulating the movement of molecules across intracellular compartments in eukaryotes (**Leung et al., 2008**); and (3) the flagellum, a highly conserved apparatus for motility in aqueous environment (**Silflow and Lefebvre, 2001**).

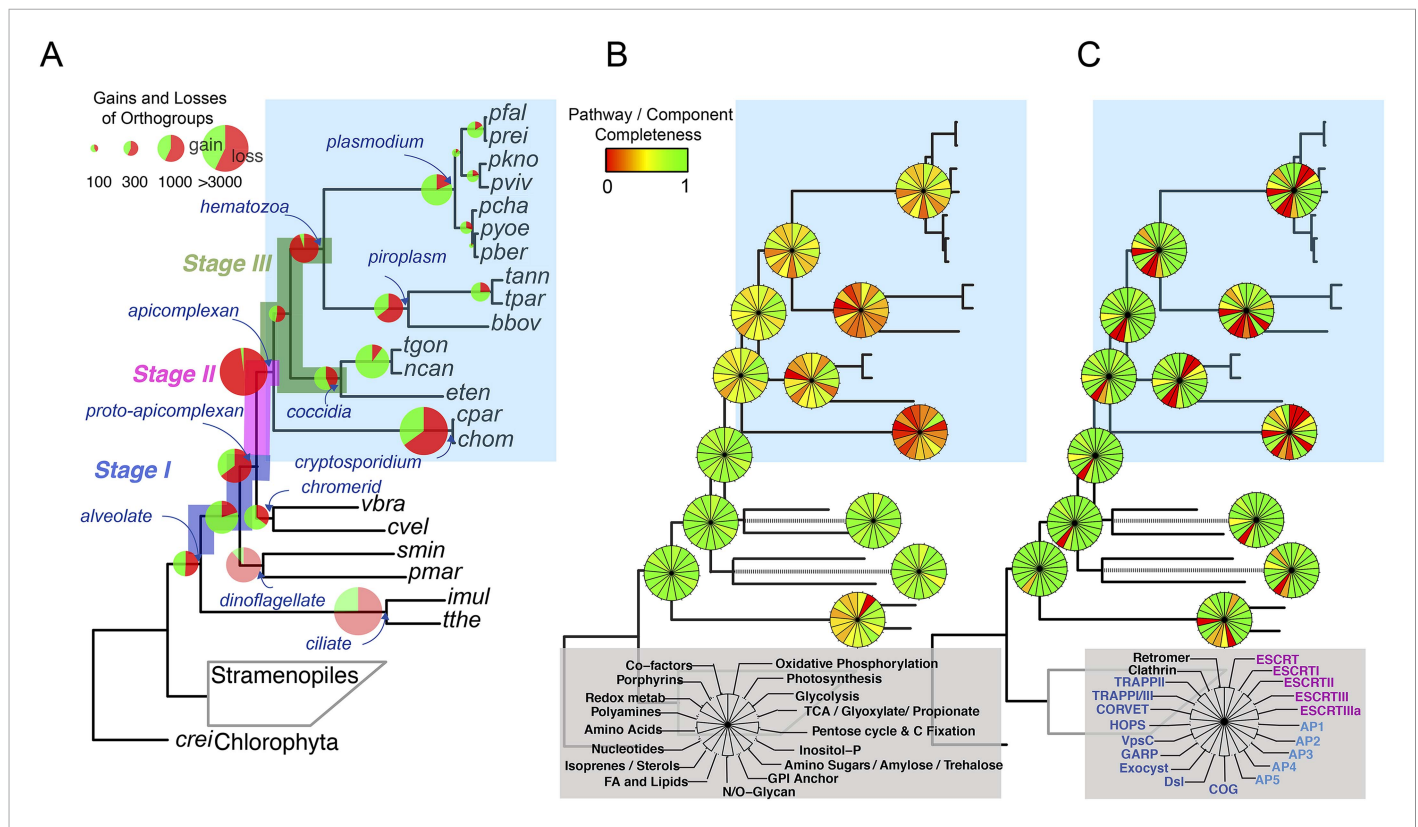


Figure 2. Gene content changes during apicomplexan evolution. **(A)** Gains and losses of orthogroups inferred based on Dollo parsimony (Csuros, 2010). Analysis based on a gene birth-and-death model provided similar results (Figure 2—figure supplement 1A). Stages I, II, and III (shown in blue, pink and green, respectively) represent groups of branches from the alveolate ancestor to apicomplexan lineage ancestors. Stage III could not be determined for *Cryptosporidium* lineage because of sparse taxon sampling. The area of a green or red section in a pie is proportional to the number of gained or lost orthogroups, respectively. **(B, C)** Overview of metabolic capabilities **(B)** and endomembrane components **(C)** in apicomplexan and chromerid ancestors. Gains and losses of enzymes and components were inferred, based on Dollo parsimony (Csuros, 2010). The pie charts are color-coded based on the fraction of enzymes or components present. Additional results from analysis of individual components and enzymes can be found in Figure 2—figure supplements 2,3,4,5, Supplementary file 3. Individual components and enzymes are listed in Figure 2—source data 1, 2. Similar analyses were performed for components encoding flagellar apparatus (Figure 2—figure supplement 5B).

DOI: 10.7554/eLife.06974.006

The following source data and figure supplements are available for figure 2:

Source data 1. Distribution of enzymes based on KEGG.

DOI: 10.7554/eLife.06974.007

Source data 2. Genes encoding subunits of the endomembrane trafficking system.

DOI: 10.7554/eLife.06974.008

Figure supplement 1. Gene gains and losses across the hypothetical ancestors of the 26 species under study.

DOI: 10.7554/eLife.06974.009

Figure supplement 2. Overview of chromerid Carbamoyl Phosphate Synthetase (CPS) and Fatty Acid Synthase I (FAS I).

DOI: 10.7554/eLife.06974.010

Figure supplement 3. Summary of metabolic pathways based on KEGG Assignments.

DOI: 10.7554/eLife.06974.011

Figure supplement 4. An overview of endomembrane trafficking components.

DOI: 10.7554/eLife.06974.012

Figure supplement 5. Evolutionary history of genes encoding cytoskeleton across 26 species.

DOI: 10.7554/eLife.06974.013

The following source data is available for figure 2s5:

Figure supplement 5—source data 1. Genes encoding components of the flagellar apparatus in the 26 species.

DOI: 10.7554/eLife.06974.014

The inferred proto-apicomplexan ancestor, like present-day chromerids, possessed complete metabolic pathways for sugar metabolism, assimilation of nitrate and sulfite, and photosynthesis-related functions (**Figure 2B**, **Figure 2—figure supplement 3**, Appendix 2, and **Supplementary file 3**). Unlike in other photosynthetic algae, both *Chromera* and *Vitrella* initiate heme synthesis in the mitochondrion using aminolevulinic synthase (C4 pathway), which thus far has been found only in a few eukaryotic heterotrophs, such as *Euglena gracilis*, dinoflagellates, and apicomplexans (Kořený et al., 2011; van Dooren et al., 2012; Danne et al., 2013) (Appendix 2 and **Supplementary file 4**). Both chromerids and apicomplexans encode modular multi-domain fatty acid synthase I (FASI)/polyketide synthase enzymes and single-domain FASII components (**Figure 2—figure supplement 2A,B**). Treatment of *Chromera* with a FASII inhibitor triclosan showed decreased production of long chain fatty acids (**Figure 2—figure supplement 2C** and Appendix 2), suggesting that *Chromera* synthesizes short-chain saturated fatty acids using the FASI pathway, which are then elongated using the FASII pathway. This was previously demonstrated in *Toxoplasma*, an apicomplexan that possesses both FASI and FASII (Mazumdar and Striepen, 2007). Likely, the proto-apicomplexan ancestor was a phototrophic alga harboring characteristic metabolic features previously found only in apicomplexan parasites, especially with regard to plastid-associated metabolic functions (see above and other examples in Appendix 2) (Kořený et al., 2011; van Dooren et al., 2012; Danne et al., 2013).

Transition to an apicomplexan ancestor (Stage II) was accompanied by the loss of metabolic processes including photosynthesis and sterol biosynthesis (**Figure 2B** and **Figure 2—figure supplement 3**). The apicomplexan ancestor appeared to possess a significant complement of enzymes in various pathways (**Figure 2B**) (Lim and McFadden, 2010). The differentiation of apicomplexan lineages (Stage III) was accompanied by further lineage-specific losses: for example, loss of FASI in *Plasmodium* spp, loss of FASII in *Cryptosporidium* spp., which has also lost the apicoplast, and loss of enzymes mediating polyamine biosynthesis in all lineages except *Plasmodium* (**Figure 2B** and **Figure 2—figure supplement 3**). These support the notion that enzymes involved in cellular metabolism critical for free-living organisms were not completely lost during the transition to the apicomplexan ancestor, but were further lost during subsequent differentiation and host-adaptation of apicomplexan lineages.

The proto-apicomplexan had a nearly complete repertoire of the endomembrane trafficking complexes, and much of this repertoire persisted through to the apicomplexan ancestor (Stage II) (Hager et al., 1999; Klinger et al., 2013a) (**Figure 2C**, **Figure 2—figure supplement 4** and Appendix 3). Differentiation of apicomplexan lineages (Stage III) was accompanied by lineage-specific losses, for example, loss of the Endosomal Sorting Complex Required for Transport II (ESCRTII) in all lineages except in piroplasms, whereas some components were retained across all lineages, such as the retromer complex components and clathrin, both systems implicated in invasion processes (Pieperhoff et al., 2013; Tomavo et al., 2013) (**Figure 2C**, **Figure 2—figure supplement 4** and Appendix 3). These lineage-specific losses have led to diverse, reduced sets of endomembrane trafficking components in present-day apicomplexans (Hager et al., 1999; Klinger et al., 2013a). Some of these components that were present in chromerids were absent in specific apicomplexan lineages as well as in dinoflagellates and ciliates, further clarifying that these losses are independent, lineage-specific events rather than ancient, shared events.

All known components of flagella were present in the proto-apicomplexan ancestor (**Figure 2—figure supplement 5A,B**). Most of the components were retained in the apicomplexan ancestor (Stage II), but losses occurred as apicomplexan lineages differentiated (Stage III). Components of intraflagellar transport, which are typically essential for assembling flagella, were lost in the other lineages except in coccidians (**Figure 2—figure supplement 5A,B**). The basal body proteins, which support an organizing center for microtubules, were lost from piroplasms. Some striated fiber assembly (SFA) proteins, typically associated with basal body rootlets, were maintained in all apicomplexan lineages including piroplasms (**Figure 2—figure supplement 5A,B,D**); their presence has been hailed as evidence that some flagellar-proteins are repurposed for new functions in apicomplexans (see below) (Francia et al., 2012).

In summary, one of the major events during apicomplexan evolution is progressive, continued loss of components important for free-living organisms. While Stage II was accompanied by a massive loss of such components including those implicated in photosynthesis, the apicomplexan ancestor still possessed many proteins, which were lost later during differentiation of lineages with diverse life strategies.

Emergent features of apicomplexans

Evolution of present-day apicomplexan parasites was accompanied not only by gene losses as noted above (**Figure 2**) but also by gene gains. We sought to determine if genes gained at a particular stage

of apicomplexan evolution, as depicted by the gray violin in **Figure 3**, would be over-represented with those involved in parasitic processes such as intracellular invasion into and egress from host cells. For *Plasmodium falciparum* and *T. gondii*, we compiled three classes of protein-coding genes directly or indirectly involved in parasitic processes of apicomplexans based on in silico prediction or information from previous functional studies ('Materials and methods'). Extracellular proteins are secreted by the apicomplexans for various parasitic processes, for example, some of them are targeted to the host cytoplasm, nucleus, and plasma membrane to modulate parasite–host interactions

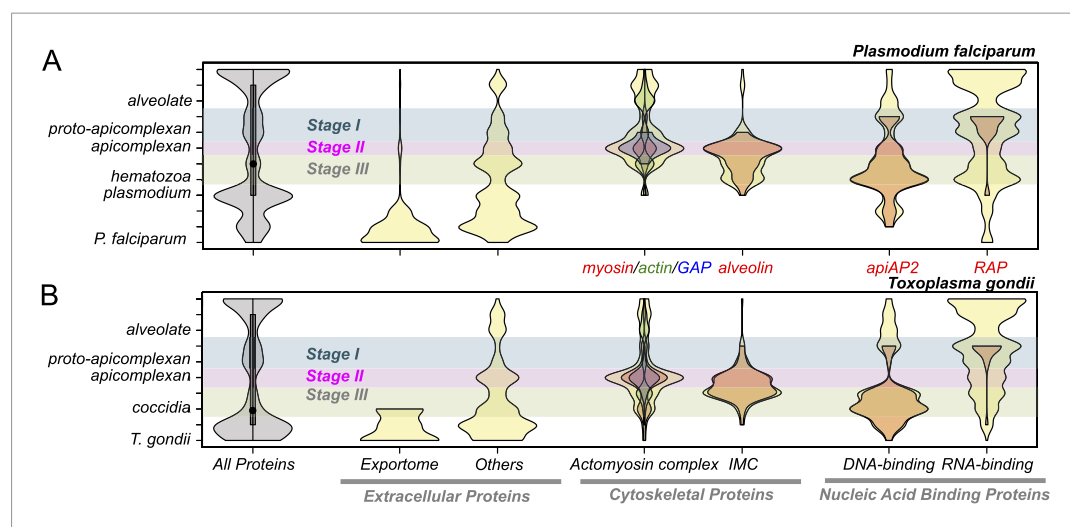


Figure 3. Evolutionary history of *Plasmodium falciparum* and *Toxoplasma gondii* genes. Violin plots showing distribution of evolutionary ages of genes (Y-axis: from species-specific (bottom) to deeply conserved (top)) in *P. falciparum* (A) and *T. gondii* (B). Evolutionary age of a gene is defined as the earliest node on the evolutionary path of the phylogenetic tree where homolog can be detected ('Materials and methods'). The horizontal thickness of a violin is proportional to the number of genes (gray) or the fraction of genes (yellow) in a functional category (X-axis) out of all with the same evolutionary age. Selected functional sub-categories are overlaid with red, green, or blue violin plots. The maximum width of each violin is scaled to be uniform across categories. Inner boxes in the gray violins indicate inter-quartile ranges and circles indicate medians. Colored shades along the X-axis indicate Stages I–III (**Figure 2**). Extracellular proteins include proteins targeted to host cytoplasm, nucleus, and plasma membrane ('exportome') and all other proteins, which are secreted or localized on the parasite surface ('others'). Cytoskeletal proteins include proteins associated with 'actomyosin motor complex' and 'IMC'. All extracellular and cytoskeletal proteins are listed in **Figure 3—source data 1, 2**. Nucleic acid-binding proteins are predicted in silico based on presence of DNA-binding domains (DBDs) and RNA-binding domains (RBDs). See 'Materials and methods' for details on how these genes are defined and compiled. Domain architectures of representative extracellular proteins in apicomplexans and chromerids are displayed as schematics in **Figure 3—figure supplement 4**. Sequence homology networks (**Figure 2—figure supplement 5E** and **Figure 3—figure supplements 1B, 2B, 3B**) and gene gains and losses on the phylogenetic tree (**Figure 3—figure supplements 1A, 2A, 3A**) provide complementary views on the evolutionary history of these genes.

DOI: [10.7554/eLife.06974.015](https://doi.org/10.7554/eLife.06974.015)

The following source data and figure supplements are available for figure 3:

Source data 1. Genes encoding extracellular proteins in *P. falciparum* and *T. gondii*.

DOI: [10.7554/eLife.06974.016](https://doi.org/10.7554/eLife.06974.016)

Source data 2. Genes encoding cytoskeletal components in the 26 species.

DOI: [10.7554/eLife.06974.017](https://doi.org/10.7554/eLife.06974.017)

Figure supplement 1. Evolutionary history of apiAP2 genes.

DOI: [10.7554/eLife.06974.018](https://doi.org/10.7554/eLife.06974.018)

Figure supplement 2. Evolutionary history of alveolins.

DOI: [10.7554/eLife.06974.019](https://doi.org/10.7554/eLife.06974.019)

Figure supplement 3. Evolutionary history of RAP genes.

DOI: [10.7554/eLife.06974.020](https://doi.org/10.7554/eLife.06974.020)

Figure supplement 4. Domain architectures of extracellular proteins in chromerids and apicomplexans.

DOI: [10.7554/eLife.06974.021](https://doi.org/10.7554/eLife.06974.021)

(Mundwiler-Pachlatko and Beck, 2013; Bougdour et al., 2014). Cytoskeletal proteins provide structural support to the cell and also the molecular machinery for motility and intracellular invasion (Baum et al., 2006; Soldati-Favre, 2008). Proteins with DNA-binding domains (DBDs) or RNA-binding domains (RBDs) can regulate various molecular processes of apicomplexan parasites. Indeed, proteins with AP2 (apiAP2) DBD have been shown to act as genetic control switches for diverse apicomplexan processes (Balaji et al., 2005; Campbell et al., 2010; Flueck et al., 2010; Radke et al., 2013; Sinha et al., 2014; Kaneko et al., 2015).

Genes encoding extracellular proteins exported into the host environments were over-represented among those gained after Stage III (Figure 3), suggesting that adaptation to specific hosts was accompanied by expansion of extracellular proteins mediating host–parasite interactions (Templeton et al., 2004a; Anantharaman et al., 2007). Stage III was accompanied by gains of those encoding DBD proteins, mostly apiAP2 proteins (Figure 3 and Figure 3—figure supplement 1A,B), suggesting extensive regulatory changes mediated by apiAP2 proteins during lineage differentiation. We note that losses of other canonical DBD proteins, for example, proteins with HSF_DNA-bind (Pfam: PF00447) domain during transition to apicomplexan ancestor (Stage II) and proteins with Tub (Pfam: PF01167) domain along the piroplasm lineage, contribute to further dominance of apiAP2 among the DBD proteins (Figure 3—figure supplement 1C). Stage II was accompanied by over-represented gains of various cytoskeletal components, including alveolins, those of the actomyosin complex (e.g., myosins) and glideosome-associated proteins with multiple membrane spans 1 and 3 (GAPM1 and GAPM3), suggesting that the molecular machinery powering gliding motility, which is essential for host cell invasion arose during evolution to apicomplexans (Frenal et al., 2010) (Figure 3, Figure 3—figure supplement 2, and Appendix 4). Gene gains during Stage I were over-represented by proteins with ‘RBD abundant in Apicomplexans’ (RAP, Pfam: PF08373) (Lee and Hong, 2004), many of which were conserved as one-to-one orthologs across descending lineages, suggesting development of evolutionarily conserved functions before apicomplexans and chromerids diverged (Figure 3, and Figure 3—figure supplement 3). Chromerid genomes encode many orthologs of apicomplexan cytoskeletal proteins (Appendix 4), including GAPM2, a member of an important protein family for apicomplexan cytoskeletal structure and gliding motility (Bullen et al., 2009), and the IMC sub-compartment protein family (ISP), implicated in establishing apical polarity and coordinating the unique cell cycle of apicomplexans (Poulin et al., 2013) (Figure 2—figure supplement 5E). These data suggest that some components existed in the free-living proto-apicomplexan ancestor and were subsequently repurposed for parasitic processes of apicomplexans.

The *Chromera* and *Vitrella* genomes encode many proteins that are specific to chromerids yet contain functional domains implicated in molecular processes of apicomplexan parasites. For example, there are chromerid-specific proteins with domain architectures similar to those in apicomplexan extracellular proteins, including those previously implicated in host interactions and described in apicomplexans only (Figure 3—figure supplement 4 and Appendix 5, and Supplementary file 5). Presence of such chromerid proteins implies some commonality in extracellular recognition and cross-species interactions and this correlates well with the presumed associations with the coral holobiont (Janoušková et al., 2012, 2013; Cumbo et al., 2013). Importantly, chromerid genomes encode numerous apiAP2 proteins, more abundant than dinoflagellates, suggesting that they have expanded in the proto-apicomplexan ancestor after it split from dinoflagellates (Figure 3—figure supplement 1D). Many of the chromerid apiAP2 proteins belong to putative paralogous clusters, suggesting that their expansion was driven by gene duplication (Figure 3—figure supplement 1D; Appendix 6). Only a small subset of the apiAP2 proteins are shared across apicomplexans, suggesting that the large apiAP2 complement in the proto-apicomplexan ancestor has diversified independently in descending lineages (Figure 3—figure supplement 1A).

In summary, genes encoding critical components of the parasitic lifestyle of apicomplexans were gained at different stages of apicomplexan evolution, some implying subsequent specialization to particular host niches, but others suggesting early adaptations before committing to parasitic lifestyle. This is evident by chromerid orthologs of many such proteins, for example, RAP proteins and specialized cytoskeletal components. Further, chromerid genomes encode chromerid-specific proteins that are not detected as orthologs of apicomplexan proteins but still have functional domains implicated in parasitic processes in apicomplexans. Together, these data imply that a molecular transition had occurred in free-living ancestors of apicomplexans, providing a foundation for host–parasite interactions and further adaptation.

Conserved gene expression programs in the proto-apicomplexan ancestor

Chromera and *Vitrella* genomes allowed us to reconstruct the gene content of the free-living ancestor of apicomplexans. To infer their putative functions using genome-wide gene expression information (Hu et al., 2010), we cultured *Chromera* under 36 different combinations of temperatures, iron and salt concentrations, and generated their gene expression profiles by RNA-seq (Box et al., 2005). In addition, we have obtained a publicly available growth perturbation data set for *P. falciparum* (Hu et al., 2010). There were 1918 orthogroups shared between the two species. We identified pairs of orthogroups that are co-expressed, that is, showing similar expression patterns across the various conditions, in both species ('Materials and methods') (Figure 4—figure supplement 1A). Such an orthogroup pair, that is, those with conserved co-expression between the two species, would include candidate genes that have been co-regulated together during apicomplexan evolution, from the free-living ancestor to present-day parasites due to conserved functions. This approach, successfully utilized by several studies in the past (Stuart et al., 2003; Mutwil et al., 2011; Gerstein et al., 2014), led to the following two observations in this study.

Many RAP genes appeared during Stage I and have been conserved across the descending phyla (Figure 3 and Figure 3—figure supplement 3), but their precise cellular roles are unknown. For 11 out of 12 orthogroups with RAP domains, co-expressed orthogroups overlapped significantly (Fisher's exact test, $p < 0.05$) between *P. falciparum* and *Chromera*, suggesting involvement of RAP proteins in cellular processes evolutionarily conserved across apicomplexans and chromerids (Figure 4A). RAP and their co-expressed orthogroups encode proteins with putative mitochondrial import signals more often than expected by chance in *Chromera* and *P. falciparum* (Fisher's exact test, $p < 0.05$) (Figure 4B), and also in other apicomplexans and chromerids (Figure 4—figure supplement 1B). We have randomly chosen three *Toxoplasma* RAP genes with predicted mitochondrial localization signals (Supplementary file 6) and confirmed experimentally by 3' endogenous gene-tagging with reporter epitopes that all three are localized to the organelle (Figure 4C). Some of the orthogroups co-expressed with orthogroups containing RAP domains encode protein products predicted to be metabolic enzymes, implying possible involvement of RAPs in mitochondrial metabolism (Figure 4—figure supplement 1C). Consistent with this, the *Cryptosporidium* lineage that has a highly reduced mitochondrion lacking both the genome and most canonical metabolic pathways (Abrahamsen et al., 2004; Xu et al., 2004) is the only apicomplexan group to have also lost its RAP repertoire (Figure 4—figure supplement 1D). Loss of RAPs along with a set of mitochondrial functions in this lineage is consistent with a mitochondrial role for RAPs. We speculate that the free-living proto-apicomplexan ancestor possessed within its mitochondrion a regulatory process mediated by RNA-binding activities of the RAP proteins, which has been retained by the extant apicomplexans and chromerids.

As discussed earlier, the proto-apicomplexan ancestor appears to have possessed genes implicated in invasion processes of present-day apicomplexans (Figure 3). Among the 1918 orthogroups, we identified 80 orthogroups comprising genes functionally annotated as implicated in invasion processes. The frequency of co-expression amongst them in the free-living *Chromera* was significantly higher than expected by chance ($p < 0.0005$), suggesting pre-existing functional relationships before transitioning to parasites (Figure 4D). We identified several modules or groups of co-expressed orthogroups (Figure 4E). In one of the co-expression modules (numbered 1 in Figures 4E), 9 out of 10 orthogroups are co-expressed with a gene encoding SFA (Cvel_872), a key protein for organizing the basal bodies of the flagellar apparatus in algae and the apical complexes in apicomplexans (Kawase et al., 2007; Francia et al., 2012) (Figure 4F). We note that SFAs are the only flagellar components found in all apicomplexans tested (Figure 2—figure supplement 5A). Also in this module, for 9 out of 10 orthogroups, their co-expressed orthogroups in *Chromera* overlapped significantly with those in *P. falciparum* (Fisher's exact test, $p < 0.05$), indicating that their regulatory programs have been evolutionarily conserved (Figure 4G). This module include various types of genes implicated in host cell invasion processes of apicomplexans such as genes encoding rhoptyr protein ROP9, apical sushi protein ASP, and gliding motility components GAP40 and GAPM2. The apical complex has been postulated to have emerged from the flagellar apparatus and associated cellular transport systems in free-living algae, based on ultrastructural evidence (Okamoto and Keeling, 2014; Portman et al., 2014). These results suggest that, in the free-living ancestor, some of the genes

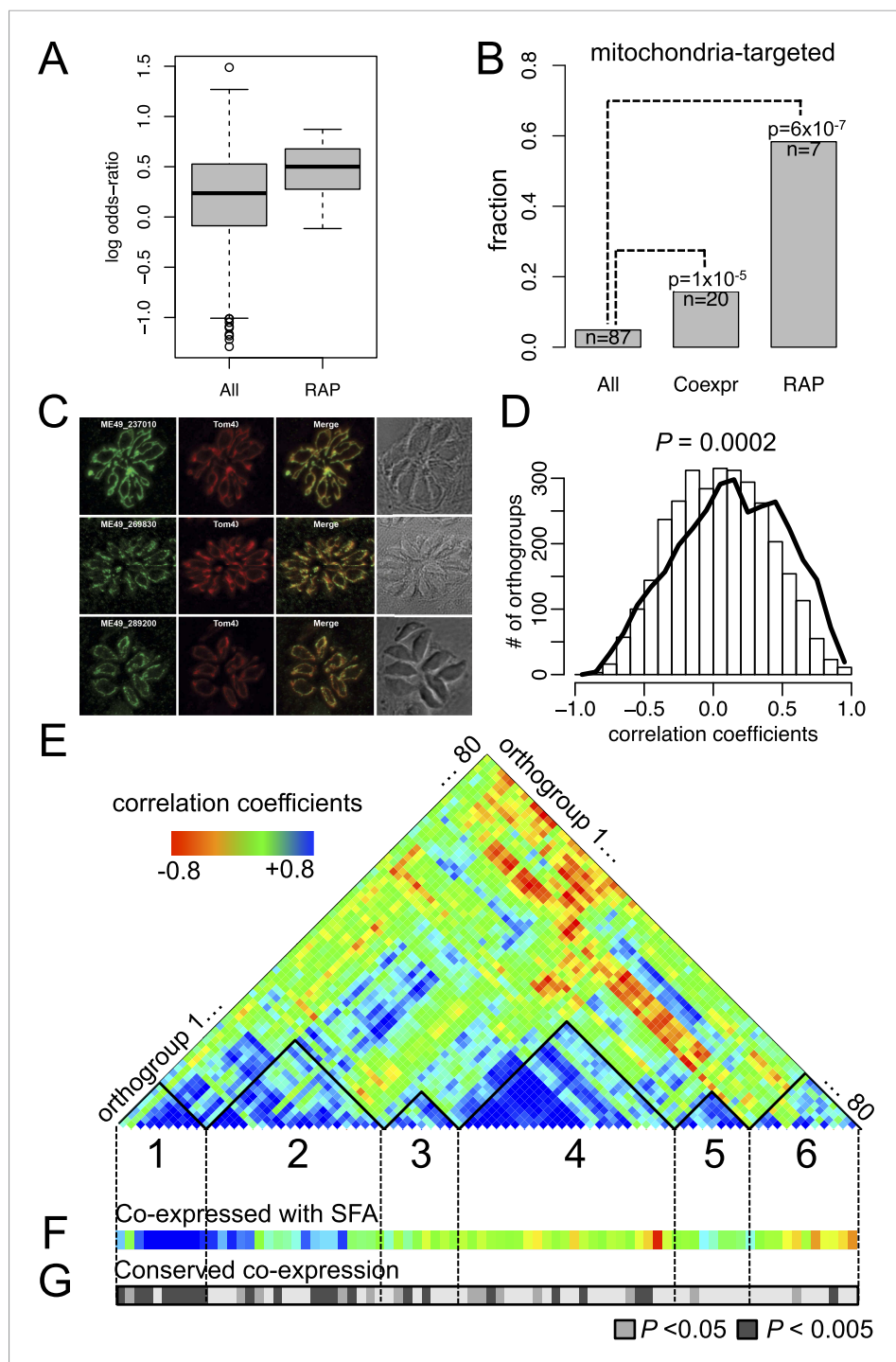


Figure 4. Conserved transcriptional programs in apicomplexans and chromerids. **(A)** Boxplot showing the extent of evolutionary conservation of transcriptional programs for all orthogroups or those with RAP domains. X-axis: 'All' (all orthogroups excluding RAP); 'RAP' (orthogroups with RAP domains). Y-axis: log-transformed odds-ratio, representing, for each orthogroup, the degree of overlap between its co-expressed orthogroups in *Chromera* and those in *P. falciparum*. **(B)** Bar chart showing the fraction of orthogroups (Y-axis) predicted to be targeted to mitochondria in both species ('Materials and methods'). The number of genes are displayed below each bar. X-axis: 'All' (all orthogroups excluding the other two categories); 'Coexpr' (orthogroups co-expressed with RAP in both species); 'RAP' (orthogroups with RAP domains). The fractions in 'Coexpr' and 'RAP' groups were compared against the fraction in 'All', and p-values based Fisher's exact test are displayed above the bar. Files deposited in European Nucleotide Archive are listed in **Figure 4—source data 1** with corresponding conditions. **(C)** Sub-cellular

Figure 4. continued on next page

Figure 4. Continued

localization of RAP proteins encoded by TGME49_237010, TGME49_269830, and TGME49_289200 was tested in *T. gondii* by 3' tagging of the endogenous genes with the coding sequence for the hemagglutinin epitope, together with a mitochondrial marker Tom40. See **Supplementary file 6** for details of the localization predictions.

(D) Distributions of Spearman's rank correlation coefficients of gene expression between all possible pairs from the 80 orthogroups implicated in invasion processes in apicomplexans (black outline) were compared against those from 80 randomly selected ones (histogram). The p value indicates statistical significance of the difference based on 10,000 random samplings. The 80 orthogroups and corresponding genes in *Chromera* and *P. falciparum* are listed in **Figure 4—source data 2**. (E) Heatmap showing a matrix of correlation coefficients amongst the 80 orthogroups. Based on a hierarchical clustering, we classified them into six co-expression modules, labeled as numeral 1–6.

(F) Heatmap showing correlation coefficients with striated fiber assemblin (SFA) (Cvel_872). The color scheme is the same as in (E). (G) Heatmap indicating statistical significance of conserved transcriptional program, that is, the odds-ratio as defined in (A) (Fisher's exact test, $p < 0.05$ (gray); $p < 0.005$ (black)).

DOI: [10.7554/eLife.06974.022](https://doi.org/10.7554/eLife.06974.022)

The following source data and figure supplement are available for figure 4:

Source data 1. RNA-seq libraries of *Chromera velia* under various growth conditions.

DOI: [10.7554/eLife.06974.023](https://doi.org/10.7554/eLife.06974.023)

Source data 2. List of genes implicated in invasion processes in apicomplexans.

DOI: [10.7554/eLife.06974.024](https://doi.org/10.7554/eLife.06974.024)

Source data 3. Evolutionary conservation of 12 orthogroups with RAP domains (for 'RAP' category in **Figure 4A**).

DOI: [10.7554/eLife.06974.034](https://doi.org/10.7554/eLife.06974.034)

Figure supplement 1. Mitochondrial targeting of RAP and its putative role in mitochondrial metabolism.

DOI: [10.7554/eLife.06974.025](https://doi.org/10.7554/eLife.06974.025)

implicated in the invasion process of present-day apicomplexans were functionally associated with those implicated in flagellar motility, providing the much-needed genetic evidence for the postulate. We speculate that a group of functionally related proteins associated with the flagellar apparatus was repurposed as a module of the apical complex and became a foundation for the invasion machinery.

Conclusion

Analysis of *Chromera* and *Vitrella* genomes has enabled insights into how apicomplexan parasites have evolved from free-living ancestors. The transition to parasitism was accompanied by massive genomic loss that continued as its descendants became specialized intracellular parasites infecting diverse hosts. The genome of free-living photosynthetic ancestors encodes many component proteins previously assumed to be restricted to the parasitic apicomplexan lineages. Such pre-existing components, including those of what would later become part of the invasion machinery, were co-opted during evolution to facilitate a successful parasitic lifestyle in multiple hosts. The genome of the proto-apicomplexan ancestor served as a molecular blueprint for evolution of the most successful group of eukaryotic parasites known to date.

Data access

Sequencing data have been deposited in the European Bioinformatics Institute under the European Nucleotide Archive (ENA) sample accession number ERP006228 for *C. velia* and ERP006229 for *V. brassicaformis* for all DNA- and RNA-seq experiments. The assembly and the annotations were submitted under accession numbers CDMZ01000001-CDMZ01005953 for *C. velia* and CDMY01000001-CDMY01001064 for *V. brassicaformis*. Some of the *Vitrella* DNA-seq experiments were done at Broad Institute and are deposited at Short Read Archive under accession numbers SRX152523 and SRX152525. The annotations and assemblies can be viewed and queried in EupathDB (<http://cryptodb.org/cryptodb/>).

Materials and methods

DNA preparation and sequencing

Genomic DNA of *C. velia* CCMP2878 (subsequently referred to as *Chromera*) and *V. brassicaformis* CCMP3155 (subsequently referred to as *Vitrella*) was extracted and then sheared into short fragment

size libraries (300–500 base pair (bp)) and large fragment size libraries (3–8 kbp fragments) by focused-ultrasonication (Covaris Inc., Woburn, USA). The last 3–8 kb libraries were prepared following Nextera mate pair protocol, following manufacturer's instructions. We used three different methods to generate the library: the Illumina (Illumina, San Diego, CA) TruSeq DNA protocol LT Sample Prep Kit (catalog no. #FC-121-2001), an amplification-free method (*Kozarewa et al., 2009*) (TruSeq DNA PCR-Free LT Sample Preparation Kit catalog no. #FC-121-3001) and the Illumina Nextera Mate Pair Sample Preparation Kit (catalog no. #FC-132-1001). The libraries were sequenced on an Illumina HiSeq2000 platform following the manufacturers standard cluster generation and sequencing protocols (*Bentley et al., 2008*; *Quail et al., 2012*). Image analysis, base-calling, and quality filtering were processed by Illumina software.

RNA preparation and sequencing

For isolation of RNAs, *Chromera* and *Vitrella* were grown under standard culture conditions (*Oborník et al., 2012*). Total RNA was extracted from the cells using TRIzol. The polyA⁺ RNA fraction was selected using oligo(dT) beads, and RNA-seq libraries were prepared using TruSeq RNA Sample Prep kit (catalog no. FC-122-1001). Strand-specific RNA-seq libraries were prepared using TruSeq Stranded mRNA LT Sample Prep Kit (catalog no. RS-122-2101) and sequenced as paired-end (2 x 100 bp) reads on a HiSeq2000 platform.

We performed additional RNA sequencing of *Chromera* subject to various environmental perturbations, to construct a global gene expression network based on transcriptomes under various perturbation conditions during in vitro growth. *Chromera* cultures were exposed to a combination of stresses (**Figure 4—figure supplement 1C**). First, six different media were prepared from the combinations of salt concentration (16.7 g/l, 33.3 g/l, 66.6 g/l) and iron deficiency by chelation (*Sutak et al., 2010*). After seeding, the cultures were maintained in the normal temperature and light condition for eleven days (*Oborník et al., 2011*). After randomization, the cultures were incubated at 26°C, 37°C, or 14°C for 0 (control), 0.5, or 2 hr. There were two biological replicates of each, in total 66 flasks of the cultures. Then, the cultures were processed with centrifugation at 3500 RPM for 15 min at 4°C to precipitate the cells. Total RNA was extracted from the 66 cultures after the treatments using Norgen RNA Extraction kit based on manufacturer's protocol (Norgen Biotek Corporation, Canada). RNA quality was assessed using Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA). RNA concentration was determined with a Qubit (Invitrogen, Carlsbad, CA). Strand-specific RNA-seq libraries were prepared from extracted high-quality RNAs (RIN \geq 8.0 as measured on an Agilent Bioanalyzer 2100) using the Illumina TrueSeq LT stranded RNA sample kit according to manufacturer's instructions. Prior to cluster generation, concentration and size of libraries were assayed using the Agilent DNA1000 kit. Libraries from all samples were sequenced as single-end (1 x 50 bp) reads on the Illumina HiSeq 2000. The RNA-seq reads were aligned to the reference genome using tophat (version 2.0.8, default parameters) and cufflinks (version 2-1.0.2, default parameters) (*Trapnell et al., 2012*). The FPKM values were \log_2 normalized with an offset of 1 and were further corrected for different distributions across the samples using the quantile normalization method (*Bolstad et al., 2003*).

Genome assembly

For *Vitrella*, the reads were corrected and assembled followed by several base correction, scaffolding and gap filling steps as briefly described below. As first step, the short insert libraries were corrected with SGA (*Simpson and Durbin, 2012*) (version 0.9.19). The corrected reads were assembled with velvet (*Zerbino and Birney, 2008*) (version 1.2.08). Iterating through different parameter settings, we choose a k-mer of 75 bp as the best parameter set. The resulting scaffolds (larger than 1 kb) were further scaffolded with SSPACE (*Boetzer et al., 2011*) using first the Illumina library (insert = 550 bp) and larger insert (1 kb) Illumina library reads. Sequencing gaps were closed with Gapfiller (*Boetzer and Pirovano, 2012*) (version 1.1.1) with two iterations, using the bowtie mapping option and PCR-Free libraries. Base pair call errors were corrected in three iterations of ICORN (*Otto et al., 2010*), using the amplification-free library. Furthermore, sequencing gaps were closed, using IMAGE (*Tsai et al., 2010*) with the amplification-free library. The assembly was quality-controlled using REAPR (*Hunt et al., 2013*), breaking the contigs at possible miss-assemblies, using the mate pair libraries. This was followed by another scaffolding step. We systematically removed 620 scaffolds containing 25.65 Mb representing the bacterial contamination. The *Vitrella* CCMP3155 assembly contains 72.7

Mb (including 931,689 N's) in 1064 scaffolds (ENA accession numbers CDMY01000001-CDMY01001064). The scaffolds were constructed from 4177 contigs.

For *Chromera*, the assembly pipeline and the algorithms used were the same as *Vitrella*, but due to the larger size, higher amount of low-complexity regions, and difficulties in generating high-quality large insert size libraries, additional steps were included to the assembly process. First, the reads of the PCR-Free library were corrected with SGA ([Simpson and Durbin, 2012](#)) and then assembled with velvet and using a k-mer of 71 (version August 2011). Next, the contigs were scaffolded, gapfilled, and corrected with ICORN, as described earlier. We mapped the reads of all large insert size libraries using SMALT (<ftp://ftp.sanger.ac.uk/pub/resources/software/smalt/>). We excluded scaffolds smaller than 1 kb. Different iterations with SSPACE were undertaken and the assembly was quality-checked with REAPR. After scaffolding, gapfiller and IMAGE were run as above, followed by ICORN. The 1725 scaffolds (spanning 16.02 Mb) representing bacterial contamination were removed. The final assembly of *Chromera* CCMP2878 contains 193.66 Mb (including 582,995 N's) in 5953 scaffolds (ENA accession numbers CDMZ01000001-CDMZ01005953). The scaffolds are constructed from 13,987 contigs.

Gene prediction

We used Augustus ([Stanke et al., 2006](#)) (version 2.5.5) for gene prediction. We manually curated 716 and 245 gene models for *Chromera* and *Vitrella*, respectively, using BLAST similarity-based approaches, and we also generated automated gene models using Cufflinks ([Trapnell et al., 2012](#)) from RNA-seq data sets, in order to use them as a 'training gene model set' for Augustus prediction. The strand-specific RNA-Seq, mapped with TopHat2 ([Kim et al., 2013](#)), was used as evidence in Augustus for intron evidence.

In summary, from the *Chromera* and *Vitrella* genome, we ab initio predicted 30,478 and 23,503 protein-coding genes, respectively, of which 18,829 and 18,240 were detected as being expressed from RNA-seq evidence as poly A+ transcripts ([Supplementary file 1](#)). Excluding putative TEs, 26,112 and 22,817 genes were predicted as protein-coding genes in *Chromera* and *Vitrella*. We annotated partial genes, when a gene probably spans more than one scaffold, located at the borders of a scaffold. We demarcated and annotated as pseudo genes if they contain in frame stop codons. We flagged gene models as transposon elements, if they overlap with the predicted TE regions and had no more than three and two intron for *Chromera* and *Vitrella*, respectively. To annotate untranslated regions (UTRs) of the predicted protein-coding genes, we used CRAIG ([Bernal et al., 2007](#)) with default parameters with mapping of the RNA-Seq data as computed by GSNAP ([Wu and Nacu, 2010](#)) (version 2013-08-19, default parameters). The annotation of both genomes has the ENA accession numbers CDMZ01000001-CDMZ01005953 and CDMY01000001-CDMY01001064 and is also available in EuPathDB ([Aurrecochea et al., 2013](#)).

Functional annotations

The predicted genes were assigned putative functions based on BLASTP (E value $<10^{-6}$) matches against UNIPROT (version March 2012). The predicted protein products were assigned protein domains using *hmmsearch* (HMMER 3.1b1, May 2013) for Pfam A v26.0. Statistical threshold defined by the Pfam ([Finn et al., 2014](#)) database was used. We aligned AP2 sequences in apicomplexan species based on PfamA AP2 (PF00847), and built apicomplexan-specific AP2 (apiAP2) hidden Markov model (HMM), and scanned the predicted protein-coding genes for apiAP2 domains; we annotated api-AP2 DNA-binding transcription factor genes with both domain and sequence E values to be less than 10^{-3} . The following Pfam RBDs were used to define RNA-binding proteins: 'CAT_RBD', 'dsRNA_bind', 'S1', 'DEAD', 'KH_1', 'KH_2', 'KH_3', 'KH_4', 'KH_5', 'RRM_1', 'RRM_2', 'RRM_3', 'RRM_4', 'RRM_5', 'RRM_6', 'SET', 'PUF', and 'RAP'. The list of DBDs was downloaded from a database of DBDs ([Wilson et al., 2008](#)). Transmembrane domains and signal peptides were assigned with the tools TMHMM 2.0 ([Krogh et al., 2001](#)) and signalP 4.0 ([Petersen et al., 2011](#)), respectively, with default parameters.

We collected several categories of genes implicated in parasitic processes in apicomplexans for two archetypal apicomplexan parasites, *Toxoplasma* and *Plasmodium*. We primarily obtained annotations from PlasmoDB ([Bahl et al., 2003](#)) and ToxoDB ([Gajria et al., 2008](#)). Information for sub-cellular localization of genes is obtained from GeneDB ([Logan-Klumpler et al., 2012](#)) and ApiLoc, a database of published protein sub-cellular localization for apicomplexan species

(<http://apiloc.biochem.unimelb.edu.au/apiloc/apiloc>). Some putative parasite genes were inferred based on orthology by OrthoMCL clustering (Li *et al.*, 2003) with closely related species with results from functional studies. We performed exhaustive literature searches to manually curate individual genes, to define rules for in silico searches across the proteomes of this study, and to categorize the identified genes based on their localization and function. The categories of parasite genes are defined as follows.

Cytoskeleton

The cytoskeleton of an organism provides the necessary structural framework for the maintenance of cell shape and integrity. We compiled two groups of cytoskeletal proteins, IMC associated proteins and actomyosin complex. First, IMC associated proteins, comprises alveolin proteins, a membrane occupation and recognition nexus protein (MORN), which associate with IMC and spindle poles and are indispensable for asexual and sexual development (Ferguson *et al.*, 2008). IMC sub-compartment proteins (ISPs) are critical for establishing apical polarity in the parasite (Poulin *et al.*, 2013). Second, components of actomyosin motor complex, which powers the characteristic gliding motility (Soldati-Favre, 2008), comprises actin, myosin, tubulin, gliding associated proteins (GAPs), aldolase, and various actin-regulatory proteins, which will assist actin in the process of quick polymerization–depolymerization cycles between F-actin and G-actin during this process. Examples of actin-regulatory proteins are Arp2/3 complex and formins (FH2) for nucleation; F-actin capping for filament regulation; coronin for cross-linking/bundling and profilin, CAP, cofilin/ADF and gelsolin for monomer treadmill (Baum *et al.*, 2006).

Extracellular proteins

Extracellular proteins are defined as parasite proteins, which are localized either on the surface or secreted off the parasite. They are released in a concerted manner to ensure successful adhesion to the surface, entry into the host cell, multiplication, and escape. Extracellular proteins can be categorized as (1) 'exportome' are proteins translocated to the host cytoplasm, membranes, and nucleus crossing the boundary membrane parasitophorous vacuole (PV); and (2) 'others', which stay on the parasite surface or released from the parasite, but not into the host intracellular space. The exportome genes are released mostly from the parasite's secretory organelles such as rhoptries and dense granules (Ravindran and Boothroyd, 2008; Trecek *et al.*, 2011; Mundwiler-Pachlatko and Beck, 2013; Bougdour *et al.*, 2014). Some of these genes possess host targeting or also known as the Plasmodium export element (PEXEL). Many PEXEL-negative proteins have been identified too (Hsiao *et al.*, 2013; Mundwiler-Pachlatko and Beck, 2013). These genes are sorted and targeted through a specialized structure known as Maurer's cleft formed in the host cytoplasm (Mundwiler-Pachlatko and Beck, 2013). These genes are mostly kinases, proteases, and surface molecules, which modulate the host and hijack the host machinery in favor of parasitic growth and host immune evasion (Trecek *et al.*, 2011; Li *et al.*, 2012; Bougdour *et al.*, 2014). The 'other' extracellular proteins consist of surface antigens (e.g., MSPs), SERAs, TRAPs, AMA-1, microneme proteins, ROPs and RONs etc.

TEs

Repeat annotation was done by using the REPET pipeline (Flutre *et al.*, 2011) and LTR finder (Xu and Wang, 2007). The overall pipeline comprises of two steps: de novo detection and classification. In the first step, the scaffolds are split into smaller batches (~1000 batches of 200 kb each). These genomic fragments were aligned against each other to detect the HSPs (High-scoring pairs) using BLASTER (Quesneville *et al.*, 2003). HSPs are then clustered using a combination of three methods such as GROUPE (Quesneville *et al.*, 2003), RECON (Bao and Eddy, 2002), and PILER (Edgar and Myers, 2005). Structure-based LTR retrotransposons (RTs) detection tools such as LTRharvest (Ellinghaus *et al.*, 2008) and LTR finder, which are based on 100–1000 bp long terminal repeats with a 1 kb–15 kb separation and target site duplication site at vicinity of 60 bp to the two terminal repeats. These LTRs detected are clustered using BlastClust. Multiple sequence alignment of each cluster was performed using MUSCLE (Edgar, 2004). Each cluster aligned was searched against Repbase (Jurka *et al.*, 2005) using BLASTER (Quesneville *et al.*, 2003) and HMMER (Johnson *et al.*, 2010). A consensus feature was detected for each aligned cluster. Further PASTEC (Flutre *et al.*, 2011), which is based on the Wicker classification, was used for consensus classification.

The repeats were annotated as follows. The genomic chunks were randomized and HSPs were detected using BLASTER (Quesneville et al., 2003), CENSOR (Jurka et al., 1996), and RepeatMasker (Tempel, 2012). These HSPs were filtered and combined. Again, full-length genomic scaffolds were compared to Repbase using MATCHER. Satellite and simple repeats were detected using the mreps (Kolpakov et al., 2003), TRF (Benson, 1999), RMSR (RepeatMasker). Finally, a long-join procedure was followed to combine the nested repeats. The whole annotation was exported to a genome-browser readable GFF3 file.

Clustering homologous genes

OrthoMCL 2.0 (Li et al., 2003) was used with a default inflation parameter ($I = 1.5$) (Chen et al., 2006) to generate groups of homologous genes (defined as orthogroups), which could have homologs from different species (putative orthologs) or from the same species (putative paralogs from gene duplications). For some genes of high interest, we manually inspected the alignments of the protein sequences within the orthogroup, which were done with MAFFT (Kato and Standley, 2013). We assigned Pfam domains to an orthogroup if more than half of the genes in an orthogroup were assigned the Pfam domains.

Sub-cellular localization prediction

There are several tools available for a general eukaryotic sub-cellular localization prediction (Du et al., 2011), but they are not applicable to alveolates due to its unique chloroplast membrane arising from secondary endosymbiosis. Therefore, HECTAR (Gschloessl et al., 2008), which was developed for the bipartite sub-cellular prediction, was used. There is no stand-alone version of HECTAR, and the online version allows only one sequence at a time. We implemented a modified HECTAR algorithm as a PERL script for batch prediction of the whole proteomes. Each protein sequence was predicted for signal sequence using SignalP 3.0 (Bendtsen et al., 2004), the signal sequence is cleaved, and the remaining amino acid sequence was used as input for the transit peptide prediction by TargetP (Emanuelsson et al., 2000). Sequences with both signal peptide and the transit peptide (either chloroplast or mitochondria) are predicted to be in the chloroplast. Sequences without the signal peptide but with the transit peptide (either chloroplast or mitochondria) are predicted to be in mitochondria. Sequences with signal peptide, without transit peptide, and predicted by TargetP to be secretory are classified as secretory proteins.

For the RAP proteins, we tested the validity of our sub-cellular localization prediction in two ways. First, we compared our in-house algorithm with other published tools: TargetP (Emanuelsson et al., 2000), MitoProt2 (Claros and Vincens, 1996), iPSORT (Bannai et al., 2002), and PredSL (Petsalaki et al., 2006) (Supplementary file 6, only mitochondrial prediction is shown). We found that our mitochondrial prediction for RAP genes is in concordance with other methods. Second, we experimentally verified mitochondrial localization in *T. gondii* by 3' tagging of the endogenous genes with the coding sequence for the hemagglutinin epitope for three RAP proteins that were predicted to target to mitochondria with high probability.

Statistical analysis

A statistical environment software R was used for most of the analyses and generating parts of figures. An R package *vioplot* was used to generate the violin plot (Hintze and Nelson, 1998). A ward algorithm on the distance matrix based on (1- correlation coefficients) in an R function *hclust* was used for all hierarchical clustering of gene expression patterns unless noted otherwise.

Evolutionary analysis

We compiled the reference proteomes of 26 alveolate and stramenopile species (Figure 1—source data 2) from public databases such as EupathDB (Aurrecochea et al., 2013) and NCBI Genome database (<http://www.ncbi.nlm.nih.gov/genome/>).

We generated a phylogenetic species tree using a data set composed of 101 one-to-one orthologs across the 26 species (see Figure 1—source data 1 for gene IDs). Amino acid sequences were aligned using MAFFT (Kato and Standley, 2013), highly variable sites were edited by trimAL (Capella-Gutierrez et al., 2009) and after manual inspection. The resulting alignment of 33,997 amino acid positions was used to construct trees by a maximum likelihood

(ML) method and Bayesian inference. The ML tree was computed using RAXML 8.1.16 by gamma corrected LG4X model (Stamatakis, 2014; Le et al., 2012). Robustness of the tree was estimated by bootstrap analysis in 1000 replicates. Bayesian tree was constructed by PhyloBayes (Lartillot and Philippe, 2004) using two-infinite mixture model CAT-GTR as implemented in PhyloBayes 3.3f. Two independent chains were run until they converged (i.e., maximum observed discrepancy was lower than 0.2), and the effective number of model parameters was at least 100 after the first 1/5 generation was omitted from topology and posterior probability inference. All clades in the tree were supported with posterior probability 1.00 and 100% bootstraps, except for one node, which representing the common ancestor of human *Plasmodium* spp. was supported by 99% bootstrap.

We performed the gene gain and loss analysis based on Dollo parsimony using Count software (Csuros, 2010). This approach allows reconstructing gene contents at observed species and at hypothetical ancestors, and gene gains and losses at branching points. The Dollo parsimony strictly prohibits multiple gains of genes. To test for validity of this assumption, we repeated analyses based on parsimony settings allowing multiple gene gains or on a phylogenetic birth-and-death model (Csuros, 2010) and reached the same conclusion (Figure 2—figure supplement 1). We have also repeated the analysis using Wagner's parsimony, allowing multiple gains per tree with gain penalty of 2 or greater, and obtained similar results (data not shown). For the analysis of metabolic enzymes, endomembrane trafficking system components, and flagellar apparatus components, the ancestral presence was inferred based on Dollo parsimony from the presence of components in the observed species. For the endomembrane trafficking component analysis, we assumed that the last common ancestor had a complete repertoire of the components.

We have inferred the evolutionary age of *P. falciparum* and *T. gondii* genes as the early node on the phylogenetic tree where the most distant species have genes with significant sequence homology (reciprocal BLASTP E value $<10^{-10}$ and clustering with OrthoMCL).

Comparison of gene expression network between *Chromera velia* and *Plasmodium falciparum*

We studied if orthologs of *Chromera* and *P. falciparum* show similar gene expression changes to physiologically equivalent growth conditions. Identifying equivalent conditions is difficult as the two species have completely different lifestyles and live in different environments. Instead, we tested if a given gene and its ortholog would show correlated expression patterns with the same set of genes (and orthologs), allowing a way to compare gene expression behavior measured under different conditions. To uncover gene-to-gene co-expression relationships, the organisms from whom transcriptomes are sampled must be exposed to various growth conditions. This approach has been successfully used in other eukaryotes (Stuart et al., 2003; Hu et al., 2010; Mutwil et al., 2011). For *Chromera*, we generated RNA-seq-based transcriptome under combinations of varying salt concentrations, iron concentrations, and temperature changes, resulting in 36 unique combinations (see 'Materials and methods' and Figure 4—figure supplement 1C). For *P. falciparum*, we obtained previously published microarray-based gene expression data sets of 144 unique conditions from 23 time series, representing stresses from various growth-inhibiting compounds (Hu et al., 2010). It has been shown that gene expression data generated using different molecular platforms are reproducible and accurate enough for cross-platform comparisons (Woo et al., 2004). Based on each data set, we calculated Spearman correlation coefficients ρ between all possible pairs from the 1918 orthogroups shared between *Chromera* and *P. falciparum* (1918×1918 matrix). We also calculated a 1918×1918 weighted adjacency matrix using CLR algorithm (Faith et al., 2007) as implemented in an R package *minet* (with parameters of method = 'clr', estimator = 'mi.shrink', and disc = 'equalfreq') (Meyer et al., 2008). Expression level of multiple genes in a given orthogroup was averaged. To rule out any potential systematic biases associated with averaging expression levels of homologous, yet distinct genes, we repeated some of the analyses with 1560 orthogroups that have one-to-one orthologs between the two species and reached the same conclusions (data not shown). A pair of genes (or orthogroup) were determined as co-expressed if the Spearman's correlation coefficient ρ is greater than 0.3 and if the value from the weighted adjacency matrix of the network is greater than 0.01. We calculated an odds-ratio to measure the

extent of conservation of co-expressed genes: (# of genes co-expressed in both species) × (# of genes co-expressed in none of the species)/[(# of genes co-expressed in *P. falciparum* only) × (# of genes co-expressed in *C. velia* only)], and Fisher's exact test was used to assess the statistical significance. For calculation of the odds-ratios, co-expression was determined based on correlation coefficient to minimize count granularity in the two-by-two table.

Acknowledgements

We thank the KAUST Bioscience Core Laboratory personnel for sequencing specific Illumina libraries used in this project, KAUST Computational Bioscience Research Center for providing computing resources, Gordon Langsley (Institut Cochin, Inserm U1016, Paris) and Anthony Holder (The Francis Crick Institute, London) for comments on the manuscript draft. AV and CB thank Florian Maumus for advice regarding TE annotation. The primary funding for this work was provided by KAUST award FIC/2010/09 to JL, MO, and AP.

Additional information

Funding

Funder	Grant reference	Author
King Abdullah University of Science and Technology (KAUST)	FIC/2010/09	Aswini K Panigrahi, Julius Lukeš, Miroslav Oborník, Arnab Pain
Council of Scientific and Industrial Research	BSC0124	Dhanasekaran Shanmugam
National Institute of Allergy and Infectious Diseases (NIAID)	HHSN272200900018C	Daniel E Neafsey
Australian Research Council (ARC)	DP120100599	Ross F Waller
Monash University		Christian Doerig
National Health and Medical Research Council (NHMRC)		Christian Doerig
Czech Science Foundation (Grantová agentura České republiky)	P506/12/1522, 13-33039S, P501/12/G055	Jan Michálek, Jaromír Cihlár, Aleš Tomčala, Julius Lukeš, Miroslav Oborník


The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.


Author contributions

YHW, Performed gene annotations; environmental perturbation and transcriptome profiling; invasion pathway and apical complex analysis; DNA- and RNA-binding protein analysis; cross-species transcriptome analysis, Coordinated the genome and transcriptome analyses, Wrote the initial manuscript, Wrote the final manuscript; HA, Performed gene annotations; invasion pathway and apical complex analysis; subcellular targeting prediction; curation of extracellular and cytoskeletal genes in apicomplexans; TDO, Performed genome assembly and gene prediction; gene annotations; gene family analysis; CMK, Performed endomembrane trafficking system analysis; MK, JC, Performed genome analysis; JM, Performed fatty acid biosynthesis; AS, Performed generation and maintenance of specimen, DNA and RNA extractions, library preparation and sequencing; validation of predicted genes; manual curation for gene predictions; DS, Performed global metabolic analysis, Commented and edited on versions of the draft manuscript; AT, Performed generation and maintenance of specimen, DNA and RNA extractions, library preparation and sequencing; manual curation for gene predictions; environmental perturbation and transcriptome profiling; AV, Performed transposable element analysis; SA, EH, JJ, MN, Performed generation and maintenance of specimen, DNA and RNA extractions, library preparation and sequencing; AB, Performed gene annotation validations; JC, Performed generation and maintenance

of specimen, DNA and RNA extractions, library preparation and sequencing; heme pathway and phylogenetic analysis; PF, Performed analysis of chromerid metabolism; commented and edited on versions of the draft manuscript; SGG, performed generation and maintenance of specimen, DNA and RNA extractions, library preparation and sequencing; commented and edited on versions of the draft manuscript; AH, Performed urea pathway and phylogenetic analyses; NJK, NS, Performed validation of RAP's localization in mitochondria; FDM, DM-S, NDR, JW, Performed comparative genome analysis; TM, Performed gene structure analysis; RN, Performed genome validation, annotation, and submission; AKP, Conceived the project; validation of predicted genes; EP-R, AR, Performed manual curation for gene predictions; AT, Performed MS and gas chromatography on fatty acid synthesis; DEN, Performed generation and maintenance of specimen, DNA and RNA extractions, library preparation and sequencing; contributed some raw sequencing reads data; CD, Performed comparative genome analysis, Commented and edited on versions of the draft manuscript; CB, Performed transposable element analysis, Commented and edited on versions of the draft manuscript; PJK, Performed genome analysis, Commented and edited on versions of the draft manuscript; DSR, Global metabolic analysis, Commented and edited on versions of the draft manuscript; JBD, Performed endomembrane trafficking system analysis, Wrote the initial manuscript, Commented and edited on versions of the draft manuscript; TJT, Performed extracellular protein analysis, Wrote the initial manuscript, Commented and edited on versions of the draft manuscript; RFW, Performed validation of RAP's localization in mitochondria, Wrote the initial manuscript, Commented and edited on versions of the draft manuscript; JL, Conceived the project, Commented and edited on versions of the draft manuscript; MO, Conceived the project, Analysis of chromerid metabolism, Commented and edited on versions of the draft manuscript; AP, Conceived the project, Wrote the initial manuscript, Commented and edited on versions of the draft manuscript, Co-ordinated the project

Author ORCIDs

Yong H Woo,  <http://orcid.org/0000-0002-0338-6493>

Javier del Campo,  <http://orcid.org/0000-0002-5292-1421>

Arnab Pain,  <http://orcid.org/0000-0002-1755-2819>

Additional files

Supplementary files

- Supplementary file 1. Summary of the genome assembly and the annotated genes of *Chromera velia*, *Vitrella brassicaformis*. Details of transposable elements on the genome are shown in **Supplementary file 2**.
DOI: [10.7554/eLife.06974.026](https://doi.org/10.7554/eLife.06974.026)
- Supplementary file 2. Summary of transposable elements on the *Chromera velia* and *Vitrella brassicaformis* genomes.
DOI: [10.7554/eLife.06974.027](https://doi.org/10.7554/eLife.06974.027)
- Supplementary file 3. Genes encoding proteins involved in forming photosystems in *Chromera velia* and *Vitrella brassicaformis*.
DOI: [10.7554/eLife.06974.028](https://doi.org/10.7554/eLife.06974.028)
- Supplementary file 4. Genes encoding enzymes involved in heme biosynthesis in chromerids.
DOI: [10.7554/eLife.06974.029](https://doi.org/10.7554/eLife.06974.029)
- Supplementary file 5. Domains of extracellular proteins and example genes in chromerids. (a) Species abbreviations: *Perkinsus marinus*, *P. mar*; *Chromera velia*, *C. vel*; *Vitrella brassicaformis*, *V. bra*; and *Cryptosporidium parvum*, *C. par*. (b) Domain accession identifiers. Domain information can be retrieved at the NCBI Conserved Domain website: (<http://www.ncbi.nlm.nih.gov/cdd>). (c) At the time of publication this accession identifier was valid but the relevant entry could not be retrieved via the NCBI Conserved Domain website: (<http://www.ncbi.nlm.nih.gov/cdd>). (d) A domain having two cysteines and thus far found only as tandem arrays in proteins of *Chromera velia* (for example, Cvel_967). (e) Cysteine-rich domain found in *Cryptosporidium* oocyst wall proteins (COWP) and in coccidians. (f) Archaeal protease-type repeats first described in the *Cryptosporidium* predicted EC protein, cgd7_4560. The domain was previously described as 'A small domain with characteristically spaced cysteine residues that is fused to a papain-like protease domain in the secreted protein

AF1946 from *Archaeoglobus fulgidus* (Templeton et al., 2004a). (g) The domain was previously described as 'Domain typically with 6 cysteines, seen thus far mainly in animals with a few occurrences in plants. It is found in the sea anemone toxin metridin and fused to animal metal proteases, plant prolyl hydroxylases and is vastly expanded in the genome of *Caenorhabditis elegans* (Templeton et al., 2004a)'. (h) The domain was previously described as 'β-strand rich domain, predicted to form a β-sandwich structure that is found in bacterial secreted levanases and glucosidases (Templeton et al., 2004a)'.

DOI: [10.7554/eLife.06974.030](https://doi.org/10.7554/eLife.06974.030)

• Supplementary file 6. Mitochondrial localization predictions of selected RAP genes. Various algorithmic methods were used to identify candidates for experimental validations in *Toxoplasma*. Classifications are given in the column 'Loc' as M-mitochondria; S- secreted; O-others.

DOI: [10.7554/eLife.06974.031](https://doi.org/10.7554/eLife.06974.031)

Major datasets

The following datasets were generated:

Author(s)	Year	Dataset title	Dataset ID and/or URL	Database, license, and accessibility information
Yong H Woo et al,	2015	Chromera velia DNA and RNA sequencing reads	http://www.ebi.ac.uk/ena/data/view/ERP006228	Publicly available at the EBI European Nucleotide Archive (Accession no: ERP006228).
Yong H Woo et al,	2015	Vitrella brassicaformis DNA and RNA sequencing reads	http://www.ebi.ac.uk/ena/data/view/ERP006229	Publicly available at the EBI European Nucleotide Archive (Accession no: ERP006229).
Yong H Woo et al,	2015	Chromera Velia Genome Assembly	http://www.ebi.ac.uk/ena/data/view/CDMZ00000000	Publicly available at the EBI European Nucleotide Archive (Accession no: CDMZ00000000).
Yong H Woo et al,	2015	Vitrella brassicaformis Genome Assembly	http://www.ebi.ac.uk/ena/data/view/CDMY01000000	Publicly available at the EBI European Nucleotide Archive (Accession no: CDMY01000000).

The following previously published dataset was used:

Author(s)	Year	Dataset title	Dataset ID and/or URL	Database, license, and accessibility information
Hu G, Cabrera A, Kono M, Mok S, Chaal BK, Haase S, Engelberg K, Cheemadan S, Spielmann T, Preiser PR, Gilberger TW, Bozdech Z	2010	Perturbation Transcriptome of Plasmodium falciparum	http://www.nature.com/nbt/journal/v28/n1/extended_data_suppl_1/nbt.1597-S2.xls	Publicly available as a part of published dataset.

References

- Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G, Lancto CA, Deng M, Liu C, Widmer G, Tzipori S, Buck GA, Xu P, Bankier AT, Dear PH, Konfortov BA, Spriggs HF, Iyer L, Anantharaman V, Aravind L, Kapur V. 2004. Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. *Science* **304**:441–445. doi: [10.1126/science.1094786](https://doi.org/10.1126/science.1094786).
- Adl SM, Leander BS, Simpson AG, Archibald JM, Anderson OR, Bass D, Bowser SS, Brugerolle G, Farmer MA, Karpov S, Kolisko M, Lane CE, Lodge DJ, Mann DG, Meisterfeld R, Mendoza L, Moestrup O, Mozley-Standridge SE, Smirnov AV, Spiegel F. 2007. Diversity, nomenclature, and taxonomy of protists. *Systematic Biology* **56**: 684–689. doi: [10.1080/10635150701494127](https://doi.org/10.1080/10635150701494127).
- Allen AE, Dupont CL, Obornik M, Horak A, Nunes-Nesi A, Mccrow JP, Zheng H, Johnson DA, Hu H, Fernie AR, Bowler C. 2011. Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. *Nature* **473**: 203–207. doi: [10.1038/nature10074](https://doi.org/10.1038/nature10074).
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* **215**:403–410. doi: [10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2).
- Anantharaman V, Iyer LM, Balaji S, Aravind L. 2007. Adhesion molecules and other secreted host-interaction determinants in *Apicomplexa*: insights from comparative genomics. *International Review of Cytology* **262**:1–74. doi: [10.1016/S0074-7696\(07\)62001-4](https://doi.org/10.1016/S0074-7696(07)62001-4).

- Aurrecoechea C**, Barreto A, Brestelli J, Brunk BP, Cade S, Doherty R, Fischer S, Gajria B, Gao X, Gingle A, Grant G, Harb OS, Heiges M, Hu S, Iodice J, Kissinger JC, Kraemer ET, Li W, Pinney DF, Pitts B, Roos DS, Srinivasamoorthy G, Stoeckert CJ Jr, Wang H, Warrenfeltz S. 2013. EuPathDB: the eukaryotic pathogen database. *Nucleic Acids Research* **41**:D684–D691. doi: [10.1093/nar/gks1113](https://doi.org/10.1093/nar/gks1113).
- Bahl A**, Brunk B, Crabtree J, Fraunholz MJ, Gajria B, Grant GR, Ginsburg H, Gupta D, Kissinger JC, Labo P, Li L, Mailman MD, Milgram AJ, Pearson DS, Roos DS, Schug J, Stoeckert CJ Jr, Whetzel P. 2003. PlasmoDB: the *Plasmodium* genome resource. A database integrating experimental and computational data. *Nucleic Acids Research* **31**:212–215. doi: [10.1093/nar/gkg081](https://doi.org/10.1093/nar/gkg081).
- Balaji S**, Babu MM, Iyer LM, Aravind L. 2005. Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains. *Nucleic Acids Research* **33**:3994–4006. doi: [10.1093/nar/gki709](https://doi.org/10.1093/nar/gki709).
- Bannai H**, Tamada Y, Maruyama O, Nakai K, Miyano S. 2002. Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics* **18**:298–305. doi: [10.1093/bioinformatics/18.2.298](https://doi.org/10.1093/bioinformatics/18.2.298).
- Bao Z**, Eddy SR. 2002. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Research* **12**:1269–1276. doi: [10.1101/gr.88502](https://doi.org/10.1101/gr.88502).
- Baum J**, Gilberger TW, Frischknecht F, Meissner M. 2008. Host-cell invasion by malaria parasites: insights from *Plasmodium* and *Toxoplasma*. *Trends in Parasitology* **24**:557–563. doi: [10.1016/j.pt.2008.08.006](https://doi.org/10.1016/j.pt.2008.08.006).
- Baum J**, Papenfuss AT, Baum B, Speed TP, Cowman AF. 2006. Regulation of apicomplexan actin-based motility. *Nature Reviews. Microbiology* **4**:621–628. doi: [10.1038/nrmicro1465](https://doi.org/10.1038/nrmicro1465).
- Bendtsen JD**, Nielsen H, von Heijne G, Brunak S. 2004. Improved prediction of signal peptides: SignalP 3.0. *Journal of Molecular Biology* **340**:783–795. doi: [10.1016/j.jmb.2004.05.028](https://doi.org/10.1016/j.jmb.2004.05.028).
- Benson G**. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**: 573–580. doi: [10.1093/nar/27.2.573](https://doi.org/10.1093/nar/27.2.573).
- Bentley DR**, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IM, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DM, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Catenazzi E, Chiara M, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Scott Furey W, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Huw Jones TA, Kang GD, Kerelska TH, Kersey AD, Khrebtkova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ling Ng B, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Pike AC, Chris Pinkard D, Pliskin DP, Podhasky J, Quijano VJ, Raczky C, Rae VH, Rawlings SR, Chiva Rodriguez A, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna Sohna J, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovskiy Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurler ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**:53–59. doi: [10.1038/nature07517](https://doi.org/10.1038/nature07517).
- Bernal A**, Crammer K, Hatzigeorgiou A, Pereira F. 2007. Global discriminative learning for higher-accuracy computational gene prediction. *PLoS Computational Biology* **3**:e54. doi: [10.1371/journal.pcbi.0030054](https://doi.org/10.1371/journal.pcbi.0030054).
- Boetzer M**, Henkel CV, Jansen HJ, Butler D, Pirovano W. 2011. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**:578–579. doi: [10.1093/bioinformatics/btq683](https://doi.org/10.1093/bioinformatics/btq683).
- Boetzer M**, Pirovano W. 2012. Toward almost closed genomes with GapFiller. *Genome Biology* **13**:R56. doi: [10.1186/gb-2012-13-6-r56](https://doi.org/10.1186/gb-2012-13-6-r56).
- Bolstad BM**, Irizarry RA, Astrand M, Speed TP. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**:185–193. doi: [10.1093/bioinformatics/19.2.185](https://doi.org/10.1093/bioinformatics/19.2.185).
- Bougdour A**, Tardieux I, Hakimi MA. 2014. *Toxoplasma* exports dense granule proteins beyond the vacuole to the host cell nucleus and rewires the host genome expression. *Cellular Microbiology* **16**:334–343. doi: [10.1111/cmi.12255](https://doi.org/10.1111/cmi.12255).
- Box GE**, Stuart Hunter J, Hunter WG. 2005. *Statistics for experimenters: design, innovation, and discovery*. 2nd ed., Wiley series in probability and statistics. Hoboken: Wiley-Interscience.
- Bullen HE**, Tonkin CJ, O'Donnell RA, Tham WH, Papenfuss AT, Gould S, Cowman AF, Crabb BS, Gilson PR. 2009. A novel family of Apicomplexan glideosome-associated proteins with an inner membrane-anchoring role. *The Journal of Biological Chemistry* **284**:25353–25363. doi: [10.1074/jbc.M109.036772](https://doi.org/10.1074/jbc.M109.036772).
- Campbell TL**, de Silva EK, Olszewski KL, Elemento O, Llinas M. 2010. Identification and genome-wide prediction of DNA binding specificities for the ApiAP2 family of regulators from the malaria parasite. *PLoS Pathogens* **6**: e1001165. doi: [10.1371/journal.ppat.1001165](https://doi.org/10.1371/journal.ppat.1001165).
- Capella-Gutierrez S**, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**:1972–1973. doi: [10.1093/bioinformatics/btp348](https://doi.org/10.1093/bioinformatics/btp348).
- Caron F**, Meyer E. 1989. Molecular basis of surface antigen variation in paramecia. *Annual Review of Microbiology* **43**:23–42. doi: [10.1146/annurev.mi.43.100189.000323](https://doi.org/10.1146/annurev.mi.43.100189.000323).

- Chaudhry F**, Little K, Talarico L, Quintero-Monzon O, Goode BL. 2010. A central role for the WH2 domain of Srv2/CAP in recharging actin monomers to drive actin turnover in vitro and in vivo. *Cytoskeleton* **67**:120–133. doi: [10.1002/cm.20429](https://doi.org/10.1002/cm.20429).
- Chen F**, Mackey AJ, Stoeckert CJ Jr, Roos DS. 2006. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Research* **34**:D363–D368. doi: [10.1093/nar/gkj123](https://doi.org/10.1093/nar/gkj123).
- Claros MG**, Vincens P. 1996. Computational method to predict mitochondrially imported proteins and their targeting sequences. *European Journal of Biochemistry* **241**:779–786. doi: [10.1111/j.1432-1033.1996.00779.x](https://doi.org/10.1111/j.1432-1033.1996.00779.x).
- Coppens I**, Sinai AP, Joiner KA. 2000. *Toxoplasma gondii* exploits host low-density lipoprotein receptor-mediated endocytosis for cholesterol acquisition. *The Journal of Cell Biology* **149**:167–180. doi: [10.1083/jcb.149.1.167](https://doi.org/10.1083/jcb.149.1.167).
- Csuros M**. 2010. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* **26**:1910–1912. doi: [10.1093/bioinformatics/btq315](https://doi.org/10.1093/bioinformatics/btq315).
- Cumbo VR**, Baird AH, Moore RB, Negri AP, Neilan BA, Salih A, van Oppen MJ, Wang Y, Marquis CP. 2013. *Chromera velia* is endosymbiotic in larvae of the reef corals *Acropora digitifera* and *A. tenuis*. *Protist* **164**:237–244. doi: [10.1016/j.protis.2012.08.003](https://doi.org/10.1016/j.protis.2012.08.003).
- Danne JC**, Gornik SG, Macrae JI, McConville MJ, Waller RF. 2013. Alveolate mitochondrial metabolic evolution: dinoflagellates force reassessment of the role of parasitism as a driver of change in apicomplexans. *Molecular Biology and Evolution* **30**:123–139. doi: [10.1093/molbev/mss205](https://doi.org/10.1093/molbev/mss205).
- Du P**, Li T, Wang X. 2011. Recent progress in predicting protein sub-cellular locations. *Expert Review of Proteomics* **8**:391–404. doi: [10.1586/epr.11.20](https://doi.org/10.1586/epr.11.20).
- Edgar RC**. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**:1792–1797. doi: [10.1093/nar/gkh340](https://doi.org/10.1093/nar/gkh340).
- Edgar RC**, Myers EW. 2005. PILER: identification and classification of genomic repeats. *Bioinformatics* **21**(Suppl 1): i152–i158. doi: [10.1093/bioinformatics/bti1003](https://doi.org/10.1093/bioinformatics/bti1003).
- Eisen JA**, Coyne RS, Wu M, Wu D, Thiagarajan M, Wortman JR, Badger JH, Ren Q, Amedeo P, Jones KM, Tallon LJ, Delcher AL, Salzberg SL, Silva JC, Haas BJ, Majoros WH, Farzad M, Carlton JM, Smith RK Jr, Garg J, Pearlman RE, Karrer KM, Sun L, Manning G, Elde NC, Turkewitz AP, Asai DJ, Wilkes DE, Wang Y, Cai H, Collins K, Stewart BA, Lee SR, Wilamowska K, Weinberg Z, Ruzzo WL, Wloga D, Gaertig J, Frankel J, Tsao CC, Gorovsky MA, Keeling PJ, Waller RF, Patron NJ, Cherry JM, Stover NA, Krieger CJ, del Toro C, Ryder HF, Williamson SC, Barbeau RA, Hamilton EP, Orias E. 2006. Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. *PLoS Biology* **4**:e286. doi: [10.1371/journal.pbio.0040286](https://doi.org/10.1371/journal.pbio.0040286).
- Ellinghaus D**, Kurtz S, Willhoeft U. 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**:18. doi: [10.1186/1471-2105-9-18](https://doi.org/10.1186/1471-2105-9-18).
- Emanuelsson O**, Brunak S, von Heijne G, Nielsen H. 2007. Locating proteins in the cell using TargetP, SignalP and related tools. *Nature Protocols* **2**:953–971. doi: [10.1038/nprot.2007.131](https://doi.org/10.1038/nprot.2007.131).
- Emanuelsson O**, Nielsen H, Brunak S, von Heijne G. 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of Molecular Biology* **300**:1005–1016. doi: [10.1006/jmbi.2000.3903](https://doi.org/10.1006/jmbi.2000.3903).
- Faith JJ**, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS. 2007. Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biology* **5**:e8. doi: [10.1371/journal.pbio.0050008](https://doi.org/10.1371/journal.pbio.0050008).
- Fedoroff NV**. 2012. Presidential address. Transposable elements, epigenetics, and genome evolution. *Science* **338**:758–767. doi: [10.1126/science.338.6108.758](https://doi.org/10.1126/science.338.6108.758).
- Fehrenbacher K**, Huckaba T, Yang HC, Boldogh I, Pon L. 2003. Actin comet tails, endosomes and endosymbionts. *The Journal of Experimental Biology* **206**:1977–1984. doi: [10.1242/jeb.00240](https://doi.org/10.1242/jeb.00240).
- Ferguson DJ**, Sahoo N, Pinches RA, Bumstead JM, Tomley FM, Gubbels MJ. 2008. MORN1 has a conserved role in asexual and sexual development across the Apicomplexa. *Eukaryot Cell* **7**:698–711. doi: [10.1128/EC.00021-08](https://doi.org/10.1128/EC.00021-08).
- Field HI**, Coulson RMR, Field MC. 2013. An automated graphics tool for comparative genomics: the Coulson plot generator. *BMC Bioinformatics* **14**:141. doi: [10.1186/1471-2105-14-141](https://doi.org/10.1186/1471-2105-14-141).
- Finn RD**, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer EL, Tate J, Punta M. 2014. Pfam: the protein families database. *Nucleic Acids Research* **42**:D222–D230. doi: [10.1093/nar/gkt1223](https://doi.org/10.1093/nar/gkt1223).
- Finn RD**, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research* **39**:W29–W37. doi: [10.1093/nar/gkr367](https://doi.org/10.1093/nar/gkr367).
- Flegontov P**, Michalek J, Janouskovec J, Lai DH, Jirku M, Hajduskova E, Tomcala A, Otto TD, Keeling PJ, Pain A, Obornik M, Lukeš J. 2015. Divergent mitochondrial respiratory chains in phototrophic relatives of apicomplexan parasites. *Molecular Biology and Evolution* **32**:1115–1131. doi: [10.1093/molbev/msv021](https://doi.org/10.1093/molbev/msv021).
- Flueck C**, Bartfai R, Niederwieser I, Witmer K, Alako BT, Moes S, Bozdech Z, Jenoe P, Stunnenberg HG, Voss TS. 2010. A major role for the *Plasmodium falciparum* ApiAP2 protein PfSIP2 in chromosome end biology. *PLoS Pathogens* **6**:e1000784. doi: [10.1371/journal.ppat.1000784](https://doi.org/10.1371/journal.ppat.1000784).
- Flutre T**, Duprat E, Feuillet C, Quesneville H. 2011. Considering transposable element diversification in de novo annotation approaches. *PLoS ONE* **6**:e16526. doi: [10.1371/journal.pone.0016526](https://doi.org/10.1371/journal.pone.0016526).
- Folch J**, Lees M, Sloane Stanley GH. 1957. A simple method for the isolation and purification of total lipides from animal tissues. *The Journal of Biological Chemistry* **226**:497–509.
- Foth BJ**, Goedecke MC, Soldati D. 2006. New insights into myosin evolution and classification. *Proceedings of the National Academy of Sciences of USA* **103**:3681–3686. doi: [10.1073/pnas.0506307103](https://doi.org/10.1073/pnas.0506307103).
- Francia ME**, Jordan CN, Patel JD, Sheiner L, Demerly JL, Fellows JD, de Leon JC, Morrisette NS, Dubremetz JF, Stripen B. 2012. Cell division in Apicomplexan parasites is organized by a homolog of the striated rootlet fiber of algal flagella. *PLoS Biology* **10**:e1001444. doi: [10.1371/journal.pbio.1001444](https://doi.org/10.1371/journal.pbio.1001444).

- Frenal K, Polonais V, Marq JB, Stratmann R, Limenitakis J, Soldati-Favre D. 2010. Functional dissection of the apicomplexan glideosome molecular architecture. *Cell Host & Microbe* **8**:343–357. doi: [10.1016/j.chom.2010.09.002](https://doi.org/10.1016/j.chom.2010.09.002).
- Frenal K, Soldati-Favre D. 2009. Role of the parasite and host cytoskeleton in Apicomplexa parasitism. *Cell Host & Microbe* **5**:602–611. doi: [10.1016/j.chom.2009.05.013](https://doi.org/10.1016/j.chom.2009.05.013).
- Gajria B, Bahl A, Brestelli J, Dommer J, Fischer S, Gao X, Heiges M, Iodice J, Kissinger JC, Mackey AJ, Pinney DF, Roos DS, Stoeckert CJ Jr, Wang H, Brunk BP. 2008. ToxoDB: an integrated *Toxoplasma gondii* database resource. *Nucleic Acids Research* **36**:D553–D556. doi: [10.1093/nar/gkm981](https://doi.org/10.1093/nar/gkm981).
- Gandhi M, Jangi M, Goode BL. 2010. Functional surfaces on the actin-binding protein coronin revealed by systematic mutagenesis. *The Journal of Biological Chemistry* **285**:34899–34908. doi: [10.1074/jbc.M110.171496](https://doi.org/10.1074/jbc.M110.171496).
- Gerstein MB, Rozowsky J, Yan KK, Wang D, Cheng C, Brown JB, Davis CA, Hillier L, Sisu C, Li JJ, Pei B, Harmanci AO, Duff MO, Djebali S, Alexander RP, Alver BH, Auerbach R, Bell K, Bickel PJ, Boeck ME, Boley NP, Booth BW, Cherbas L, Cherbas P, Di C, Dobin A, Drenkow J, Ewing B, Fang G, Fastuca M, Feingold EA, Frankish A, Gao G, Good PJ, Guigo R, Hammonds A, Harrow J, Hoskins RA, Howald C, Hu L, Huang H, Hubbard TJ, Huynh C, Jha S, Kasper D, Kato M, Kaufman TC, Kitchen RR, Ladewig E, Lagarde J, Lai E, Leng J, Lu Z, MacCoss M, May G, McWhirter R, Merrihew G, Miller DM, Mortazavi A, Murad R, Oliver B, Olson S, Park PJ, Pazin MJ, Perrimon N, Pervouchine D, Reinke V, Reymond A, Robinson G, Samsonova A, Saunders GI, Schlesinger F, Sethi A, Slack FJ, Spencer WC, Stoiber MH, Strasbourger P, Tanzer A, Thompson OA, Wan KH, Wang G, Wang H, Watkins KL, Wen J, Wen K, Xue C, Yang L, Yip K, Zaleski C, Zhang Y, Zheng H, Brenner SE, Graveley BR, Celniker SE, Gingeras TR, Waterston R. 2014. Comparative analysis of the transcriptome across distant species. *Nature* **512**:445–448. doi: [10.1038/nature13424](https://doi.org/10.1038/nature13424).
- Gordon JL, Sibley LD. 2005. Comparative genome analysis reveals a conserved family of actin-like proteins in apicomplexan parasites. *BMC Genomics* **6**:179. doi: [10.1186/1471-2164-6-179](https://doi.org/10.1186/1471-2164-6-179).
- Gournier H, Goley ED, Niederstrasser H, Trinh T, Welch MD. 2001. Reconstitution of human Arp2/3 complex reveals critical roles of individual subunits in complex structure and activity. *Molecular Cell* **8**:1041–1052. doi: [10.1016/S1097-2765\(01\)00393-8](https://doi.org/10.1016/S1097-2765(01)00393-8).
- Gouy M, Guindon S, Gascuel O. 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution* **27**:221–224. doi: [10.1093/molbev/msp259](https://doi.org/10.1093/molbev/msp259).
- Gschloessl B, Guermeur Y, Cock JM. 2008. HECTAR: a method to predict subcellular targeting in heterokonts. *BMC Bioinformatics* **9**:393. doi: [10.1186/1471-2105-9-393](https://doi.org/10.1186/1471-2105-9-393).
- Hafner MS, Sudman PD, Villablanca FX, Spradling TA, Demastes JW, Nadler SA. 1994. Disparate rates of molecular evolution in cospeciating hosts and parasites. *Science* **265**:1087–1090. doi: [10.1126/science.8066445](https://doi.org/10.1126/science.8066445).
- Hager KM, Striepen B, Tilney LG, Roos DS. 1999. The nuclear envelope serves as an intermediary between the ER and Golgi complex in the intracellular parasite *Toxoplasma gondii*. *Journal of Cell Science* **112**:2631–2638.
- Hintze JL, Nelson RD. 1998. Violin plots: a box plot-density trace synergism. *American Statistician* **52**:181–184. doi: [10.2307/2685478](https://doi.org/10.2307/2685478).
- Hirst J, Barlow LD, Francisco GC, Sahlender DA, Seaman MNJ, Dacks JB, Robinson MS. 2011. The fifth adaptor protein complex. *PLOS Biology* **9**:e1001170. doi: [10.1371/journal.pbio.1001170](https://doi.org/10.1371/journal.pbio.1001170).
- Hsiao CH, Luisa Hiller N, Haldar K, Knoll LJ. 2013. A HT/PEXEL motif in *Toxoplasma dense granule* proteins is a signal for protein cleavage but not export into the host cell. *Traffic* **14**:519–531. doi: [10.1111/tra.12049](https://doi.org/10.1111/tra.12049).
- Hu G, Cabrera A, Kono M, Mok S, Chaal BK, Haase S, Engelberg K, Cheemadan S, Spielmann T, Preiser PR, Gilberger TW, Bozdech Z. 2010. Transcriptional profiling of growth perturbations of the human malaria parasite *Plasmodium falciparum*. *Nature Biotechnology* **28**:91–98. doi: [10.1038/nbt.1597](https://doi.org/10.1038/nbt.1597).
- Hu K, Johnson J, Florens L, Fraunholz M, Suravajjala S, DiLullo C, Yates J, Roos DS, Murray JM. 2006. Cytoskeletal components of an invasion machine—the apical complex of *Toxoplasma gondii*. *PLOS Pathogens* **2**:e13. doi: [10.1371/journal.ppat.0020013](https://doi.org/10.1371/journal.ppat.0020013).
- Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD. 2013. REAPR: a universal tool for genome assembly evaluation. *Genome Biology* **14**:R47. doi: [10.1186/gb-2013-14-5-r47](https://doi.org/10.1186/gb-2013-14-5-r47).
- Janouškovec J, Horák A, Barott KL, Rohwer FL, Keeling PJ. 2012. Global analysis of plastid diversity reveals apicomplexan-related lineages in coral reefs. *Current Biology* **22**:R518–R519. doi: [10.1016/j.cub.2012.04.047](https://doi.org/10.1016/j.cub.2012.04.047).
- Janouškovec J, Horák A, Barott KL, Rohwer FL, Keeling PJ. 2013. Environmental distribution of coral-associated relatives of apicomplexan parasites. *The ISME Journal* **7**:444–447. doi: [10.1038/ismej.2012.129](https://doi.org/10.1038/ismej.2012.129).
- Janouškovec J, Horák A, Oborník M, Lukeš J, Keeling PJ. 2010. A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids. *Proceedings of the National Academy of Sciences of USA* **107**:10949–10954. doi: [10.1073/pnas.1003335107](https://doi.org/10.1073/pnas.1003335107).
- Janouškovec J, Tikhonenkov DV, Burki F, Howe AT, Kolisko M, Mylnikov AP, Keeling PJ. 2015. Factors mediating plastid dependency and the origins of parasitism in apicomplexans and their close relatives. *Proceedings of the National Academy of Sciences of USA*. doi: [10.1073/pnas.1423790112](https://doi.org/10.1073/pnas.1423790112).
- Johnson LS, Eddy SR, Portugaly E. 2010. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11**:431. doi: [10.1186/1471-2105-11-431](https://doi.org/10.1186/1471-2105-11-431).
- Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research* **110**:462–467. doi: [10.1159/000084979](https://doi.org/10.1159/000084979).
- Jurka J, Klonowski P, Dagman V, Pelton P. 1996. CENSOR—a program for identification and elimination of repetitive elements from DNA sequences. *Computers and Chemistry* **20**:119–121. doi: [10.1016/S0097-8485\(96\)80013-1](https://doi.org/10.1016/S0097-8485(96)80013-1).
- Kafsack BF, Rovira-Graells N, Clark TG, Bancells C, Crowley VM, Campino SG, Williams AE, Drought LG, Kwiatkowski DP, Baker DA, Cortes A, Llinas M. 2014. A transcriptional switch underlies commitment to sexual development in malaria parasites. *Nature* **507**:248–252. doi: [10.1038/nature12920](https://doi.org/10.1038/nature12920).

- Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2014. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Research* **42**:D199–D205. doi: [10.1093/nar/gkt1076](https://doi.org/10.1093/nar/gkt1076).
- Kaneko I, Iwanaga S, Kato T, Kobayashi I, Yuda M. 2015. Genome-wide identification of the target genes of AP2-O, a plasmodium AP2-family transcription factor. *PLOS Pathogens* **11**:e1004905. doi: [10.1371/journal.ppat.1004905](https://doi.org/10.1371/journal.ppat.1004905).
- Katinka MD, Duprat S, Cornillot E, Metenier G, Thomarat F, Prensier G, Barbe V, Peyretailade E, Brottier P, Wincker P, Delbac F, El Alaoui H, Peyret P, Saurin W, Gouy M, Weissenbach J, Vivares CP. 2001. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature* **414**:450–453. doi: [10.1038/35106579](https://doi.org/10.1038/35106579).
- Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**:772–780. doi: [10.1093/molbev/mst010](https://doi.org/10.1093/molbev/mst010).
- Kawase O, Nishikawa Y, Bannai H, Zhang H, Zhang G, Jin S, Lee EG, Xuan X. 2007. Proteomic analysis of calcium-dependent secretion in *Toxoplasma gondii*. *Proteomics* **7**:3718–3725. doi: [10.1002/pmic.200700362](https://doi.org/10.1002/pmic.200700362).
- Keeling PJ. 2004. Reduction and compaction in the genome of the apicomplexan parasite *Cryptosporidium parvum*. *Developmental Cell* **6**:614–616.
- Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* **14**:R36. doi: [10.1186/gb-2013-14-4-r36](https://doi.org/10.1186/gb-2013-14-4-r36).
- Klinger CM, Klute MJ, Dacks JB. 2013a. Comparative genomic analysis of multi-subunit tethering complexes demonstrates an ancient pan-eukaryotic complement and sculpting in Apicomplexa. *PLOS ONE* **8**:e76278. doi: [10.1371/journal.pone.0076278](https://doi.org/10.1371/journal.pone.0076278).
- Klinger CM, Nisbet RE, Ouologuem DT, Roos DS, Dacks JB. 2013b. Cryptic organelle homology in apicomplexan parasites: insights from evolutionary cell biology. *Current Opinion in Microbiology* **16**:424–431. doi: [10.1016/j.mib.2013.07.015](https://doi.org/10.1016/j.mib.2013.07.015).
- Kolpakov R, Bana G, Kucherov G. 2003. mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Research* **31**:3672–3678. doi: [10.1093/nar/gkg617](https://doi.org/10.1093/nar/gkg617).
- Kono M, Herrmann S, Loughran NB, Cabrera A, Engelberg K, Lehmann C, Sinha D, Prinz B, Ruch U, Heussler V, Spielmann T, Parkinson J, Gilberger TW. 2012. Evolution and architecture of the inner membrane complex in asexual and sexual stages of the malaria parasite. *Molecular Biology and Evolution* **29**:2113–2132. doi: [10.1093/molbev/mss081](https://doi.org/10.1093/molbev/mss081).
- Kořený L, Oborník M. 2011. Sequence evidence for the presence of two tetrapyrrole pathways in *Euglena gracilis*. *Genome Biology and Evolution* **3**:359–364. doi: [10.1093/gbe/evr029](https://doi.org/10.1093/gbe/evr029).
- Kořený L, Sobotka R, Janouskovec J, Keeling PJ, Oborník M. 2011. Tetrapyrrole synthesis of photosynthetic chromerids is likely homologous to the unusual pathway of apicomplexan parasites. *The Plant Cell* **23**:3454–3462. doi: [10.1105/tpc.111.089102](https://doi.org/10.1105/tpc.111.089102).
- Koumandou VL, Dacks JB, Coulson RM, Field MC. 2007. Control systems for membrane fusion in the ancestral eukaryote; evolution of tethering complexes and SM proteins. *BMC Evolutionary Biology* **7**:29. doi: [10.1186/1471-2148-7-29](https://doi.org/10.1186/1471-2148-7-29).
- Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ. 2009. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nature Methods* **6**:291–295. doi: [10.1038/nmeth.1311](https://doi.org/10.1038/nmeth.1311).
- Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology* **305**:567–580. doi: [10.1006/jmbi.2000.4315](https://doi.org/10.1006/jmbi.2000.4315).
- Kucera K, Koblansky AA, Saunders LP, Frederick KB, De La Cruz EM, Ghosh S, Modis Y. 2010. Structure-based analysis of *Toxoplasma gondii* profilin: a parasite-specific motif is required for recognition by Toll-like receptor 11. *Journal of Molecular Biology* **403**:616–629. doi: [10.1016/j.jmb.2010.09.022](https://doi.org/10.1016/j.jmb.2010.09.022).
- Kursula I, Kursula P, Ganter M, Panjekar S, Matuschewski K, Schuler H. 2008. Structural basis for parasite-specific functions of the divergent profilin of *Plasmodium falciparum*. *Structure* **16**:1638–1648. doi: [10.1016/j.str.2008.09.008](https://doi.org/10.1016/j.str.2008.09.008).
- Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution* **21**:1095–1109. doi: [10.1093/molbev/msh112](https://doi.org/10.1093/molbev/msh112).
- Le SQ, Dang CC, Gascuel O. 2012. Modeling protein evolution with several amino acid replacement matrices depending on site rates. *Molecular Biology and Evolution* **29**:2921–2936. doi: [10.1093/molbev/mss112](https://doi.org/10.1093/molbev/mss112).
- Lee I, Hong W. 2004. RAP—a putative RNA-binding domain. *Trends in Biochemical Sciences* **29**:567–570. doi: [10.1016/j.tibs.2004.09.005](https://doi.org/10.1016/j.tibs.2004.09.005).
- Leonardi R, Zhang YM, Rock CO, Jackowski S. 2005. Coenzyme A: back in action. *Progress in Lipid Research* **44**:125–153. doi: [10.1016/j.plipres.2005.04.001](https://doi.org/10.1016/j.plipres.2005.04.001).
- Leung KF, Dacks JB, Field MC. 2008. Evolution of the multivesicular body ESCRT machinery; retention across the eukaryotic lineage. *Traffic* **9**:1698–1716. doi: [10.1111/j.1600-0854.2008.00797.x](https://doi.org/10.1111/j.1600-0854.2008.00797.x).
- Li H, Child MA, Bogoy M. 2012. Proteases as regulators of pathogenesis: examples from the Apicomplexa. *Biochimica et Biophysica Acta* **1824**:177–185. doi: [10.1016/j.bbapap.2011.06.002](https://doi.org/10.1016/j.bbapap.2011.06.002).
- Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research* **13**:2178–2189. doi: [10.1101/gr.1224503](https://doi.org/10.1101/gr.1224503).
- Lim L, McFadden GI. 2010. The evolution, metabolism and functions of the apicoplast. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **365**:749–763. doi: [10.1098/rstb.2009.0273](https://doi.org/10.1098/rstb.2009.0273).
- Liu J, Guo W. 2012. The exocyst complex in exocytosis and cell migration. *Protoplasma* **249**:587–597. doi: [10.1007/s00709-011-0330-1](https://doi.org/10.1007/s00709-011-0330-1).

- Logan-Klumpler FJ**, de Silva N, Boehme U, Rogers MB, Velarde G, McQuillan JA, Carver T, Aslett M, Olsen C, Subramanian S, Phan I, Farris C, Mitra S, Ramasamy G, Wang H, Tivey A, Jackson A, Houston R, Parkhill J, Holden M, Harb OS, Brunk BP, Myler PJ, Roos D, Carrington M, Smith DF, Hertz-Fowler C, Berriman M. 2012. GeneDB—an annotation database for pathogens. *Nucleic Acids Research* **40**:D98–D108. doi: [10.1093/nar/gkr1032](https://doi.org/10.1093/nar/gkr1032).
- Machesky LM**, Atkinson SJ, Ampe C, Vandekerckhove J, Pollard TD. 1994. Purification of a cortical complex containing two unconventional actins from *Acanthamoeba* by affinity chromatography on profilin-agarose. *The Journal of Cell Biology* **127**:107–115. doi: [10.1083/jcb.127.1.107](https://doi.org/10.1083/jcb.127.1.107).
- Magnani E**, Sjolander K, Hake S. 2004. From endonucleases to transcription factors: evolution of the AP2 DNA binding domain in plants. *The Plant Cell* **16**:2265–2277. doi: [10.1105/tpc.104.023135](https://doi.org/10.1105/tpc.104.023135).
- Mazumdar J**, Striepen B. 2007. Make it or take it: fatty acid metabolism of apicomplexan parasites. *Eukaryot Cell* **6**:1727–1735. doi: [10.1128/EC.00255-07](https://doi.org/10.1128/EC.00255-07).
- McFadden GI**, Reith ME, Munholland J, Lang-Unnasch N. 1996. Plastid in human parasites. *Nature* **381**:482. doi: [10.1038/381482a0](https://doi.org/10.1038/381482a0).
- Meyer PE**, Lafitte F, Bontempi G. 2008. minet: a R/Bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics* **9**:461. doi: [10.1186/1471-2105-9-461](https://doi.org/10.1186/1471-2105-9-461).
- Miranda K**, Pace DA, Cintron R, Rodrigues JCF, Fang J, Smith A, Rohloff P, Coelho E, de Haas F, de Souza W, Coppens I, Sibley LD, Moreno SNJ. 2010. Characterization of a novel organelle in *Toxoplasma gondii* with similar composition and function to the plant vacuole. *Molecular Microbiology* **76**:1358–1375. doi: [10.1111/j.1365-2958.2010.07165.x](https://doi.org/10.1111/j.1365-2958.2010.07165.x).
- Moore RB**, Obornik M, Janouškovec J, Chrudimský T, Vancová M, Green DH, Wright SW, Davies NW, Bolch CJ, Heimann K, Slapeta J, Hoegh-Guldberg O, Logsdon JM, Carter DA. 2008. A photosynthetic alveolate closely related to apicomplexan parasites. *Nature* **451**:959–963. doi: [10.1038/nature06635](https://doi.org/10.1038/nature06635).
- Moriya Y**, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. 2007. KAAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research* **35**:W182–W185. doi: [10.1093/nar/gkm321](https://doi.org/10.1093/nar/gkm321).
- Morrison HG**, McArthur AG, Gillin FD, Aley SB, Adam RD, Olsen GJ, Best AA, Cande WZ, Chen F, Cipriano MJ, Davids BJ, Dawson SC, Elmendorf HG, Hehl AB, Holder ME, Huse SM, Kim UU, Lasek-Nesselquist E, Manning G, Nigam A, Nixon JE, Palm D, Passamaneck NE, Prabhu A, Reich CI, Reiner DS, Samuelson J, Svard SG, Sogin ML. 2007. Genomic minimalism in the early diverging intestinal parasite *Giardia lamblia*. *Science* **317**:1921–1926. doi: [10.1126/science.1143837](https://doi.org/10.1126/science.1143837).
- Morrisette NS**, Sibley LD. 2002. Cytoskeleton of apicomplexan parasites. *Microbiology and Molecular Biology Reviews* **66**:21–38. doi: [10.1128/MMBR.66.1.21-38.2002](https://doi.org/10.1128/MMBR.66.1.21-38.2002).
- Mullins RD**, Stafford WF, Pollard TD. 1997. Structure, subunit topology, and actin-binding activity of the Arp2/3 complex from *Acanthamoeba*. *The Journal of Cell Biology* **136**:331–343. doi: [10.1083/jcb.136.2.331](https://doi.org/10.1083/jcb.136.2.331).
- Mundwiler-Pachlatko E**, Beck HP. 2013. Maurer's clefts, the enigma of *Plasmodium falciparum*. *Proceedings of the National Academy of Sciences of USA* **110**:19987–19994. doi: [10.1073/pnas.1309247110](https://doi.org/10.1073/pnas.1309247110).
- Mutwil M**, Klie S, Tohge T, Giorgi FM, Wilkins O, Campbell MM, Fernie AR, Usadel B, Nikoloski Z, Persson S. 2011. PlaNet: combined sequence and expression comparisons across plant networks derived from seven species. *The Plant Cell* **23**:895–910. doi: [10.1105/tpc.111.083667](https://doi.org/10.1105/tpc.111.083667).
- Nevin WD**, Dacks JB. 2009. Repeated secondary loss of adaptin complex genes in the *Apicomplexa*. *Parasitology International* **58**:86–94. doi: [10.1016/j.parint.2008.12.002](https://doi.org/10.1016/j.parint.2008.12.002).
- Obornik M**, Modrý D, Lukeš M, Cernotiková-Stříbrná E, Cihlář J, Tesařová M, Kotabová E, Vancová M, Prášil O, Lukeš J. 2012. Morphology, ultrastructure and life cycle of *Vitrella brassicaformis* n. sp., n. gen., a novel chromerid from the Great Barrier Reef. *Protist* **163**:306–323. doi: [10.1016/j.protis.2011.09.001](https://doi.org/10.1016/j.protis.2011.09.001).
- Obornik M**, Vancová M, Lai DH, Janouškovec J, Keeling PJ, Lukeš J. 2011. Morphology and ultrastructure of multiple life cycle stages of the photosynthetic relative of *Apicomplexa*, *Chromera velia*. *Protist* **162**:115–130. doi: [10.1016/j.protis.2010.02.004](https://doi.org/10.1016/j.protis.2010.02.004).
- Okamoto N**, Keeling PJ. 2014. The 3D structure of the apical complex and association with the flagellar apparatus revealed by serial TEM tomography in *Psammoma pacifica*, a distant relative of the *Apicomplexa*. *PLOS ONE* **9**:e84653. doi: [10.1371/journal.pone.0084653](https://doi.org/10.1371/journal.pone.0084653).
- Otto TD**, Sanders M, Berriman M, Newbold C. 2010. Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. *Bioinformatics* **26**:1704–1707. doi: [10.1093/bioinformatics/btq269](https://doi.org/10.1093/bioinformatics/btq269).
- Pawlowski J**, Audic S, Adl S, Bass D, Belbahri L, Berney C, Bowser SS, Cepicka I, Decelle J, Dunthorn M, Fiore-Donno AM, Gile GH, Holzmann M, Jahn R, Jirku M, Keeling PJ, Kostka M, Kudryavtsev A, Lara E, Lukeš J, Mann DG, Mitchell EA, Nitsche F, Romeralo M, Saunders GW, Simpson AG, Smirnov AV, Spouge JL, Stern RF, Stoeck T, Zimmermann J, Schindler D, de Vargas C. 2012. CBOL protist working group: barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PLOS Biology* **10**:e1001419. doi: [10.1371/journal.pbio.1001419](https://doi.org/10.1371/journal.pbio.1001419).
- Pelletier L**, Stern C, Pypaert M, Sheff D. 2002. Golgi biogenesis in *Toxoplasma gondii*. *Nature* **418**:1–5. doi: [10.1038/nature00946](https://doi.org/10.1038/nature00946).
- Petersen TN**, Brunak S, von Heijne G, Nielsen H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods* **8**:785–786. doi: [10.1038/nmeth.1701](https://doi.org/10.1038/nmeth.1701).
- Petsalaki EI**, Bagos PG, Litou ZI, Hamodrakas SJ. 2006. PredSL: a tool for the N-terminal sequence-based prediction of protein subcellular localization. *Genomics, Proteomics & Bioinformatics* **4**:48–55. doi: [10.1016/S1672-0229\(06\)60016-8](https://doi.org/10.1016/S1672-0229(06)60016-8).
- Pieperhoff MS**, Schmitt M, Ferguson DJ, Meissner M. 2013. The role of clathrin in post-Golgi trafficking in *Toxoplasma gondii*. *PLOS ONE* **8**:e77620. doi: [10.1371/journal.pone.0077620](https://doi.org/10.1371/journal.pone.0077620).
- Pombert JF**, Blouin NA, Lane C, Boucias D, Keeling PJ. 2014. A lack of parasitic reduction in the obligate parasitic green alga *Helicosporidium*. *PLOS Genetics* **10**:e1004355. doi: [10.1371/journal.pgen.1004355](https://doi.org/10.1371/journal.pgen.1004355).

- Pollard TD**, Borisy GG. 2003. Cellular motility driven by assembly and disassembly of actin filaments. *Cell* **112**: 453–465. doi: [10.1016/S0092-8674\(03\)00120-X](https://doi.org/10.1016/S0092-8674(03)00120-X).
- Portman N**, Foster C, Walker G, Slapeta J. 2014. Evidence of intraflagellar transport and apical complex formation in a free-living relative of the Apicomplexa. *Eukaryot Cell* **13**:10–20. doi: [10.1128/EC.00155-13](https://doi.org/10.1128/EC.00155-13).
- Poulin B**, Patzewitz EM, Brady D, Silvie O, Wright MH, Ferguson DJ, Wall RJ, Whipple S, Guttery DS, Tate EW, Wickstead B, Holder AA, Tewari R. 2013. Unique apicomplexan IMC sub-compartment proteins are early markers for apical polarity in the malaria parasite. *Biology Open* **2**:1160–1170. doi: [10.1242/bio.20136163](https://doi.org/10.1242/bio.20136163).
- Quail MA**, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13**:341. doi: [10.1186/1471-2164-13-341](https://doi.org/10.1186/1471-2164-13-341).
- Quesneville H**, Nouaud D, Anxolabehere D. 2003. Detection of new transposable element families in *Drosophila melanogaster* and *Anopheles gambiae* genomes. *Journal of Molecular Evolution* **57**(Suppl 1):S50–S59. doi: [10.1007/s00239-003-0007-2](https://doi.org/10.1007/s00239-003-0007-2).
- Quigg A**, Kotabova E, Jaresova J, Kana R, Setlik J, Sediva B, Komarek O, Prasil O. 2012. Photosynthesis in *Chromera velia* represents a simple system with high efficiency. *PLOS ONE* **7**:e47036. doi: [10.1371/journal.pone.0047036](https://doi.org/10.1371/journal.pone.0047036).
- Radke JB**, Lucas O, de Silva EK, Ma Y, Sullivan WJ Jr, Weiss LM, Llinas M, White MW. 2013. ApiAP2 transcription factor restricts development of the *Toxoplasma* tissue cyst. *Proceedings of the National Academy of Sciences of USA* **110**:6871–6876. doi: [10.1073/pnas.1300059110](https://doi.org/10.1073/pnas.1300059110).
- Raffaele S**, Kamoun S. 2012. Genome evolution in filamentous plant pathogens: why bigger can be better. *Nature Reviews. Microbiology* **10**:417–430. doi: [10.1038/nrmicro2790](https://doi.org/10.1038/nrmicro2790).
- Ravindran S**, Boothroyd JC. 2008. Secretion of proteins into host cells by Apicomplexan parasites. *Traffic* **9**: 647–656. doi: [10.1111/j.1600-0854.2008.00723.x](https://doi.org/10.1111/j.1600-0854.2008.00723.x).
- Reid AJ**, Blake DP, Ansari HR, Billington K, Browne HP, Bryant J, Dunn M, Hung SS, Kawahara F, Miranda-Saavedra D, Malas TB, Mourier T, Naghra H, Nair M, Otto TD, Rawlings ND, Rivaille P, Sanchez-Flores A, Sanders M, Subramaniam C, Tay YL, Woo Y, Wu X, Barrell B, Dear PH, Doerig C, Gruber A, Ivens AC, Parkinson J, Rajandream MA, Shirley MW, Wan KL, Berriman M, Tomley FM, Pain A. 2014. Genomic analysis of the causative agents of coccidiosis in domestic chickens. *Genome Research* **24**:1676–1685. doi: [10.1101/gr.168955.113](https://doi.org/10.1101/gr.168955.113).
- Roiko MS**, Carruthers VB. 2009. New roles for perforins and proteases in apicomplexan egress. *Cellular Microbiology* **11**:1444–1452. doi: [10.1111/j.1462-5822.2009.01357.x](https://doi.org/10.1111/j.1462-5822.2009.01357.x).
- Roos DS**. 2005. Genetics. Themes and variations in apicomplexan parasite biology. *Science* **309**:72–73. doi: [10.1126/science.1115252](https://doi.org/10.1126/science.1115252).
- Russell K**, Hasenkamp S, Emes R, Horrocks P. 2013. Analysis of the spatial and temporal arrangement of transcripts over intergenic regions in the human malarial parasite *Plasmodium falciparum*. *BMC Genomics* **14**:267. doi: [10.1186/1471-2164-14-267](https://doi.org/10.1186/1471-2164-14-267).
- Rybakin V**, Clemen CS. 2005. Coronin proteins as multifunctional regulators of the cytoskeleton and membrane trafficking. *Bioessays* **27**:625–632. doi: [10.1002/bies.20235](https://doi.org/10.1002/bies.20235).
- Sakharkar KR**, Dhar PK, Chow VT. 2004. Genome reduction in prokaryotic obligatory intracellular parasites of humans: a comparative analysis. *International Journal of Systematic and Evolutionary Microbiology* **54**: 1937–1941. doi: [10.1099/ijs.0.63090-0](https://doi.org/10.1099/ijs.0.63090-0).
- Shoguchi E**, Shinzato C, Kawashima T, Gyoja F, Mungpakdee S, Koyanagi R, Takeuchi T, Hisata K, Tanaka M, Fujiwara M, Hamada M, Seidi A, Fujie M, Usami T, Goto H, Yamasaki S, Arakaki N, Suzuki Y, Sugano S, Toyoda A, Kuroki Y, Fujiyama A, Medina M, Coffroth MA, Bhattacharya D, Satoh N. 2013. Draft assembly of the *Symbiodinium minutum* nuclear genome reveals dinoflagellate gene structure. *Current Biology* **23**:1399–1408. doi: [10.1016/j.cub.2013.05.062](https://doi.org/10.1016/j.cub.2013.05.062).
- Stifflow CD**, Lefebvre PA. 2001. Assembly and motility of eukaryotic cilia and flagella. Lessons from *Chlamydomonas reinhardtii*. *Plant Physiology* **127**:1500–1507. doi: [10.1104/pp.010807](https://doi.org/10.1104/pp.010807).
- Simpson JT**, Durbin R. 2012. Efficient de novo assembly of large genomes using compressed data structures. *Genome Research* **22**:549–556. doi: [10.1101/gr.126953.111](https://doi.org/10.1101/gr.126953.111).
- Singh BK**, Sattler JM, Chatterjee M, Huttu J, Schuler H, Kursula I. 2011. Crystal structures explain functional differences in the two actin depolymerization factors of the malaria parasite. *The Journal of Biological Chemistry* **286**:28256–28264. doi: [10.1074/jbc.M111.211730](https://doi.org/10.1074/jbc.M111.211730).
- Sinha A**, Hughes KR, Modrzynska KK, Otto TD, Pfander C, Dickens NJ, Religa AA, Bushell E, Graham AL, Cameron R, Kafsack BF, Williams AE, Llinas M, Berriman M, Billker O, Waters AP. 2014. A cascade of DNA-binding proteins for sexual commitment and development in *Plasmodium*. *Nature* **507**:253–257. doi: [10.1038/nature12970](https://doi.org/10.1038/nature12970).
- Skillman KM**, Diraviyam K, Khan A, Tang K, Sept D, Sibley LD. 2011. Evolutionarily divergent, unstable filamentous actin is essential for gliding motility in apicomplexan parasites. *PLOS Pathogens* **7**:e1002280. doi: [10.1371/journal.ppat.1002280](https://doi.org/10.1371/journal.ppat.1002280).
- Soldati-Favre D**. 2008. Molecular dissection of host cell invasion by the apicomplexans: the glideosome. *Parasite* **15**:197–205.
- Stamatakis A**. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**:1312–1313. doi: [10.1093/bioinformatics/btu033](https://doi.org/10.1093/bioinformatics/btu033).
- Stanke M**, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research* **34**:W435–W439. doi: [10.1093/nar/gkl200](https://doi.org/10.1093/nar/gkl200).
- Stevens JM**, Galyov EE, Stevens MP. 2006. Actin-dependent movement of bacterial pathogens. *Nature Reviews. Microbiology* **4**:91–101. doi: [10.1038/nrmicro1320](https://doi.org/10.1038/nrmicro1320).
- Struck NS**, Herrmann S, Schmuck-Barkmann I, de Souza Dias S, Haase S, Cabrera AL, Treeck M, Bruns C, Langer C, Cowman AF, Marti M, Spielmann T, Gilberger TW. 2008. Spatial dissection of the cis- and trans-Golgi

- compartments in the malaria parasite *Plasmodium falciparum*. *Molecular Microbiology* **67**:1320–1330. doi: [10.1111/j.1365-2958.2008.06125.x](https://doi.org/10.1111/j.1365-2958.2008.06125.x).
- Stuart JM**, Segal E, Koller D, Kim SK. 2003. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**:249–255. doi: [10.1126/science.1087447](https://doi.org/10.1126/science.1087447).
- Sutak R**, Slapeta J, San Roman M, Camadro JM, Lesuisse E. 2010. Nonreductive iron uptake mechanism in the marine alveolate *Chromera velia*. *Plant Physiology* **154**:991–1000. doi: [10.1104/pp.110.159947](https://doi.org/10.1104/pp.110.159947).
- Tempel S**. 2012. Using and understanding RepeatMasker. *Methods in Molecular Biology* **859**:29–51. doi: [10.1007/978-1-61779-603-6_2](https://doi.org/10.1007/978-1-61779-603-6_2).
- Templeton TJ**, Iyer LM, Anantharaman V, Enomoto S, Abrahante JE, Subramanian GM, Hoffman SL, Abrahamsen MS, Aravind L. 2004a. Comparative analysis of Apicomplexa and genomic diversity in eukaryotes. *Genome Research* **14**:1686–1695. doi: [10.1101/gr.2615304](https://doi.org/10.1101/gr.2615304).
- Templeton TJ**, Lancto CA, Vigdorovich V, Liu C, London NR, Hadsall KZ, Abrahamsen MS. 2004b. The *Cryptosporidium* oocyst wall protein is a member of a multigene family and has a homolog in *Toxoplasma*. *Infection and Immunity* **72**:980–987. doi: [10.1128/IAI.72.2.980-987.2004](https://doi.org/10.1128/IAI.72.2.980-987.2004).
- Tenter AM**, Heckerth AR, Weiss LM. 2000. *Toxoplasma gondii*: from animals to humans. *International Journal for Parasitology* **30**:1217–1258. doi: [10.1016/S0020-7519\(00\)00124-7](https://doi.org/10.1016/S0020-7519(00)00124-7).
- Tomavo S**, Slomianny C, Meissner M, Carruthers VB. 2013. Protein trafficking through the endosomal system prepares intracellular parasites for a home invasion. *PLoS Pathogens* **9**:e1003629. doi: [10.1371/journal.ppat.1003629](https://doi.org/10.1371/journal.ppat.1003629).
- Trapnell C**, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* **7**:562–578. doi: [10.1038/nprot.2012.016](https://doi.org/10.1038/nprot.2012.016).
- Trecek M**, Sanders JL, Elias JE, Boothroyd JC. 2011. The phosphoproteomes of *Plasmodium falciparum* and *Toxoplasma gondii* reveal unusual adaptations within and beyond the parasites' boundaries. *Cell Host & Microbe* **10**:410–419. doi: [10.1016/j.chom.2011.09.004](https://doi.org/10.1016/j.chom.2011.09.004).
- Tsai IJ**, Otto TD, Berriman M. 2010. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biology* **11**:R41. doi: [10.1186/gb-2010-11-4-r41](https://doi.org/10.1186/gb-2010-11-4-r41).
- Van de Peer Y**, Frickey T, Taylor J, Meyer A. 2002. Dealing with saturation at the amino acid level: a case study based on anciently duplicated zebrafish genes. *Gene* **295**:205–211. doi: [10.1016/S0378-1119\(02\)00689-3](https://doi.org/10.1016/S0378-1119(02)00689-3).
- van Dooren GG**, Kennedy AT, McFadden GI. 2012. The use and abuse of heme in apicomplexan parasites. *Antioxidants & Redox Signaling* **17**:634–656. doi: [10.1089/ars.2012.4539](https://doi.org/10.1089/ars.2012.4539).
- Wilson D**, Charoensawan V, Kummerfeld SK, Teichmann SA. 2008. DBD—taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Research* **36**:D88–D92.
- Wong W**, Webb AI, Olshina MA, Infusini G, Tan YH, Hanssen E, Catimel B, Suarez C, Condrón M, Angrisano F, Nebi T, Kovar DR, Baum J. 2014. A mechanism for actin filament severing by malaria parasite actin depolymerizing factor 1 via a low affinity binding interface. *The Journal of Biological Chemistry* **289**:4043–4054. doi: [10.1074/jbc.M113.523365](https://doi.org/10.1074/jbc.M113.523365).
- Woo Y**, Affourtit J, Daigle S, Viale A, Johnson K, Naggert J, Churchill G. 2004. A comparison of cDNA, oligonucleotide, and Affymetrix GeneChip gene expression microarray platforms. *Journal of Biomolecular Techniques* **15**:276–284.
- Woo YH**, Li WH. 2011. Gene clustering pattern, promoter architecture, and gene expression stability in eukaryotic genomes. *Proceedings of the National Academy of Sciences of USA* **108**:3306–3311. doi: [10.1073/pnas.1100210108](https://doi.org/10.1073/pnas.1100210108).
- Wu TD**, Nacu S. 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**:873–881. doi: [10.1093/bioinformatics/btq057](https://doi.org/10.1093/bioinformatics/btq057).
- Xu P**, Widmer G, Wang Y, Ozaki LS, Alves JM, Serrano MG, Puiu D, Manque P, Akiyoshi D, Mackey AJ, Pearson WR, Dear PH, Bankier AT, Peterson DL, Abrahamsen MS, Kapur V, Tzipori S, Buck GA. 2004. The genome of *Cryptosporidium hominis*. *Nature* **431**:1107–1112. doi: [10.1038/nature02977](https://doi.org/10.1038/nature02977).
- Xu Z**, Wang H. 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research* **35**:W265–W268. doi: [10.1093/nar/gkm286](https://doi.org/10.1093/nar/gkm286).
- Zahradnickova H**, Tomcala A, Berkova P, Schneedorferova I, Okrouhlik J, Simek P, Hodkova M. 2014. Cost effective, robust, and reliable coupled separation techniques for the identification and quantification of phospholipids in complex biological matrices: application to insects. *Journal of Separation Science* **37**:2062–2068.
- Zdobnov EM**, Apweiler R. 2001. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**:847–848. doi: [10.1093/bioinformatics/17.9.847](https://doi.org/10.1093/bioinformatics/17.9.847).
- Zerbino DR**, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* **18**:821–829. doi: [10.1101/gr.074492.107](https://doi.org/10.1101/gr.074492.107).

Appendix 1

Genome characteristics.

The statistics of the genome assembly and annotation are shown in **Supplementary file 1**. There was bacterial contamination in 20% and 80% of the sequence reads in *Chromera* and *Vitrella*, respectively. There was a high amount of low-complexity DNA sequence repeats and TEs in *Chromera* (**Supplementary file 1**). By various bioinformatics methods ('Materials and methods'), we generated assemblies containing 5953 and 1064 scaffolds for *Chromera* and *Vitrella*, respectively. The total number of predicted genes differed between *Chromera* and *Vitrella* primarily due to significant differences in TE gene content between the two chromerids but the number of expressed genes was similar (**Supplementary file 1**).

We examined how genomes of the chromerids and other species were organized (**Supplementary file 1**). The median gene length is roughly the same between the two chromerids. The number of introns in a given gene was similar between the chromerids, although the size of introns was larger in *Chromera* than in *Vitrella* (**Supplementary file 1**). Compared to these chromerids, the number of introns in Apicomplexa was drastically less, raising the possibility that introns were compacted and reduced during apicomplexan evolution, which would need to be confirmed with further detailed investigation. For many genes (13,912 and 17,569 respectively for *Chromera* and *Vitrella*), we were able to assign 5' and 3' UTRs, using strand-specific transcriptome (RNA-seq) data sets. The distance between the protein-coding genes in *Vitrella* was short (median 92 base-pairs (bp)), indicating compactness of its genome. On the other hand, such distance was longer in *Chromera* (median 989 bp). Determining whether the common ancestor of chromerids had a compact genome or not would require analysis of genomes from more closely related species. There are three possible orientations by closely spaced neighboring genes can be clustered, that is, those with short intergenic spaces between the gene boundaries: tandem, head-to-head, or tail-to-tail. In both *Chromera* and *Vitrella* genomes, closely spaced (<1000 bp) genes were in head-to-head orientation more often than expected by chance (data not shown). It was previously shown that many neighboring genes in head-to-head clusters showed correlated expressions across various conditions; however, most of the co-expressions were modest; instead, head-to-head clustering is a major mechanism for stabilizing transcription of genes in fundamental cellular processes rather than for co-regulating the two genes (**Woo and Li, 2011; Russell et al., 2013**). Head-to-head clustering probably provided evolutionary and regulatory stability to genes involved in fundamental cellular processes. Other related species had different gene orientations, for example, the dinoflagellate *Symbiodinium microtinum* has tandem clusters driven by tandem gene duplication (**Shoguchi et al., 2013**). Given the dynamic nature of genome organization, we propose that different groups of species evolved different strategies for genome organization (**Woo and Li, 2011**).

Repetitive sequences constitute a significant proportion of eukaryotic genomes (**Fedoroff, 2012**). Thus, they play a significant role in evolution of host genomes. Systematic TE clustering, classification, and annotation were performed on 1064 *Vitrella* scaffolds (72.7 Mb genome—72,700,666 bp) and 5953 *Chromera* scaffolds (193.6 Mb genome—193,664,168 bp) *Chromera*. In both species, Class I elements (**Tempel, 2012**) make up a larger proportion of the genome than Class II elements (**Tempel, 2012**) (**Supplementary file 2**). The RT domain variation shows that *Eimeria tenella* TEs grouped separately and are not related to chromerid TEs (**Supplementary file 2**), suggesting gains of TEs in *E. tenella* (**Reid et al., 2014**) independently from chromerids. *Vitrella* forms a separate clade in the phylogenetic analysis of the RT domains.

Appendix 2

Metabolism.

Materials and methods

Reconstructing global metabolic map based on KEGG

Global metabolic pathways were mapped to KEGG metabolic pathways for the predicted protein-coding gene sets for the 26 species. KEGG ortholog (KO) assignments for the respective proteomes were made using the KO identification tools available on the KEGG website (<http://genome.jp/tools/kaas>) (Moriya *et al.*, 2007; Kanehisa *et al.*, 2014), and then the assigned KO numbers were used to identify and map metabolic pathways using the 'Search and color pathway' tool available on the KEGG site (http://genome.jp/kegg/tool/map_pathway2.html). The output of this mapping exercise was then manually inspected to compile the set of enzymes present in all major metabolic pathways (Figure 2—source data 1).

As KEGG is very strict in mapping orthologs and assigning KO numbers (so as to minimize false positives), we found numerous pathway holes (missing enzymes), many of which were readily apparent as false negatives. In order to resolve this, we then resorted to identifying orthologs from OrthoMCL-DB for all the genomes compared here (Chen *et al.*, 2006). The resulting ortholog assignments were then used to manually verify presence/absence of missing enzymes for filling pathway holes where possible. This curated data, based on both KEGG and OrthoMCL assignments, were used to generate the final mapping of enzymes to pathways, and using this info a metabolic pathway network was drawn to represent all major pathways involved in carbohydrate, energy, fatty acid, lipid, isoprene, steroid, amino acid, nucleotide, cofactor, polyamine, and redox metabolism (Figure 2—figure supplement 3).

Based on the enzymes mapped, we calculated the completeness of metabolic pathways by comparing the fraction of enzymes present for each pathway in each species. The complete set of enzymes mapped to each pathway (originally taken from KEGG and further curated to eliminate non-specific entries) is given in column B of **Supplementary file 3**. The fractional values were then color-coded and the resulting data are shown in **Figure 2B**. In order to visualize the retention, loss or gain of higher level metabolic functions, the fraction of enzymes mapped to these pathways is indicated as a pie chart for hypothetical ancestors of selected apicomplexan groups and chromerids (Figure 2B). We used presence of enzymes across the species and the phylogenetic relationship to infer presence of enzymes in the hypothetical ancestors based on Dollo parsimony (Csuros, 2010). Dollo parsimony is based on an assumption that it is unlikely that the same enzymes were gained multiple times independently in different lineages.

Phylogeny of heme pathway enzymes, the urea pathway CPS and enzymes involved in fatty acid biosynthesis.

Predicted proteins from *Vitrella* (Chromera heme pathway is already published [Kořený and Oborník, 2011; Kořený *et al.*, 2011]) were searched for enzymes involved in the synthesis of tetrapyrroles (aminolevulinic acid [ALA] synthase, ALAS; ALA dehydratase, ALAD; Porphobilinogen deaminase, PBGD; Uroporphyrinogen synthase UROS; Uroporphyrinogen decarboxylase, UROD; Coproporphyrinogen oxidase, CPOX; Protoporphyrinogen oxidase, PPOX; and Ferrochelatase FeCH). All genes identified were aligned to the homologs available in public databases such as NCBI and JGI using Muscle (Edgar, 2004), with the alignments further edited in SeaView (Gouy *et al.*, 2010). The results from these analyses are shown in **Supplementary file 4**. The same procedure but searching in the predicted proteomes of both chromerids was used to construct the alignment of carbamoyl phosphate synthases (CPS). Genes coding for enzymes containing ketoacyl synthase domain were searched using BLAST. Functional domains were searched using InterProScan (Zdobnov and Apweiler, 2001). Phylogenetic trees of all investigated enzymes were constructed using the ML approach

(RAxML [Stamatakis, 2014]), Bayesian inference (PHYLOBAYES [Lartillot and Philippe, 2004]), and a method designed to deal with amino acid saturation (AsaturA [Van de Peer et al., 2002]). ML trees were computed under the gamma corrected LG4X model of evolution as implemented in RAxML 7.4.8a using the rapid-bootstrap optimization algorithm in 1000 replicates. Bayesian phylogeny was inferred using empirical site-heterogeneous model C40 as implemented in Phylobayes 3.2f. Two independent chains were run until they converged (i.e., maximum observed discrepancy was lower than 0.2), and the effective number of model parameters was at least 100 after the first 1/5 generation was omitted from topology and posterior probability inference.

AsaturA trees were computed using a Poisson corrected LG model and the support was assessed from 1000 replicates. Sequences from *Vitrella* (all enzymes under investigation) and *Chromera* (CPS and FAS enzymes) were inspected for the presence of N-terminal leader sequences using SignalP (Bendtsen et al., 2004) and TargetP (Emanuelsson et al., 2007) software respectively, suggesting targeting to either mitochondrion (with mitochondrial transit peptide) or plastid (with bipartite leader composed of ER signal peptide and transit peptide).

Fatty acid synthesis pathway

C. velia cells were grown in the f2 medium. Cultures were kept in 25 cm² flasks under artificial light with photoperiod 12/12, light exposure between 70 and 120 μmol/m²/s and temperature of 26°C. 1 ml of *C. velia* stationary culture was added to each flask with 20 ml of f2 solution. The cultures were grown for one month to reach a high density of cells. Since triclosan is not soluble in water, dimethyl sulfoxid (DMSO) was used as a soluble mediator. Four experimental groups were established: control, control with DMSO, *Chromera* treated with triclosan in concentrations of 1 mM and 0.5 mM, respectively. After 16 days of incubation, cultures were harvested via centrifugation, and pellets were stored in –20°C for subsequent lipid extraction. Homogenization of algal sample was achieved by Mini-beadbeater (Biospec Products). Homogenates were dried and weighted. Lipids were extracted using on chloroform and methanol, as described before (Folch et al., 1957). An aliquot of 100 μl volume was subjected to HPLC ESI/MS. The technique was performed on an ion trap LTQ mass spectrometer coupled to Allegro ternary HPLC system equipped with Accela autosampler with the thermostat chamber (all by Thermo, San Jose, CA, USA). 5 μl of sample was injected into a Gemini column 250 × 2 mm i.d. 3 μm (Phenomenex, Torrance, CA, USA). The mobile phase consisted of (A) 5 mmol/l ammonium acetate in methanol, (B) water, and (C) 2-propanol. The analysis was completed within 80 min with a flow rate of 200 μl/min by following gradient of 92% A and 8% B in 0–5 min, then 100% A till the 12th minute, subsequently increasing the phase C to 60% till 50 min and holding for 15 min and then in at the 65th minute returning back to the 92:8% A:B mixture and 10 min to column conditioning. The column temperature was maintained at 30°C. The mass spectrometer was operated in the positive and negative ion detection modes at +4 kV and –4 kV with capillary temperature at 220°C. Nitrogen was employed as shielding and auxiliary gas for both polarities. Mass range of 140–1400 Da was scanned every 0.5 s to obtain the full scan ESI mass spectra of lipids. For investigation of the lipid molecules structures the collisionally induced decomposition multi-stage ion trap tandem mass spectra MS² in both polarity settings were simultaneously recorded with a 3 Da isolation window. Maximum ion injection time was 100 ms, and normalized collision energy was 35%. Major phospholipids, galactolipids, and neutral lipids molecular species that are detected were separated by reversed-phase HPLC. The structure of each entity was identified by MS² experiments in positive or negative mode. Peak areas for each detected lipid component were summarized and their relative contents estimated to sum of all obtained peaks.

Raw extracted lipids have to be transformed to methylesters of fatty acids (FAMES) to enable application of the GC technique. For this purpose sodium methoxide was employed as a transesterification reagent, as previously described (Zahradnickova et al., 2014). FAMES were then analyzed by GC/FID. Hydrocarbon with 26-carbon chain was chosen as an internal standard. The chromatography was performed using gas chromatograph GC-2014 (Shimadzu) equipped by with column BPX70 (SGE)—0.22 mm ID; 0.25 μm film; 30 m length. μl of derivatized

sample was injected via autosampler and injector AOC—20i (Shimadzu) to the column in split mode (split ratio 10). The temperature of the injector was 220°C. The starting temperature of the column was 120°C holding for 4 min. Then, the temperature increased to the 180°C in at the rate of 10°C per minute, and after that 7°C per minute to 230°C. Temperature of the flame ionization detector was 260°C. The whole analysis takes took approximately 20 min. H₂ was used as a carrier gas. For the identification of particular FAs, a mixture of 37 standards purchased from Supelco Inc. was used.

Results and discussion

Global metabolic map

Metabolic annotations based on ortholog assignments with KEGG and OrthoMCL database showed that chromerids contain all major primary metabolic pathways typically found in free-living unicellular eukaryotes (**Figure 2—figure supplement 3**). 2918 *Chromera* proteins and 2985 *Vitrella* proteins were assigned KO numbers, from which 432 *Chromera* (1.3% of proteome) and 425 *Vitrella* proteins (1.8% of proteome), respectively, were identified as enzymes with a metabolic function based on EC number association.

In support of their autotrophic lifestyle, the chromerids appear to be capable of generating de novo all primary carbon metabolites such as the various sugars and other reduced carbon compounds (presumably via photosynthesis and associated carbon fixation pathways), amino acids, nucleotides, fatty acids and lipids, isoprene and steroid derivatives, and most vitamins (except biotin and vitamin B12). These organisms are also capable of assimilating both nitrate and sulfite and can generate energy from photosynthesis as well as mitochondrial respiration. A full complement of enzymes involved in sugar and sugar derivative metabolism, such as glycolysis, Krebs's cycle, pentose phosphate pathway, inositol mono- and poly-phosphate formation, polysaccharide formation, and amino- and nucleotide-sugar formation, is encoded by the chromerids. Chromerids are also capable of synthesizing sulfoquinovosyl-diacyl-glycerol lipids, which are found associated with the chloroplast in photosynthetic organisms.

Figure 2—figure supplement 3 illustrates a complete representation of all major pathways mapped to chromerids in comparison to selected apicomplexan lineages.

Generally, *Chromera* and *Vitrella* have similar sets of metabolic enzymes. Enzymes for the oxidative arm of pentose phosphate pathway, conversion of diacyl-glycerol to phosphatidyl ethanolamine, phosphatidyl ethanolamine to phosphatidyl serine, and XppppX to XTP are absent in *Chromera*, while, on the other hand, the enzymes involved in conversion of glucose-1P to UDP-glucose, and cytidine to uridine are missing in *Vitrella*. One major difference between the two chromerids is that the complex III of the mitochondrial respiratory chain (cytochrome c reductase) is missing in *Chromera*, but present in *Vitrella* (**Flegontov et al., 2015**). This feature of the *Chromera* mitochondrion, that is, absence of complex III but presence of complex IV, makes it unique among all mitochondria and mitochondria-derived organelles.

The crucial enzyme for the urea pathway, mitochondrially targeted carbamoyl phosphate synthase (CPS) (**Allen et al., 2011**), is absent from both chromerids. However, while *Chromera* contains single CPS involved in pyrimidine biosynthesis, *Vitrella* genome encodes two CPSs. But both these genes are closely related suggesting they are recent duplicates (ML tree is shown in **Figure 2—figure supplement 2A**) and they lack a mitochondrial leader sequence at the N-terminus (data not shown). This means that *Vitrella* duplicated CPSs are not likely to be involved in urea cycle. In contrast to *Vitrella*, *Chromera* lacks the gene encoding argininosuccinate lyase (ASL), an enzyme of the ornithin cycle.

Plastid-related metabolic pathways

Chromerid photosystems have a reduced set of genes similarly to that of other algae with a complex plastid. The patterns of reduction were lineage-specific, even between the two known chromerids. We found *psbM*, *petN*, *psaJ*, *Psb27*, *Ycf4*, and *Ycf44* genes in *Vitrella* but absent in *Chromera*; vice versa, *Ycf39* and *Ycf54* are absent in *Vitrella* but present in *Chromera* (**Supplementary file 3**). This demonstrates that the plastid of *Chromera* is more diverse and reduced when compared to *Vitrella* with respect to both the composition of photosystems and the number of genes encoded in the plastid genome (**Janoušková et al., 2010**). In spite of substantial reduction of photosystems, photosynthesis in *Chromera* is highly efficient (**Quigg et al., 2012**).

The heme pathway in *Vitrella* is homologous to that found in *Chromera* (**Kořený et al., 2011**) for most of the involved enzymes (ALAS, ALAD, PBGD, UROS, PPOX), however, *Vitrella* and *Chromera* do not constitute sister groups in the CPOXs and FeCH trees (trees not shown). Some enzymes are not present in single copies (UROD) in *Vitrella*, in contrast to *Chromera*, where three orthologs originating from cyanobacteria, endosymbiont nucleus, and exosymbiont nucleus are present (**Kořený et al., 2011**). For some investigated enzymes (UROD, CPOX), only the endoplasmic reticulum signal peptide was found with transit peptide missing from the sequence, suggesting their possible location in the endoplasmic reticulum or periplastidal space.

Genes containing the ketoacyl synthase domain and thus likely involved in fatty acid or polyketide synthesis were searched in the genomes of chromerids. We found that both algae possess multi-modular enzymes responsible for fatty acid synthesis type I (FASI), similar to some apicomplexan parasites, such as *Cryptosporidium* spp. and *T. gondii*. The longest multi-modular enzyme found in *Chromera* contains five multi-domain modules, reaching over 11,000 amino acids in length (**Figure 2—figure supplement 2B**).

Evolution of metabolic pathways

Apicomplexan parasites differ drastically from each other in their metabolic functions, and have a significantly reduced metabolic capability in comparison to the chromerids. Apicomplexans are non-photosynthetic and therefore lack all associated metabolic activities including photosynthetic carbon fixation. Interestingly, however, all plastid-bearing parasites have retained only the ferredoxin-NADP⁺ reductase (FNR)/ferredoxin redox system of the photosynthetic electron transport (**Lim and McFadden, 2010**). In photosynthetic organisms, this redox system mediates the transfer of electrons originating from water to NADP⁺, resulting in the formation of NADPH (cofactor for fatty acid biosynthesis and Calvin cycle), and it is likely that this role is conserved in chromerids. In apicomplexans, the source of electrons for generating reduced ferredoxin is not clear, but it is evident that reduced ferredoxin is required for generating reducing equivalents and is a cofactor for several plastid-associated enzymes, including those involved in isoprene synthesis (**Lim and McFadden, 2010**). Other notable pathways missing in apicomplexans but present in the chromerids include the following: glyoxalate pathway; steroid biosynthesis; synthesis of aromatic and branched-chain amino acids; purine synthesis; and synthesis of cofactors such as thiamine, riboflavin, and nicotinate ribonucleotide.

Despite the reduced metabolic capabilities, certain core metabolic functions are conserved in chromerids as well as in all apicomplexan parasites, and many of these are likely to be essential. These pathways include: glycolysis; synthesis of ubiquinone, inositol-P derivatives, GPI-anchor, mono-, di- and tri-acyl glycerol, isoprene derivatives, and N-glycans; a subset of scavenge reactions for purine and pyrimidine bases and their conversion to nucleotides; glycine-serine inter-conversion; one-carbon folate cycle and S-adenosyl-methionine formation. There are many metabolic pathways that are retained in specific apicomplexan lineages but shared with the chromerids (see **Figure 2B** and **Figure 2—figure supplement 3**). The following are notable examples of pathways shared between chromerids and an apicomplexan lineage: with *Plasmodium*, polyamine synthesis; with *Cryptosporidium*, conversion of serine to tryptophan; with *Toxoplasma*, branched-chain amino acid degradation, synthesis of aspartate, lysine, and methionine; synthesis of molybdopterin, bipterin, pyridoxal-phosphate, and pantothenate

cofactors. Surprisingly, with respect to the chromerids and other apicomplexans, *Cryptosporidium* appears to be the only apicomplexan lineage to have gained a metabolic function of conversion of thymidine to dTMP by thymidine kinase. In addition, the type I and II pathways for fatty acids biosynthesis show lineage-specific distribution in apicomplexans and chromerids (**Figure 2—figure supplement 3**).

We can also find example of pathways that have been lost in a lineage-specific manner. For example, the ability to synthesize the di-saccharide trehalose is missing only in *Plasmodium*. However, the most dramatic loss of metabolic function in a single lineage can be found in cryptosporidia. These parasites are devoid of all plastid- and mitochondria-associated metabolic functions and other pathways involved in the synthesis of ribose-P, pyrimidine, most amino acids, heme, fatty acids (de novo), and isoprene units. It seems that the lack of mitochondrial oxidative pathways in cryptosporidia led to loss of the ability to generate flavin nucleotide (FMN/FAD) and lipoic acid cofactors.

In order to cope with loss of metabolic pathways, parasites have evolved various mechanisms for scavenging the required nutrients and metabolites from their respective hosts. For example, metabolites such as heme, fatty acids, steroids (specifically cholesterol), and sphingolipids are known to be scavenged by various apicomplexans as needed from their respective hosts.

According to the metabolic pathway maps, certain metabolic functions, which are coupled to each other, have been either retained or lost concomitantly in various species (**Figure 2—figure supplement 3**). For example, piroplasms and cryptosporidia lack de novo fatty acid biosynthesis along with the pyruvate dehydrogenase enzyme complex (plastid-associated in apicomplexans), which is known to supply acetyl CoA units for fatty acid synthesis. On the other hand, these parasites have retained the ability to convert pantothenate to co-enzyme A, which is required for the activation of fatty acids scavenged from the hosts (**Leonardi et al., 2005**). Similarly, activities of the serine hydroxyl methyl transferase and thymidylate synthase enzymes are coupled to each other and to one-carbon folate metabolism. Therefore, these three metabolic functions are retained in all parasites.

Appendix 3

Endomembrane trafficking system.

Materials and methods

The predicted proteomes of the 26 species in **Figure 1** have been searched for endomembrane trafficking components. Initial homology searching was carried out using BLAST (**Altschul et al., 1990**). Known sequences from human (*Homo sapiens*) and yeast (*Saccharomyces cerevisiae*) were used to search the proteomes of each organism including *Chromera* and *Vitrella* to identify potential homologs of proteins implicated in endomembrane trafficking. Any sequences scoring an initial E value of 0.05 or lower were subjected to confirmation by reciprocal BLASTP. This involved the use of candidate homologous sequences as queries against the relevant *H. sapiens* or *S. cerevisiae* genome. Sequences that retrieved the query sequence, or named homologs/paralogs/isoforms thereof, first with an E value of 0.05 or lower were considered true homologs.

Additional searches were carried out using HMMER (**Finn et al., 2011**). The HMMs for the initial queries were built and used to search each proteome. Top hits based on BLASTP results with E values less than 0.05 were considered confirmed homologs, and not subjected to further analysis. Subunits with significant HMMER hits were further investigated by the reciprocal BLASTP as described above. Further HMMER searches were carried out with the addition of homologous sequences from *Bigelowiella natans*, *Phytophthora infestans*, and *T. gondii* to the original HMMs. Results were analyzed identically to the first round. All identified endomembrane components are listed in **Figure 2—source data 2**.

To identify homologous proteins not predicted by the gene prediction software, we used TBLASTN with the homologous protein from the closest related organism in our data set against scaffolds and contigs; E value cut-off was identical to BLASTP analysis. We utilized BLASTP to search either genome with an identified homolog from the other, if it was present. The final results are summarized in **Figure 2—figure supplement 4** using the Coulson Plot Generator software (**Field et al., 2013**).

Results and discussion

Apicomplexa possess numerous unusual features in their membrane trafficking systems. Non-canonical membranous inclusions such as the invasion organelles, the micronemes, rhoptries, and dense granules are present (**Baum et al., 2008**). Though canonical, stacked, Golgi bodies are present in *T. gondii* (**Pelletier et al., 2002**), other apicomplexan species possess Golgi bodies with aberrant morphology and unusual characteristics (**Struck et al., 2008**). Combined with other organelle destinations such as mitochondria, digestive vacuoles involved in hemoglobin catabolism in *P. falciparum*, and plant-like lytic vacuoles in *T. gondii* (**Miranda et al., 2010**), specificity of protein and lipid components of these various organelles suggest a need for unique trafficking pathways mediated by distinct protein machinery.

Interestingly, previous studies demonstrated the loss of trafficking machinery in Apicomplexa, including three key sets of proteins in the ESCRT machinery (**Leung et al., 2008**), adaptor protein complex (AP) families (**Nevin and Dacks, 2009; Hirst et al., 2011**), and multi-subunit tethering complexes (MTCs) (**Koumandou et al., 2007; Klinger et al., 2013a**) have been published. Several of the aforementioned families are involved in trafficking within the late endosomal system in opisthokont models and so may be associated with the evolution of the rhoptries and micronemes within the apicomplexan or myzozoan lineage. Consistent with this

idea, some cases of reduction were not limited to Apicomplexa, and could be observed in the sister phyla of the ciliates and dinoflagellates.

This pattern of loss raises the question of what losses correlate with the transition to parasitism and which are pre-adaptive, arising more deeply in the lineage. The unique phylogenetic position of chromerids (*Janouškovec et al., 2010, 2012; Oborník et al., 2012*) allows finer dissection of the patterns of retention/loss observed previously. Hence, we chose to focus on detailed characterization of the three previously studied sets of membrane trafficking machinery in the predicted proteomes of *Chromera* and *Vitrella*, together with 24 closely related organisms for comparison.

ESCRT machinery

The ESCRT machinery is a set of five sub-complexes involved in recognition of ubiquitylated proteins and recruitment to the multi-vesicular body (MVB)/late endosome for degradation (*Leung et al., 2008*). Most eukaryotes, including *Chlamydomonas reinhardtii* and our representative stramenopile taxa (*Thalassiosira pseudonanna*, *Phaeodactylum tricornutum*, *Ectocarpus siliculosus*, and *Pythium ultimum*), have a complete set of the ESCRT machinery, suggesting that the ancestor of alveolates, and indeed the Last Eukaryotic Common Ancestor (LECA), likely had it. Though this ancestral complement appears to have been reduced in ciliates in the ESCRTI and III complexes, and a few components are missing from dinoflagellate taxa, numerous gene duplications have occurred as well, suggesting sculpting of the machinery. By comparison, apicomplexan parasites exhibit significant reductions in their ESCRT machinery (*Leung et al., 2008*). Cryptosporidia, coccidia, and plasmodia appear to lack any subunits of the ESCRTI and II complexes. ESCRTIII conservation is better, though no apicomplexan encodes Vps24, and multiple taxa have lost Vps20 as well. A similar pattern is seen for the ESCRTIII-a machinery, with piroplasmids encoding only Vps46 and Vps4. Coccidia additionally encode Vps31, and cryptosporidia Vps60, whereas plasmodia encode all subunits (rodent parasites like *Plasmodium chabaudi*), or lack Vps31 (human or simian parasites like *P. falciparum*). *Chromera* and *Vitrella* possess all ESCRT subunits except for the ESCRT-III component CHMP7, which is rarely found outside the opisthokont supergroup (*Leung et al., 2008*). This observation suggests two conclusions regarding the evolution of the ESCRT machinery within alveolates: massive gene loss within the Apicomplexa occurred recently, after the split from the proto-apicomplexan ancestor, and some losses of machinery shared between apicomplexans and other alveolates are due to independent losses. An excellent example of this latter case is that of Vps37, which is present only in chromerids, but in no other alveolate included in the current study, suggesting its function was dispensable in a large number of lineages.

APs

The APs are heterotetrameric complexes that select cargo for inclusion into transport vesicles at organelles of the late secretory system and endocytic system. AP1 and AP3 are involved in the transport between the trans-Golgi network (TGN) and endosomes. AP2 is involved in the transport from the cell surface. AP4 is involved in TGN transport to either endosomes or the cell surface, while the recently described AP5 complex is involved in the transport from late endosomes back to early endosomes. All five complexes are ancient, having likely been present in the LECA (*Nevin and Dacks, 2009; Hirst et al., 2011*). However, the complexes have also been secondarily lost on multiple occasions as well. Outgroup taxa in our data set possess AP1-4 complexes, with the exception of *C. reinhardtii* lacking AP3, but only *Symbiodinium minutum* possesses an AP5 complex.

Apicomplexa display higher variability in AP complex retention. With the exception of AP2M in cryptosporidia, all taxa retain full AP1, 2, and 4 complexes. Piroplasmids lack all subunits of the AP3 complex, and together with *P. falciparum* and *Plasmodium reichenowi*, lack AP5 as well. Other plasmodia possess all AP5 subunits with the exception of the mu subunit. This result was unexpected, based on the usual patterns of conservation seen across *Plasmodium* species.

Presence of AP5 in the majority of these organisms suggests the exciting possibility of a novel trafficking pathway absent from the comparatively well-studied human parasite *P. falciparum*. Additionally, our increased taxon sampling has suggested that AP5 may be well conserved across Myzozoa, a result otherwise indeterminable from previous studies of this protein family (**Hirst et al., 2011**). Cryptosporidia also lack AP3, but unlike piroplasmids, they possess almost a complete AP5 complex, missing only the sigma subunit. Coccidia are the exception, possessing all five AP complexes in their entirety. Excitingly, *Chromera* and *Vitrella*, like coccidia, possess a complete complement of adaptin subunits, suggestive of a more complete set of trafficking pathways to endosomal organelles in these organisms.

MTCs

The MTCs are an assembly of heteromeric protein complexes involved in the first stage of vesicle fusion and delivery of contents from a transport vesicle to a destination organelle. Each one is specific to an organelle or transport pathway and all eight complexes have been deduced as present in the LECA, with some interesting cases of secondary loss. While *C. reinhardtii* and the stramenopiles encode a complete set of MTC machinery, several of these MTCs have interesting patterns of conservation, specifically in the Apicomplexa (**Klinger et al., 2013a**).

The conservation of the TRAPP I–II complexes is unclear through eukaryotes and clear patterns are difficult to draw. However, the apparent absence of the entire TRAPP II complex in *Vitrella* may be due to gaps/biases/absences in sequencing, protein prediction, or analysis, but has interesting ramifications if proven to be a real biological phenomenon.

Exocyst is involved in diverse processes, all of which involve polarized exocytosis (**Liu and Guo, 2012**). *Tetrahymena* appears to encode only four of the Exocyst subunits. None of the eight subunits were identifiable in *Chromera*, *Vitrella*, nor in any of the Apicomplexa or dinoflagellates. This confirms, and extends, a previous result suggesting the absence of this complex within the Myzozoa, suggesting a bona fide ancestral loss concurrent with the acquisition of an apical complex that could have served an analogous tethering function for secretory organelles.

COG is an octameric complex involved in tethering at the Golgi body (**Tomavo et al., 2013; Klinger et al., 2013b**). The COG complex is poorly conserved in Apicomplexa and a ciliate *Tetrahymena thermophila* only encodes half of the COG subunits. In contrast, all eight COG subunits are present in *Chromera* and *Vitrella*. The retention of a complete COG complex in both *Chromera* and *Vitrella* contrasts with the substantial loss of subunits in Apicomplexa, especially outside the coccidians (**Klinger et al., 2013b**) (**Figure 2—figure supplement 4**). Notably, this conservation is consistent with the presence of robust, stacked Golgi bodies in *Chromera* (**Oborník et al., 2011**) and *T. gondii* (**Pelletier et al., 2002**), compared to aberrant morphology in other Apicomplexa.

The complexes of CORVET and HOPS mediate tethering at the early and late endosomes (**Tomavo et al., 2013; Klinger et al., 2013b**). They share a core of four subunits with complex-specific proteins (Vps3 and 8 for CORVET and Vps39/41 for HOPS). Though all taxa encode the complete VpsC core of the HOPS/CORVET complex, all taxa except for *T. gondii* only appear to encode the CORVET-specific interactor Vps3. *Chromera* and *Vitrella*, like Apicomplexa, possess the entire VpsC core complements as well as the HOPS component Vps41 and both CORVET components.

Chromerids exhibit complex life cycles, from immotile vegetative cells to multi-cellular sporangia, and occasionally motile flagellated cells. Both lineages contain numerous potential locales for intracellular trafficking including mitochondria, plastid, starch granules, flagella, micronemes, and, in *Chromera*, the chromerosome. Additionally, vesicular traffic to the sporangial/cyst wall has been visualized in both lineages (**Oborník et al., 2012**). Our results indicate that chromerids possess an appropriately complex complement of membrane trafficking machinery to achieve these requirements.

Though MVBs have not been explicitly imaged or characterized in either lineage to date, both *Chromera* and *Vitrella* encode a complete set of ESCRT machinery, suggestive of the presence of functional MVBs. These may play a key role in modulating surface protein expression in various life cycle stages. Importantly, the close evolutionary position of *Chromera* and *Vitrella* to Apicomplexa suggests that the extensive decrease in ESCRT subunit conservation in Apicomplexa occurred in the immediate ancestor and is not an ancestral feature of a more inclusive group (Leung et al., 2008) (Figure 2—figure supplement 4). Particularly, the lack of some ESCRT subunits such as Vps37 in ciliates and dinoflagellates is most parsimoniously attributed to multiple independent losses. Further evidence for a complete set of ESCRT machinery in the last common alveolate ancestor comes from the conservation of all subunits to the exclusion of CHMP7 in the outgroup stramenopile taxa and in *C. reinhardtii*. The absence of CHMP7 in all taxa is not unusual, as it is lost in numerous taxa across eukaryotes (Leung et al., 2008).

Conservation of adaptin subunits is striking, particularly the complete retention of AP5 in chromerids. In an initial study of seven organisms from the SAR supergroup (the group in which chromerids belong to), only two (*B. natans* and *T. gondii*) were found to encode the complex; conservation across eukaryotes was similarly sparse (Hirst et al., 2011). The presence of a complete AP5 complex in chromerids and coccidians may be indicative of a conserved function in both lineages. Likewise, the retention of an almost complete AP5 in cryptosporidia and plasmodia may have functional significance or may simply represent a reductive evolutionary process that has not yet reached completion. The complete lack of AP5 in *P. falciparum* and *P. reichenowi* supports the latter view. As with the ESCRT complexes, the presence of AP1-5 in chromerids suggests the loss of AP3 and AP5 observed in some Apicomplexa is secondary, as well as the loss of AP5 in *Perkinsus marinus*, and in both ciliate lineages.

Presence of a complete VpsC core along with an additional CORVET subunit Vps3 in the majority of apicomplexan genomes suggests the potential for a modified HOPS/CORVET complex that interacts with Rab5 to direct tethering at the micronemes/rhoptries. This is in keeping with the view of rhoptries/micronemes as divergent endolysosomal organelles (Klinger et al., 2013b). However, chromerids do not appear to possess rhoptries, although chromerids possess cellular components analogous to micronemes (Obornik et al., 2011, 2012). More HOPS/CORVET subunits were found to be conserved in *T. gondii*, which are the only apicomplexan to date to be described as possessing a canonical lysosome-like compartment⁵, suggesting that complete complexes are retained in these lineages because they are required for trafficking to canonical lysosome-related organelles as well. Additionally, *Chromera* possesses the chromerosome, which often displays intraluminal vesicles similar to MVBs, suggesting it may also be derived from endosome-like organelles (Obornik et al., 2011).

In conclusion, apicomplexans possess unusual endomembrane compartments including atypical Golgi and endosome-derived invasion organelles such as micronemes and rhoptries (Klinger et al., 2013b). Modifications in the complement of membrane trafficking machinery, including the loss of key protein complexes found in most eukaryotes, have been observed in the apicomplexan lineage, potentially associated with the specialization of the endomembrane system. The absence of some components (Exocyst, Vps39, Trs120, Tip20) within *Chromera* and *Vitrella* suggests pre-adaptation to parasitism deeper in the apicomplexan lineage. By contrast, the presence of near complete complements of key machinery (AP1-5, ESCRTs, COG) absent in many apicomplexans, pinpoints the timing of the losses at the colpodellid/apicomplexan transition.

Appendix 4

Apical complex and cytoskeleton.

Motility is an essential feature of many living organisms. Some organisms utilize microtubule-based specialized structures such as flagella and cilia for locomotion. Some use actin-based structures like filopodia, lamellipodia, and pseudopodia (**Frenal and Soldati-Favre, 2009**), which are exploited in the amoeboid crawling (**Pollard and Borisy, 2003**) or bacterial and viral movement into and between cells (**Stevens et al., 2006**). Apicomplexan parasites use an unconventional actin-based mode of locomotion known as gliding motility (**Morrisette and Sibley, 2002**). This mechanism allows the parasites to move fast in the absence of canonical microtubular and actin-based structures. Gliding motility is mediated by the apical complex, which is a cellular structure common to all apicomplexan parasites. In the apical complex, proteins secreted from specialized secretory organelles, microneme and rhoptries, mediate adhesion to the cell substrate during motility and invasion or formation of a PV (**Baum et al., 2006**).

Actin-based gliding motility is essential for apicomplexan invasion (**Skillman et al., 2011**). Apicomplexan gliding motility undergoes actin polymerization/depolymerization for their directional motility with other associated protein classes such as actin-like proteins (ALP), actin-related protein (ARP), capping protein (CP), formin, profilin and cofilin/ADF. Actins elongate in the form of filaments and push the membrane forward. Arp2/3 complex (one of the ARPs) mediates the initiation of new branches on pre-existing filaments. After some growth, CP terminates the elongation of the filaments. Cofilin/ADF promotes de-branching and depolymerization. Profilin mediates the refilling of ATP-actin monomer pools, which are used for elongation through catalyzing ADP-ATP exchange (**Baum et al., 2006; Foth et al., 2006**).

We identified and compared genes encoding actins and other related components in the 26 species according to a method described by a previous study (**Baum et al., 2006**). Chromerids share homology with Apicomplexa for most of the actin, ALP and ARP classes. For example, both chromerids possess actin 1 (ACT1), actin-related (ARP), and actin-like (ALP) homologs. There were fewer actin genes in apicomplexans than in chromerids, indicating losses during apicomplexan evolution. The patterns of losses were the same for closely related species, suggesting non-random, lineage-specific losses (**Figure 2—figure supplement 5C**).

Arp2/3 complex, a nucleator of actins (**Machesky et al., 1994**), consists of seven subunits that regulate actin polymerization (**Mullins et al., 1997; Fehrenbacher et al., 2003**). Initially identified in *Acanthamoeba* (**Machesky et al., 1994**), Arp2/3 complex is conserved in most eukaryotes (**Gordon and Sibley, 2005**). We could not identify genes encoding subunits of Arp2/3 complex in both chromerids (**Figure 2—figure supplement 5C**). Also, all subunits were not found in apicomplexan species, consistent with a previous study (**Gordon and Sibley, 2005**). Individual subunits are important, as subunit ARPC4/p20 was shown to be essential for a complete, functional Arp2/3 complex (**Gournier et al., 2001**). Different subunits were identified in different phyla (**Figure 2—figure supplement 5C**). Within Apicomplexa, ARPC1 and ARPC4 were present in *Cryptosporidium hominis* and *Cryptosporidium parvum*, and ARPC1 and ARPC2 in *Plasmodium* spp. *S. minutum*, a dinoflagellate, contains genes for most of the subunits except ARPC1. This suggests that the common ancestor of Myzozoa (chromerids, apicomplexans, and dinoflagellates) had all the subunits, and they were lost in different lineages. Genes encoding ALP1, hypothesized to function as Arp2/3-like nucleator (**Gordon and Sibley, 2005**), were found in apicomplexans and also in *Vitrella* (Vbra_266.t1). FH2-domain (Pfam-PF02181) containing formins are members of another actin nucleator gene family. They produce unbranched filaments unlike Arp2/3 complex, which induces branched filaments. Both chromerids possess formin1 (FRM1) and formin 2 (FRM2) homologs, which are conserved in all the other studied species as well. Although *Plasmodium* spp. maintained a 1-1 orthology for both FRM1 and FRM2, we found a coccidian-specific FRM3 (TGME49_213370), suggesting a lineage-specific expansion. Maintenance of some formins across chromerids and

apicomplexans and lack of Arp2/3 complex suggest their importance, perhaps reflecting a switch from Arp2/3 complex to formins for actin nucleation during the evolution of Apicomplexa. Taken together, it seems that an Arp2/3 independent actin nucleation mechanism had already evolved before Apicomplexa and chromerids, and losses of ARP have probably begun too, although inferring the exact timing and sequence of losses will require studying more closely related species such as *Colpodella*.

We analyzed coronins, a major conserved gene family with a multifunctional role in actin regulation and vesicular transport (**Rybakin and Clemen, 2005**). These are WD40-repeat containing proteins, which represent the only candidate for actin bundling in apicomplexan parasites. Coronins inhibit the nucleating activity of Arp2/3, unlike other known Arp2/3-binding proteins. We observed absence of coronins in both chromerids, which is consistent with the notion that they, functionally linked to the Arp2/3 complex, were lost (**Figure 2—figure supplement 5C**). Although parasite homologs do not have the microtubule-binding domain of canonical coronins, but essential amino acid residues are conserved (**Gandhi et al., 2010**). Thus, coronin could be playing a role in stabilizing F-actin scaffolds or having an alternative role in vesicular transport in apicomplexan parasites.

Profilins are actin-binding proteins that supply pools of readily polymerizable actin monomers (**Baum et al., 2006**). Genes encoding profilins were found in all 26 species studied except for an oomycete (*P. ultimum*) and diatoms (*P. tricornutum*, *Thalassiosira pseudonana*). Apicomplexa-specific profilins have β mini1 and β mini2 domains, which provide an extended interface with actin and formed the structural basis of their actin-binding function in *Toxoplasma* (**Kucera et al., 2010**) and *Plasmodium* (**Kursula et al., 2008**). These domains are not found in other eukaryotes. Sequence alignment of these profilins reveals an intriguing observation that *Vitrella* (Vbra_7301.t1) had these β -domains previously thought to be specific to Apicomplexa, with partial conservation in *Chromera* (Cvel_18957.t1) and in dinoflagellate *P. marinus* (XP_002774080). The β -domains were not detected in other non-apicomplexan species. All species studied has had only one profilin gene except for chromerids where we observed 2 in *Chromera* and 3 in *Vitrella*, including an one-to-one ortholog of the apicomplexan profilin in both chromerids.

Cyclase-associated proteins (CAPs) are evolutionary conserved G-actin-binding proteins, which participate in filament turnover regulation by acting on actin monomers (**Chaudhry et al., 2010**). CAP proteins are made up of three significant regions: N-terminal adenylate cyclase binding domain (CAP_N, linked to the cAMP-RAS signaling), a central proline-rich segment, and a C-terminal actin-binding domain (CAP_C). Apicomplexans do not possess the N-terminal (CAP_N) domain altogether with few genes in stramenopiles and in *Vitrella* (Vbra_7026.t1) also showing a similar pattern of loss. However, the *Chromera* gene Cvel_8488.t1 possesses both domains. This suggests the dispensable nature of CAP_N domain (**Figure 2—figure supplement 5C**), and we speculate that in parasites CAP functions are reduced to actin sequestration only.

The F-actin capping, CapZ duplex, a dimer of α - and β -CPs, prevents polymerization from the 'barbed' (plus) end. It is conserved across Apicomplexa except for in piroplasmids. It is also conserved in most of the species studied including stramenopiles, dinoflagellates, and both chromerids. In Apicomplexa, several gelsolin domain-containing proteins were found but they are unlikely to be functionally related and are speculated to be Sec23/Sec24-like proteins, which function in vesicular transport (**Baum et al., 2006**).

Cofilin/ADF genes promote de-branching of actin filaments and are well conserved among species studied. However, plasmodia differ from the rest of the Apicomplexa species in having an additional copy of the ADF gene. Phylogenetic analysis shows that ADF in plasmodia has duplicated and diverged with respect to the rest of the Apicomplexa, and recent structural studies explain the mechanism of action of *Plasmodium* ADFs (**Singh et al., 2011; Wong et al., 2014**). This represents a clear example of additional innovations of actin regulation in certain apicomplexan clades.

In addition, we identified myosin families in the 26 species using a myosin HMM model (Foth et al., 2006) (Figure 2—figure supplement 5C). Members of piroplasmids such as *Theileria annulata* and *Theileria parva* have the fewest genes among the apicomplexan species examined, likely because piroplasms do not require motility for intracellular invasion. On the other hand, we detected the most complete myosin family repertoire in *Chromera* and *Vitrella*. We detected certain myosin families in some apicomplexan genera, but not among non-apicomplexan species, indicating lineage-specific gains (data not shown). In summary, combinations of lineage-specific losses and gains have led to streamlined, unique repertoires of actins and myosins in various apicomplexan species.

Appendix 5

Extracellular proteins in chromerids.

We curated the chromerid genomes for genes with extracellular domains and domain architectures like similar to those of apicomplexans (**Figure 3—figure supplement 4; Supplementary file 5**). Both chromerids possess mucin-like proteins having long stretches of threonine and serine residues with predicted O-linked glycosylation, as well as the enzyme pathways involved in O-linked glycosylation (**Templeton et al., 2004a; Anantharaman et al., 2007**). Proteins encoding combinations of von Willibrand factor A (vWA) and thrombospondin 1 (TSP1) were also observed, although none with apparent orthologous relationship to the vWA and TSP1 domain proteins (TRAP) that serve as receptors mediating gliding motility in apicomplexans. The chromerid genomes possess numerous secreted proteins with domains predicted to participate in binding of sugar moieties (**Figure 3—figure supplement 4**). Chromerids share FRINGE domains with *Cryptosporidium*, and HINT domains with *Cryptosporidium* and *Gregarina*, to the exclusion of other apicomplexans, in support of early divergence of these genera within the Apicomplexa (**Figure 3—figure supplement 4B**). *Vitrella* genome contains multiple copies of proteins, which have arrays of the cysteine-rich oocyst wall protein (OWP) domain found in *Cryptosporidium* and coccidians, which are associated with forming environmentally durable walls of oocysts (**Templeton et al., 2004b**).

Several EC domain architectures thought to be distributed apicomplexan-wide have homologs in the chromerids; for example, the LCCL domain-containing proteins, CCp1 and CCp2/3, as well as the CPW-WPC domain proteins (**Figure 3—figure supplement 4C**). Ultrastructures reminiscent of micronemes have been observed in both chromerids (**Obornik et al., 2012**); consistent with this, we identified EC proteins having domains and architectures typical of *Toxoplasma* and *Plasmodium* micronemal secretory proteins. Examples include expansions of proteins containing SUSHI, EGF, TSP1, and vWA domains (data not shown). Chromerids possess unique architectures of proteins containing the macrophage perforin (MacPerf) domain (**Figure 3—figure supplement 4E**), which, previously found in apicomplexans and ciliates (as large expansions), are thought to function in apicomplexans to mediate membrane lysis during host cell egress and tissue traversal (**Roiko and Carruthers, 2009**). The *Chromera* MacPerf domain proteins also contain arrays of a domain, WSC, thus far not found in other alveolates, as well as a unique C-terminal DERM domain. *Chromera* possesses four MacPerf domain proteins with various domain architectures, whereas *Vitrella* a single MacPerf protein with a stand alone MacPerf domain (Vbra_18070.t1).

The ciliate genomes possess highly amplified and antigenically diverse repertoires of GPI-linked proteins termed 'immobilization antigens' (**Caron and Meyer, 1989**). We did not see amplifications of GPI-linked gene families in either chromerid species. Lineage-specific gene amplifications include a predicted secreted protein in *Vitrella*, which contains an arenylsulfonase domain (**Figure 3—figure supplement 4E**). The chromerids possess highly amplified gene family, annotated as 'CAST multi-domain protein' in the ciliate, *Oxytricha* (e.g., UniProt ID: OXYTRI_15408), and which comprises conserved cysteine-rich domains in the extracellular region, a single transmembrane domain, and a conserved predicted coiled-coil region in the cytoplasmic domain (e.g., Cvel_3066.t1). Representatives of this protein are found in the ciliate *Oxytricha*, but not in *Tetrahymena* and *Paramecium*; in stramenopiles, choanoflagellates and coccidians, but are absent from other apicomplexans such as *Cryptosporidium*, *Theileria*, *Babesia*, and *Plasmodium*.

Appendix 6

ApiAP2 proteins.

We examined the abundance of apicomplexan-specific AP2 (apiAP2) genes, transcription factors that play regulatory roles in key aspects of apicomplexan biology (**Campbell et al., 2010; Flueck et al., 2010; Radke et al., 2013; Kafsack et al., 2014; Sinha et al., 2014**). We scanned the protein-coding gene sets of the 26 species using the apicomplexan-specific apiAP2 HMM, which was constructed with the AP2 domain sequences from apicomplexan species. We found that apiAP2 genes were abundant in both chromerids and in all apicomplexans. ApiAP2 genes were moderately abundant in the two dinoflagellates and rare or absent in ciliates and stramenopiles, respectively (**Figure 3—figure supplement 1D**). There were very few apiAP2 genes that were shared between apicomplexans and non-apicomplexan species; most were shared between closely related species, that is, from the same clade (**Figure 3—figure supplement 1B**). These lineage-specific apiAP2 genes in the present-day species could have arisen from de novo gene birth or modification of the full-length sequences of existing genes beyond recognition. In the former case, the proto-apicomplexan ancestor had a small set of apiAP2 genes. In the latter case, the common ancestor already had a large set of apiAP2 genes, which continued to change, giving the appearance of ‘new’ clade-specific genes. The latter case, the turnover scenario, is more parsimonious because, according to the de novo gene birth scenario, apiAP2 genes must have expanded independently in every descending lineage from the proto-apicomplexan ancestor. In summary, our data support the notion that massive apiAP2 expansion occurred in the common ancestor before Apicomplexa and chromerids split, and the apiAP2s continued to change as the common ancestor split into chromerids and apicomplexans, which continued to radiate and adapt to their host niches and life cycle strategies.

We sought to determine if gene duplication and divergence was a significant mechanism for the expansion and the turnover of apiAP2 genes. The number of apiAP2 genes that have other homologous apiAP2 genes within the species based on OrthoMCL clustering, which are likely mediated by paralogous expansions, was high (93 out of 136) in chromerids (**Figure 3—figure supplement 1D**). In *Vitrella*, we identified one cluster of 50 homologous apiAP2 genes. This means that gene duplication played an important role in expanding apiAP2 gene repertoire in chromerids. The number was significantly less (5 out of 13) in dinoflagellates than in chromerids (**Figure 3—figure supplement 1D**). We suspect that gene duplication and diversifications drove expansion of apiAP2 genes significantly after the split of dinoflagellates. In apicomplexan species, evidence for recent duplications was sparse, as only 4 out of 409 apiAP2 genes had homologous copies in the same species. This does not necessarily mean that apiAP2 genes do not duplicate readily in apicomplexans, but rather that redundant copies of apiAP2 are quickly lost or diversified beyond recognition in part by selective pressure to reduce gene repertoires and genome sizes (**Katinka et al., 2001**) and due to higher rate of sequence divergence in parasites (**Hafner et al., 1994**).

Previous studies have shown that plant genomes contain a large repertoire of AP2 genes, and that plant AP2 domains evolved from an endonuclease domain in a cyanobacteria (**Magnani et al., 2004**). According to our phylogenetic analysis, AP domains among bacteria are many and diverse, with both plant-like and apicomplexan-like AP2s (data not shown). We did not find significant homology with bacterial AP2 genes at the full gene length level. It is not clear if the originally transferred AP2 gene has evolved beyond recognition or if only the domain has been transferred to these eukaryotes. The exact genetic events that led to acquisition of AP2s in apicomplexans are not clear. However, what is the most probable scenario is that AP2 domains in alveolates came from bacteria and have expanded in myzozoans, independent of those in plants. Both functional studies and more taxon sampling would be required for elucidating how AP2s in alveolates were acquired in the first place.

Paper III

**RE-EVALUATING THE GREEN VERSUS RED SIGNAL IN EUKARYOTES WITH
SECONDARY PLASTID OF RED ALGAL ORIGIN**

Burki F, Flegontov P, Oborník M, Cihlář J, Pain A, Lukes J, Keeling PJ.

Genome Biology and Evolution 4(6):626-35 (2012)

Re-evaluating the Green versus Red Signal in Eukaryotes with Secondary Plastid of Red Algal Origin

Fabien Burki^{1,†}, Pavel Flegontov^{2,†}, Miroslav Oborník^{2,3,4}, Jaromír Cihlár², Arnab Pain⁵, Julius Lukeš^{2,3}, and Patrick J. Keeling^{1,*}

¹Canadian Institute for Advanced Research, Department of Botany, University of British Columbia, Vancouver, Canada

²Biology Centre, Institute of Parasitology, Czech Academy of Sciences, České Budějovice, Czech Republic

³Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic

⁴Institute of Microbiology, Czech Academy of Sciences, Třeboň, Czech Republic

⁵Computational Bioscience Research Center (CBRC), Chemical Life Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

[†]These authors contributed equally to this work.

*Corresponding author: E-mail: pkeeling@mail.ubc.ca.

Accepted: May 9, 2012

Abstract

The transition from endosymbiont to organelle in eukaryotic cells involves the transfer of significant numbers of genes to the host genomes, a process known as endosymbiotic gene transfer (EGT). In the case of plastid organelles, EGTs have been shown to leave a footprint in the nuclear genome that can be indicative of ancient photosynthetic activity in present-day plastid-lacking organisms, or even hint at the existence of cryptic plastids. Here, we evaluated the impact of EGT on eukaryote genomes by reanalyzing the recently published EST dataset for *Chromera velia*, an interesting test case of a photosynthetic alga closely related to apicomplexan parasites. Previously, 513 genes were reported to originate from red and green algae in a 1:1 ratio. In contrast, by manually inspecting newly generated trees indicating putative algal ancestry, we recovered only 51 genes congruent with EGT, of which 23 and 9 were of red and green algal origin, respectively, whereas 19 were ambiguous regarding the algal provenance. Our approach also uncovered 109 genes that branched within a monocot angiosperm clade, most likely representing a contamination. We emphasize the lack of congruence and the subjectivity resulting from independent phylogenomic screens for EGT, which appear to call for extreme caution when drawing conclusions for major evolutionary events.

Key words: Endosymbiotic gene transfer, plastid evolution, protist, algae, chromera.

The photosynthetic organelles of plants and algae (plastids) are the product of endosymbioses, where once free-living organisms were engulfed and retained by eukaryotic host cells (Reyes-Prieto et al. 2007; Gould et al. 2008). Initially, primary endosymbiosis involved the integration of a photosynthetic prokaryote related to modern-day cyanobacteria, most likely in the common ancestor of glaucophytes, red algae, and green plants (green algae and land plants), resulting in the Plantae supergroup (Palmer et al. 2004). Subsequently, primary plastids spread to other eukaryotes by means of secondary endosymbioses, where a green or red alga was taken up by another lineage, and the process was repeated yet again as tertiary endosymbioses in some dinoflagellates (Keeling 2010).

Plastid genomes rarely encode more than 200 proteins, which represent a small fraction of the proteins required for full functionality, and an even smaller fraction of the few thousand proteins found in modern-day cyanobacteria (Martin et al. 1998). It is widely assumed that most endosymbiont genes were either lost or transferred to the host nucleus during the course of plastid integration (Lane and Archibald 2008). This migration of genes between two genomes is known as endosymbiotic gene transfer (EGT), a special case of horizontal gene transfer (HGT). The products of the transferred genes that are essential for plastid function are targeted back across the plastid membranes to reside in their original compartment, a process that played a fundamental role in the integration of endosymbiont and host (Patron and Waller

© The Author(s) 2012. Published by Oxford University Press on behalf of the *Society for Molecular Biology and Evolution*.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

2007). However, not all nucleus-encoded genes inferred to be of endosymbiotic origin are plastid targeted; in the land plant *Arabidopsis thaliana*, for example, >50% of identified EGTs have evolved functions unrelated to the plastid (Martin et al. 2002).

The impact of EGTs on the host nuclear genome is generally considered to be significant. Estimates for cyanobacterial genes in the nucleus range from 6% in the green alga *Chlamydomonas reinhardtii* (Moustafa and Bhattacharya 2008), to about 11% in the glaucophyte *Cyanophora paradoxa* (Reyes-Prieto et al. 2006), and to as high as 18% in *A. thaliana* (Martin et al. 2002). Secondary endosymbioses complicate the prediction of EGTs because not only the host nucleus potentially integrated genes from the secondary plastid, but also from the nucleus of the green or red algal endosymbiont, itself the recipient of cyanobacterial genes previously transferred from the primary plastid (Archibald 2009). Nevertheless, genome-scale analyses have begun to analyze the extent of EGTs in taxa with plastids of secondary origin, with complex and sometimes contradictory results. Diatoms possess a red algal plastid, and in *Phaeodactylum tricorutum* 171 genes (1.6% of the gene catalog) were predicted to be of red algal origin (Bowler et al. 2008). A much less anticipated result came from another analysis of diatoms, which suggested that over 1700 genes, representing 16% of the nuclear genes, were derived from green algae, compared with only about 400 genes with red algal affinity (Moustafa et al. 2009). A green phylogenetic signal of such magnitude led Moustafa et al. (2009) to build on other similar findings of fewer genes (Becker et al. 2008; Frommolt et al. 2008) and propose that these genes are in fact evidence of an ancient, cryptic green algal endosymbiont predating the acquisition of the red algal plastid that we observe today.

A similar approach was employed to study the phylogenetic origins of *Chromera velia* expressed nuclear genes (Woehle et al. 2011). *Chromera velia* has attracted much attention because it is a photosynthetic relative of apicomplexan parasites, whose highly reduced, non-photosynthetic plastid has been a puzzling evolutionary issue (Moore et al. 2008; Janouskovec et al. 2010; Obornik et al. 2011). Woehle et al. (2011) produced 29,856 contigs from a 454 Titanium GS FLX (Roche) cDNA sequencing, of which they drastically reduced the redundancy to 3,151 clusters. As expected for an alga with a red algal-derived plastid, 263 genes were found to indicate a red photosynthetic ancestry, but they also found a prominent signal of 250 genes apparently reflecting a green ancestry (Woehle et al. 2011). In this case, however, the authors cautiously attributed this signal to limited sampling of red algal genomes and phylogenetic artifacts rather than to a green endosymbiont, as in the diatom analysis (Moustafa et al. 2009).

In a Blast-based survey of *C. velia* clusters, we found indication of contamination from land plants (specifically from monocots). This prompted us to re-evaluate the ratio of

putatively red and green genes in *C. velia* using a slightly different phylogenomic protocol (see Materials and Methods), which allowed us to investigate how methodological variations can affect the phylogenomic profiles of the same dataset. To identify putative red or green algal genes in *C. velia*, we first generated maximum likelihood phylogenetic trees for 2,146 genes and automatically searched for topologies consistent with EGT. This procedure identified 362 genes showing exclusive affinity between secondary plastid-bearing lineages (including *C. velia*) and red algae, viridiplantae (green algae and/or land plants), or glaucophytes (bootstrap support $\geq 80\%$). This represented our initial pool of candidate genes for EGT. As controls, we also evaluated the signal uniting *C. velia* with alveolates (apicomplexans, dinoflagellates, and/or ciliates), which are closely related to *C. velia* and therefore expected to be the dominant signal. We found *C. velia* united with alveolates in 448 trees. Lastly, we scanned our set of trees for monophyletic grouping between *C. velia* and prokaryotes, and identified 53 cases as possible evidence of HGT.

At face value, these figures might be taken to suggest a large contribution of EGT to the *C. velia* genome. However, automated computational pipelines used for searching HGT/EGT in genomic data can be misleading and detailed curation of the resulting phylogenies is absolutely necessary to avoid false positives. In the case of hypothetical EGT from red or putative cryptic green endosymbionts, the expected relationships are known: the transferred genes should be most closely related to either red or green algae (ideally nested within either group if a diverse sample of algal sequences is available) to the exclusion of all other eukaryotic or prokaryotic groups. If the genes were ancestrally derived from the cyanobacterial progenitor that gave rise to the primary plastids in red and green algae, a cyanobacterial outgroup should also be recovered. Realistically, it cannot be expected that such theoretical topologies will be inferred or will be robustly supported for every real case of EGT, even with the help of complex evolutionary models. Indeed, the considerable evolutionary distances, inappropriate taxon sampling, lack of genuine phylogenetic signal, and various artifacts such as compositional biases, extreme rate variation among sites, or heterotachy will negatively impact the resolution of most trees (Philippe and Laurent 1998; Philippe et al. 2005; Lockhart et al. 2006; Jeffroy et al. 2006; Stiller 2011). Accordingly, the conditions for the detailed verification of the trees were slightly relaxed so that more than one algal type was allowed in the monophyly (see Material and Methods).

The above conditions were applied to the initial pool of 362 candidate algal genes to refine the assessment of putative EGT, resulting in a different picture than the automated sort. First, 109 genes (almost one-third of the genes identified as possibly "algal") showed strong similarity to land plants, with *C. velia* clearly belonging to a monocotyledon clade (Supplementary fig. S1 and table S1, Supplementary Material online). It cannot be ruled out that these represent

HGTs from land plant to *C. velia*, but the high level of sequence identity to homologs from monocotyledons (90 *C. velia* sequences displayed >90% identity, among which 22 showed 100% identity), favors the simpler explanation of a contamination in the *C. velia* dataset.

More interestingly, out of the remaining 253 candidate genes of algal origin, only 23 were found to support a red algal origin (fig. 1 and [supplementary fig. S2, Supplementary Material online](#); table 1) and 9 supported a green algal origin (fig. 2 and [supplementary fig. S3, Supplementary Material online](#); table 1). An outgroup and representatives of both green and red algae were required to be included in the tree, which are necessary conditions to distinguish between red and green signals. Other genes produced more ambiguous signals because *C. velia* fell within a clade of mixed algal types: in 11 trees red and green algae were mixed; in 3 trees red and glaucophyte algae were mixed; and in 5 trees red, glaucophyte, and green algae were mixed (fig. 3 and [supplementary fig. S4, Supplementary Material online](#); table 1). The coverage of *C. velia* in these putative algal genes ranged from 27% to 100% of the length of the trimmed alignments, but for the majority (65%) *C. velia* covered >90%, limiting possible phylogenetic artifacts associated with incomplete genes (table 1). Finally, 18 trees showing possible evidence of exclusive HGT from bacteria remained after manual curation ([supplementary table S2, Supplementary Material online](#)).

All in all, detailed inspection of automatically parsed trees recovered a mere 51 genes in this *C. velia* EST dataset possibly supporting transfers from an algal endosymbiont, although sampling is often so limited as to preclude any strong conclusions about the direction of the transfer. Interestingly, 47% (24/51) of these EGT candidates were also predicted to encode an N-terminal plastid targeting presequence (Woehle et al. 2011), providing an independent evidence of their link to the plastid (table 1). Other aspects of these trees are not so easily explained. For example, 12 genes inferred to be of red algal origin included chlorarachniophytes in the “red” clade, but these algae possess green secondary plastids (Rogers et al. 2007). Although compatible with the nested phylogenetic position of the chlorarachniophyte host among the red algal plastid-containing groups stramenopile, alveolate, and hatophytes (Burki et al. 2007; 2012), it implies additional HGT events either before or after the establishment of its green plastid (Archibald et al. 2003).

Most importantly, however, these analyses show that large-scale phylogenomic pipelines can result in drastic differences: from the same transcriptome data we identified 51 putative algal-derived genes, versus 513 identified by Woehle et al. (2011). But this is only part of the problem, because the overlap in genes identified by the two analyses was only eight genes, meaning that 43 (84%) of the genes that we identified were not recovered by Woehle et al. (2011), whereas 505 (98%) of the genes they identified did not meet our criteria (fig. 4). We see a number of explanations

for this discrepancy, some of which compound the effects of others. (1) The database used in Woehle et al. (2011) to populate the phylogenetic trees led to misleading results. Very limited sampling for land plants (only two representatives, *A. thaliana* and *Physcomitrella patens*) did not permit to recover the monocot signal in 109 genes, 10 of which were wrongly classified as contributing to the green signal in Woehle et al. (2011) (fig. 4; [supplementary table S1, Supplementary Material online](#)). The absence of prokaryotes was also problematic and precluded the identification of several instances of complicated phylogenetic patterns (including non-exclusive HGTs) rather than evidence of red and green signals. [Supplementary figure S5 \(Supplementary Material online\)](#) shows examples of such phylogenies impacted by the inclusion of prokaryotes that do not support an algal ancestry in *C. velia*, but were inferred to do so in Woehle et al. (2011). (2) The procedure to select the final taxa entering the phylogenetic reconstruction step in Woehle et al. (2011) interfered with the interpretation of the resulting trees. Specifically, all taxa except *C. velia*, red and green alga, and an outgroup were removed from clusters of homologous sequences prior to the phylogenetic reconstructions, which likely exacerbated the problem outlined above. (3) No statistical support was used to evaluate the robustness of the trees, resulting in many trees showing only weak affinity to red or green algae yet classified as contributing to the overall photosynthetic signal.

The case of *C. velia* is not unique: a number of recent studies have described contrasting reinterpretations of the same datasets. For example, the imposing 1,700 genes inferred to be of green algal origin in diatoms (Moustafa et al. 2009) was reduced to only 144 genes after more stringent criteria were applied, notably the mandatory presence of red algal sequences in the trees (Dorrell and Smith 2011). These differences are important, because the presence of EGTs is not only used to infer the contribution of extant endosymbiotic organelles to their host, but have also been used as evidence for photosynthetic ancestry in plastid-lacking lineages, or even the presence of cryptic plastids. Oomycetes and ciliates are two heterotrophic groups sharing undisputable common ancestry with red algal plastid-containing lineages. In the case of oomycetes, the complete genomes of two *Phytophthora* species revealed the existence of 855 genes with putative red algal or cyanobacterial origins that were presented as evidence for the ancient presence of a red algal plastid (Tyler 2006). However, a reanalysis of this dataset, specifically testing for EGTs, showed no such evidence for red algal contributions to the oomycete genome (Stiller et al. 2009). Similarly, based on the identification of 16 genes of apparent algal origin in the genomes of *Tetrahymena thermophila* and *Paramecium tetraurelia*, ciliates were proposed to have once been photosynthetic (Reyes-Prieto et al. 2008), despite a previous assessment that *T. thermophila* displayed no signal of plastid descent above the expected background noise (Eisen et al. 2006).

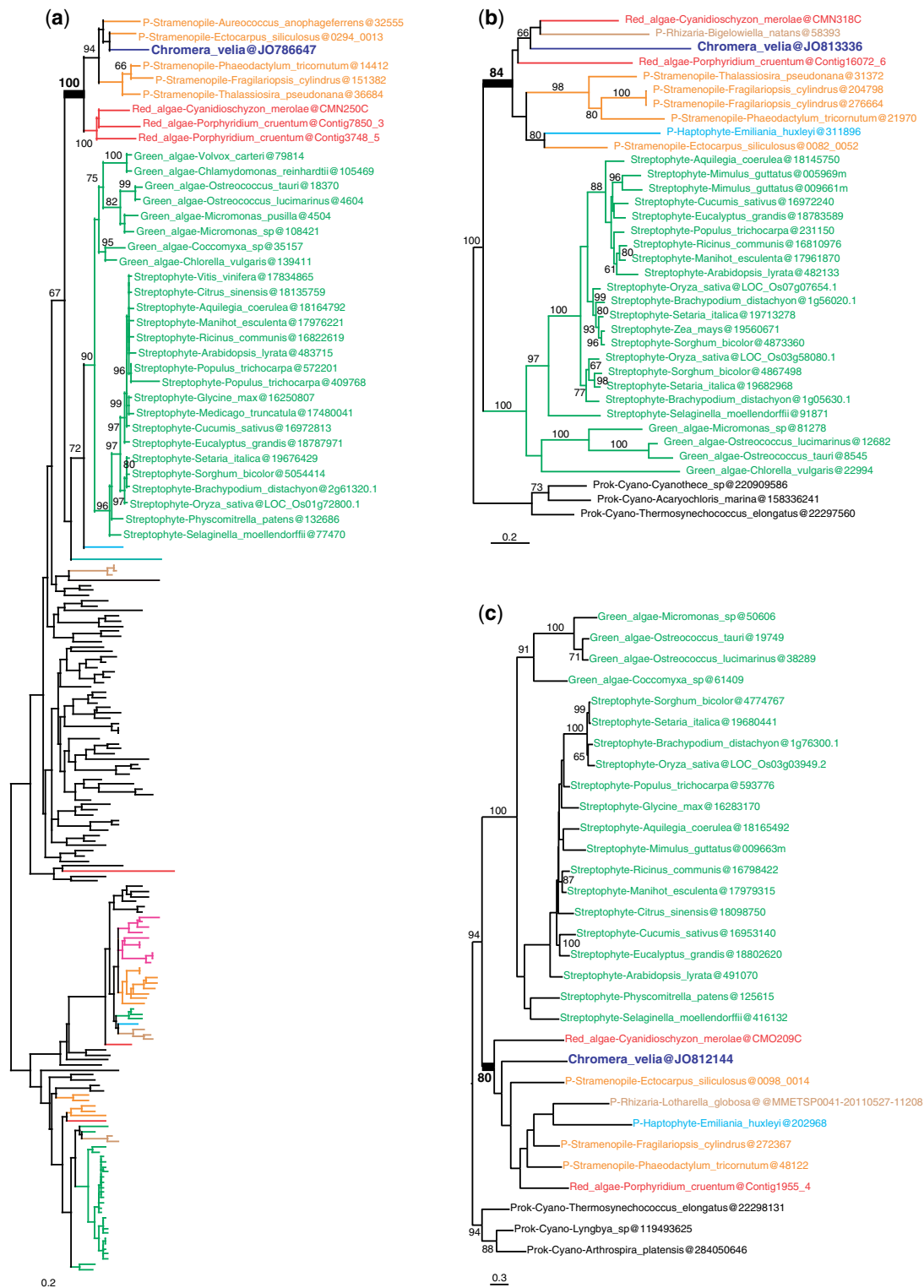


FIG. 1.—Examples of maximum likelihood trees congruent with EGT from a red algal endosymbiont. (a) Signal recognition particle-docking protein. (b) Folate biopterin transporter. (c) Vitamin k epoxide reductase. Numbers at nodes represent bootstrap proportion; only values higher than 60% are shown. For clarity, only the relevant taxa are shown (complete taxon list is available in [Supplementary Material online](#)); branches and taxa are colored according to their taxonomy: dark blue: *C. velia*; red: red algae; green: viridiplantae; orange: stramenopiles; light blue: haptophytes, cryptophytes; brown: Rhizaria; pink: alveolates; black: prokaryotes, animals, fungi, Amoebozoa. All trees congruent with EGT from a red algal endosymbiont are found in [supplementary figure S2](#) ([Supplementary Material online](#)).

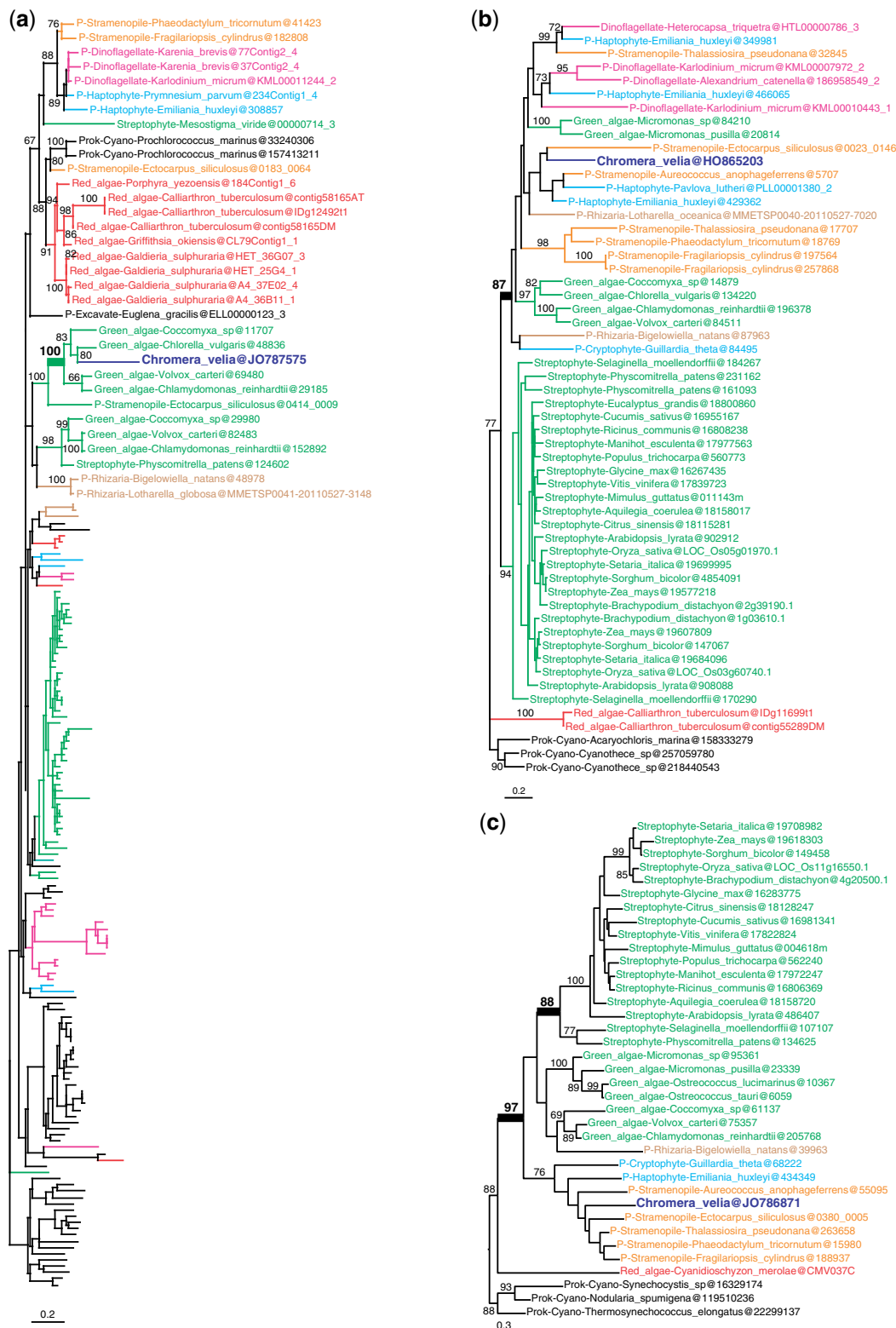


FIG. 2.—Examples of maximum likelihood trees congruent with EGT from a green algal endosymbiont. (a) Fructose-bisphosphate aldolase c. (b) No function prediction. (c) Gun4 domain protein. Numbers at nodes represent bootstrap proportion; only values higher than 60% are shown. For clarity, only the relevant taxa are shown (complete taxon list is available in [Supplementary Material online](#)); branches and taxa are colored according to their taxonomy:

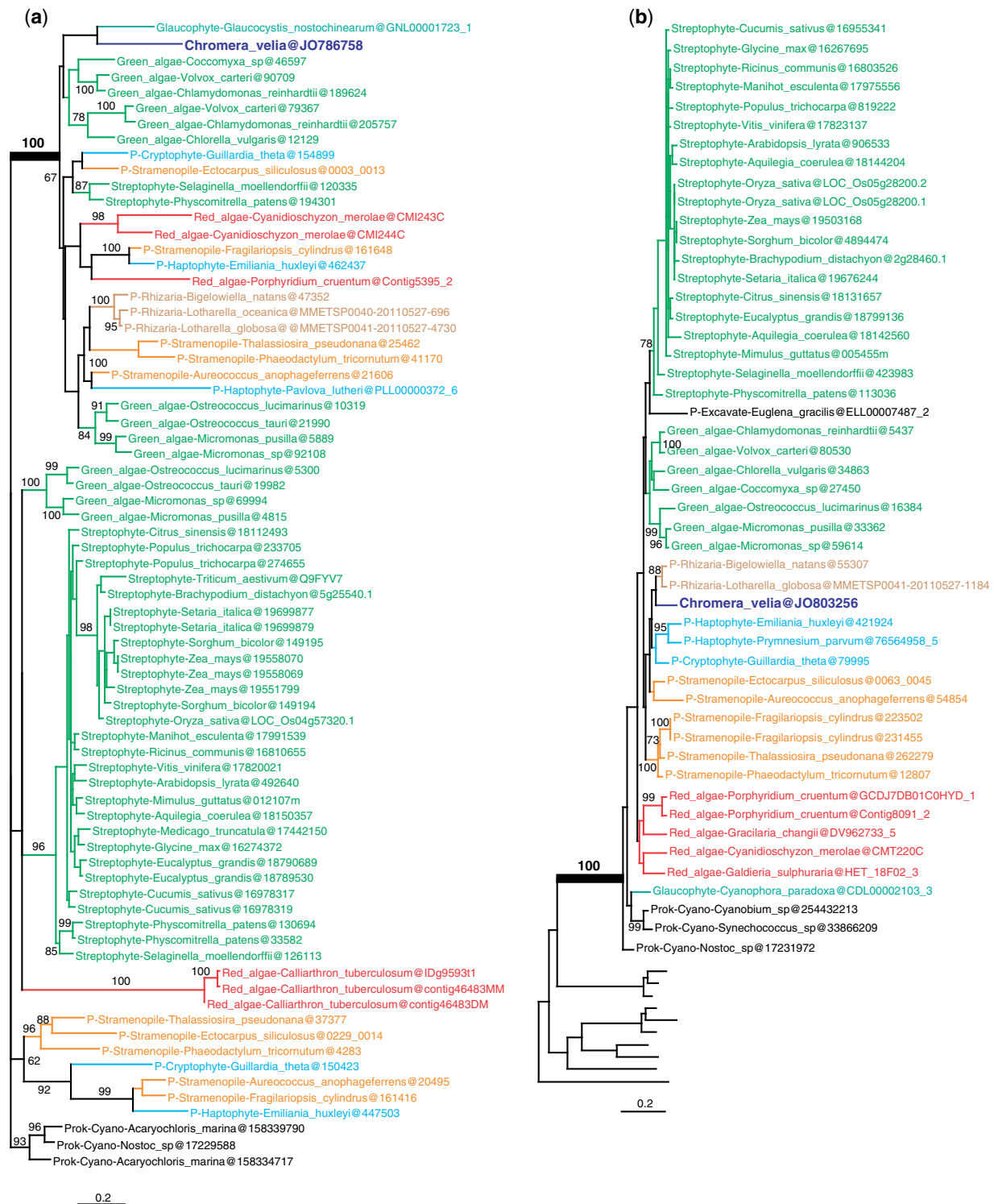


FIG. 3.—Examples of maximum likelihood trees congruent with EGT from an algal endosymbiont, but the algal type could not be determined. (a) Plastid terminal oxidase. (b) Chlorophyll synthetase. Numbers at nodes represent bootstrap proportion; only values higher than 60% are shown. For clarity, only the relevant taxa are shown (complete taxon list is available in [Supplementary Material online](#)); branches and taxa are colored according to their taxonomy: dark blue: *C. velia*; red: red algae; green: viridiplantae; orange: stramenopiles; light blue: haptophytes, cryptophytes; brown: Rhizaria; turquoise green: glaucophytes; black: prokaryotes, animals, fungi, Amoebozoa. All trees congruent with an algal origin are found in [supplementary figure S4](#) ([Supplementary Material online](#)).

Table 1

Genes with tree topologies concordant with an algal origin

Seq. ID	Seq. Function	E-value	Algal Origin ^a	Plastid Targeted ^b	Coverage ^c
JO786647	Signal recognition particle-docking protein	7.43E-80	R	Yes	0.99
JO786663	NA	1.15E-99	R	Yes	0.99
JO786667	Ferredoxin (2fe-2s)	1.94E-40	R	Yes	1
JO786670	ATP-dependent clp proteolytic subunit	4.89E-90	R	Yes	0.98
JO786681	<i>ATP-dependent clp protease proteolytic subunit</i>	1.12E-72	R	Yes	1
JO786748 ^d	<i>ATP-dependent clp protease proteolytic subunit</i>	5.54E-112	R	Yes	1
JO786683	Integral membrane protein	3.24E-77	R	Yes	1
JO786729	NA	6.57E-37	R	Yes	0.98
JO786744	Fructosamine kinase	2.96E-87	R	Yes	0.98
JO786766 ^d	Tyrosyl-tRNA synthetase	9.57E-47	R	Yes	1
JO786779	NA	9.34E-102	R	Yes	1
JO789192	Glycerol-3-phosphate dehydrogenase	9.79E-118	R	No	1
JO790726	Adenosine trna methylthiotransferase	2.22E-70	R	No	0.53
JO792696	<i>Nad-dependent epimerase dehydratase</i>	2.91E-51	R	No	0.99
JO803234	<i>Nad-dependent epimerase dehydratase</i>	2.74E-140	R/GI	No	1
JO794159	Oxygen-evolving enhancer protein	5.78E-51	R	No	0.47
JO795745	Aspartyl glutamyl-trna amidotransferase subunit b	1.05E-37	R	No	0.5
JO800417	Peptide chain release factor 3	0	R	No	1
JO805350 ^d	Peptide chain release factor 1	1.92E-130	R	No	0.96
JO807105 ^d	Electroneutral sodium bicarbonate exchanger 1	7.37E-50	R	No	0.34
JO807782	<i>Aldo keto reductase</i>	9.82E-48	R	No	0.97
JO799950	<i>Aldo keto reductase</i>	4.99E-75	R/G	No	0.87
JO812144	Vitamin k epoxide reductase	1.79E-46	R	No	1
JO813336	Folate biopterin transporter	2.09E-27	R	No	0.27
JO813530	Magnesium chelatase atpase subunit d	1.90E-127	R	No	0.41
JO814400	Zinc-binding dehydrogenase	3.01E-46	R	No	0.98
HO865203	NA	6.57E-49	G	Yes	0.73
JO786726 ^d	Coproporphyrinogen iii oxidase	0	G	Yes	0.99
JO786781	NA	1.12E-120	G	Yes	0.97
JO786871	Gun4 domain protein	3.26E-59	G	Yes	0.99
JO787575 ^d	Fructose-bisphosphate aldolase c	1.91E-75	G	No	0.76
JO794110	Light-dependent protochlorophyllide oxido-reductase	7.19E-41	G	No	0.93
JO798116	Vacuolar atp synthase 16 kda proteolipid subunit	8.48E-31	G	No	0.51
JO803246	Glucose-methanol-choline oxidoreductase	1.04E-152	G	No	0.99
JO812733 ^d	NA	1.91E-91	G	No	0.54
HO865098	Flavodoxin	1.11E-38	R/G	Yes	0.99
JO786648	<i>Uroporphyrinogen decarboxylase</i>	0	R/G	Yes	1
JO786655	<i>Uroporphyrinogen decarboxylase</i>	0	R/G	Yes	1
JO786721	Permeases of the major facilitator superfamily	3.41E-44	R/G	Yes	0.96
JO786743	NA	3.42E-60	R/G	Yes	0.95
JO786758	Plastid terminal oxidase	4.12E-87	R/G	Yes	0.93
JO786778	Zeta-carotene desaturase	5.86E-171	R/G	Yes	0.73
JO786874 ^d	Tryptophanyl-tRNA synthetase	1.72E-71	R/G	Yes	0.58
JO793833	Fe-s metabolism associated	1.92E-40	R/G	No	0.88
JO802386	Amine oxidase	1.48E-93	R/G	No	0.47
JO803256	Chlorophyll synthetase	8.61E-160	R/G	No	1
JO806278	Leucyl aminopeptidase	4.86E-59	R/G	No	0.41
JO806648	Phosphoserine aminotransferase	8.53E-92	R/G	No	0.98
JO807737	NA	6.40E-58	R/G	No	0.99
JO814175	<i>Methyltransferase type 11</i>	2.48E-59	R/G	No	0.65
JO786792	<i>Methyltransferase type 11</i>	1.20E-102	R/G/GI	Yes	1

NOTE.—Italic characters denote ancient paralogs, that is, duplication occurred in the algal donor, and both copies were possibly acquired via EGT.

^aPossible origins in *C. velia*. R: Red algae; G: Green algae; R/G: Red and/or Green algae; R/GI: Red and/or Glaucophyte algae; R/G/GI: Red and/or Green and/or Glaucophyte algae.

^bAs inferred in Woehle et al. (2011).

^cCoverage is defined here as the length of the *C.velia* gene fragment divided by the total length of the alignment after masking of the poorly aligned sites (trimal).

^dAlso recovered in Woehle et al. (2011).

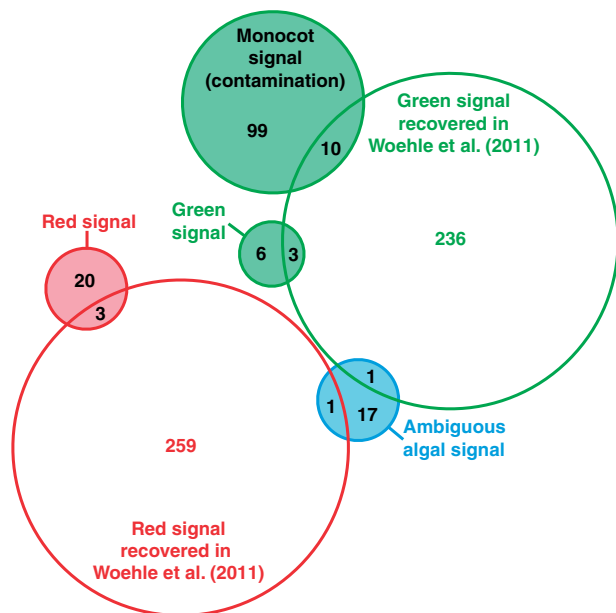


FIG. 4.—Venn diagram showing the number of overlapping genes between this study and Woehle et al. (2011). The filled circles correspond to the genes recovered in this study.

These discrepancies aside, all algae do contain some genes of endosymbiotic origin, raising a fundamental question: How many cases of EGT are enough to be considered evidence for past presence of endosymbionts? There is no clear answer because each lineage is different. For example, a mere seven genes of cyanobacterial or algal origin were identified in the apicomplexan parasite *Cryptosporidium parvum* (Huang et al. 2004), which lacks a plastid (Abrahamsen 2004). But because *Cryptosporidium*'s close relatives all possess plastids, these few genes were interpreted as supporting the view that *Cryptosporidium* evolved from a plastid-containing lineage (Huang et al. 2004). In contrast, over 100 genes of possible algal origin were inferred in the unicellular choanoflagellate *Monosiga* (Sun et al. 2010), but because there is no other evidence to suggest that choanoflagellates ever had a plastid, these genes were interpreted as HGT, reflecting feeding behaviors rather than plastid losses (Sun et al. 2010).

Another example is the chlamydial footprint found in Plantae; two studies reported that at least 21 and 55 genes, respectively, were transferred between chlamydiae and the ancestor of primary photosynthetic eukaryotes, the majority of which are putatively plastid targeted and as such were proposed to have contributed to the establishment of the cyanobacterial endosymbiont (Huang and Gogarten 2007; Moustafa et al. 2008). But because there is no unambiguous rule to distinguish between HGTs from related sources and EGTs, Huang and Gogarten (2007) interpreted these genes as evidence for an ancient chlamydial endosymbiont, whereas

Moustafa et al. (2008) raised the possibility that mixotrophy and multiple HGTs may have instead played an important role. Overall, independent phylogenomic analyses are not only leading to different results, but often reach different conclusions based on similar results.

These contrasting cases are symptomatic of the current situation and attest that the interpretation of unexpected phylogenetic patterns is often subjective and influenced by *a priori* expectation. They call for a better use of experimental controls and explicit testing of predictions of HGTs/EGTs to distinguish between genuine signal and noise (Stiller 2011). The task of analyzing thousands of trees that genome data have made possible is complex and improved methods need to be developed to help identifying the trees that strongly support the HGTs/EGTs scenarios under investigation. Increasing availability of genomic data for key taxa will permit us to specifically test these scenarios and examine alternative explanations for phylogenetic signal deviating from vertical inheritance.

Materials and Methods

A workflow diagram describing the procedure of sequence retrieval, alignment, tree reconstruction, and sorting can be found in [supplementary figure S6 \(Supplementary Material online\)](#). *Chromera velia* 3,151 clusters from Woehle et al. (2011) were used as query in a BLASTP search against protein sets from complete genomes and EST datasets (see [supplementary table S3, Supplementary Material online](#) for the complete list of taxa included in the analysis). CDHIT (Li and Godzik 2006) was used to reduce redundancy within each protein dataset prior to Blast in order to facilitate the subsequent tree interpretation by removing recent paralogs (clustering threshold: 90% identity). The Blast output was then parsed with a stringent *e*-value threshold of 1e-20 to minimize the inclusion of paralogs and hits were collected for each *C. velia* protein and multiple fasta files created. To prevent the inclusion of several closely related prokaryotic species, only the three best hits in each prokaryotic group were included ([supplementary table S3, Supplementary Material online](#)). MAFFT-LINSI (Katoh et al. 2005) was used for aligning sequences and TRIMAL (Capella-Gutiérrez et al. 2009) for selecting aligned positions, with sites containing more than 10% of gaps removed. Multiple sequence alignment files with less than five species were discarded at this stage. RAXML 7.2.8 (Stamatakis 2006) was used to reconstruct trees, with the LG substitution matrix + Γ 4 + F evolutionary model and 100 bootstrap replicates.

This approach resulted in 2,143 trees containing at least five species (including *C. velia*). The pre-sorting of these trees was first done automatically with a text-parsing Perl script used in Chan, Reyes-Prieto, et al. (2011) and Chan et al. (2011), with the initial condition that *C. velia* be monophyletic with members of plants (red algae, green algae, streptophytes, and/or glaucophytes) and/or members of secondary

plastid-bearing lineages of alveolates, stramenopiles, Rhizaria, haptophytes and cryptophytes, and/or Cyanobacteria (supplementary table S3, Supplementary Material online). An arbitrary bootstrap threshold of 80% was applied to restrict the sorting to trees with moderate to high statistical support. This constituted the initial pool of EGT candidates with 362 trees. We also extended the condition to include the plastid-lacking stramenopiles (oomycetes, *Blastocystis*), alveolates (ciliates), and Rhizaria (*Reticulomyxa filosa*, *Gromia sphaerica*, and *Paracercomonas longicauda*) to account for the prediction that endosymbioses might have occurred in their common ancestors, but found no additional trees. Then, we manually scanned each tree for topologies consistent with EGTs and discarded the ones that did not contain at least *C. velia*, red and green algal representatives, and an outgroup. We used prokaryotic lineages as outgroup when possible, or alternatively members of animals, Fungi, or Amoebozoa. We also discarded trees with ≤ 10 taxa to reduce potential phylogenetic artifacts associated with poor taxon sampling (which ultimately did not contribute to the differences between our results and those of Woehle et al. [2011]). In parallel, we also evaluated the extent of land plant contamination by pooling the trees showing *C. velia* nested within monocotyledons (bootstrap support $\geq 80\%$). Finally, we monitored the alveolate and prokaryotic signals from the remaining 1,781 trees by searching for exclusive monophyletic grouping including *C. velia* and apicomplexans, dinoflagellates and/or ciliates, and *C. velia* and prokaryotes (bootstrap support $\geq 80\%$). Functional annotation of the EGT candidates was done with BLAST2GO (Götz et al. 2008).

Supplementary Material

Supplementary tables S1–S3, figures S1–S6 and supplementary materials are available at *Genome Biology and Evolution* online (<http://www.gbe.oxfordjournals.org/>).

Acknowledgments

We thank Sven B. Gould for sharing the *C. velia* cluster. This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada (227301) to P.J.K., by a grant from the Tula Foundation to the Centre for Microbial Diversity and Evolution, by the Czech Science Foundation, projects P506/12/1522 and P501/12/G055 to M.O., the Praemium Academiae award to J.L., and by Award IC/2010/09 made by the King Abdullah University of Science and Technology (KAUST) to A.P., M.O., and J.L. P.J.K. and J.L. are Fellows of the Canadian Institute for Advanced Research.

Literature Cited

Abrahamsen MS. 2004. Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. *Science* 304:441–445.

- Archibald JM. 2009. The puzzle of plastid evolution. *Curr Biol*. 19: R81–R88.
- Archibald JM, Rogers MB, Toop M, Ishida K-I, Keeling PJ. 2003. Lateral gene transfer and the evolution of plastid-targeted proteins in the secondary plastid-containing alga *Bigeloviella natans*. *Proc Natl Acad Sci U S A*. 100:7678–7683.
- Becker B, Hoef-Emden K, Melkonian M. 2008. Chlamydial genes shed light on the evolution of photoautotrophic eukaryotes. *BMC Evol Biol*. 8:203.
- Bowler C, et al. 2008. The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature* 456:239–244.
- Burki F, et al. 2007. Phylogenomics reshuffles the eukaryotic supergroups. *PLoS ONE* 2:e790.
- Burki F, Okamoto N, Pombert JF, Keeling PJ. 2012. The evolutionary history of haptophytes and cryptophytes: phylogenomic evidence for separate origins. *Proc R Soc B*. 279:2246–2254.
- Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25:1972–1973.
- Chan CX, Reyes-Prieto A, Bhattacharya D. 2011. Red and green algal origin of diatom membrane transporters: insights into environmental adaptation and cell evolution. *PLoS ONE* 6:e29138.
- Chan CX, et al. 2011. Red and green algal monophyly and extensive gene sharing found in a rich repertoire of red algal genes. *Curr Biol*. 1–6. doi: 10.1016/j.cub.2011.01.037.
- Dorrell RG, Smith AG. 2011. Do red and green make brown? Perspectives on plastid acquisitions within the chromalveolates. *Euk Cell*. 10: 856–868.
- Eisen JA, et al. 2006. Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. *PLoS Biol*. 4:e286.
- Frommolt R, et al. 2008. Ancient recruitment by chromists of green algal genes encoding enzymes for carotenoid biosynthesis. *Mol Biol Evol*. 25:2653–2667.
- Götz S, et al. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res*. 36:3420–3435.
- Gould SB, Waller RF, McFadden GI. 2008. Plastid evolution. *Annu Rev Plant Biol*. 59:491–517.
- Huang J, Gogarten JP. 2007. Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? *Genome Biol*. 8: R99.
- Huang J, et al. 2004. Phylogenomic evidence supports past endosymbiosis, intracellular and horizontal gene transfer in *Cryptosporidium parvum*. *Genome Biol*. 5:R88.
- Janouškovec J, Horák A, Obornik M, Lukeš J, Keeling PJ. 2010. A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids. *Proc Natl Acad Sci U S A*. 107:10949–10954.
- Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: the beginning of incongruence? *Trends Genet*. 22:225–231.
- Katoh K, Kuma K-I, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*. 33: 511–518.
- Keeling PJ. 2010. The endosymbiotic origin, diversification and fate of plastids. *Phil Trans R. Soc B*. 365:729–748.
- Lane CE, Archibald JM. 2008. The eukaryotic tree of life: endosymbiosis takes its TOL. *Trends Ecol Evol*. 23:268–275.
- Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22: 1658–1659.
- Lockhart P, et al. 2006. Heterotachy and tree building: a case study with plastids and eubacteria. *Mol Biol Evol*. 23:40–45.
- Martin W, et al. 1998. Gene transfer to the nucleus and the evolution of chloroplasts. *Nature* 393:162–165.
- Martin W, et al. 2002. Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of

- cyanobacterial genes in the nucleus. *Proc Natl Acad Sci U S A*. 99: 12246–12251.
- Moore RB, et al. 2008. A photosynthetic alveolate closely related to apicomplexan parasites. *Nature* 451:959–963.
- Moustafa A, Bhattacharya D. 2008. PhyloSort: a user-friendly phylogenetic sorting tool and its application to estimating the cyanobacterial contribution to the nuclear genome of *Chlamydomonas*. *BMC Evol Biol*. 8:7.
- Moustafa A, Reyes-Prieto A, Bhattacharya D. 2008. Chlamydiae has contributed at least 55 genes to Plantae with predominantly plastid functions. *PLoS ONE* 3:e2205.
- Moustafa A, et al. 2009. Genomic footprints of a cryptic plastid endosymbiosis in diatoms. *Science* 324:1724–1726.
- Oborník M, et al. 2011. Morphology, ultrastructure and life cycle of *Vitrella brassicaformis* n. sp., n. gen., a Novel Chromerid from the Great Barrier Reef. *Protist* 163:306–323.
- Palmer JD, Soltis D, Chase M. 2004. The plant tree of life: An overview and some points of view. *Am J Bot*. 91:1437–1445.
- Patron NJ, Waller RF. 2007. Transit peptide diversity and divergence: A global analysis of plastid targeting signals. *Bioessays* 29:1048–1058.
- Philippe H, Laurent J. 1998. How good are deep phylogenetic trees? *Curr Opin Genet Dev*. 8:616–623.
- Philippe H, Zhou Y, Brinkmann H, Rodrigue N, Delsuc F. 2005. Heterotachy and long-branch attraction in phylogenetics. *BMC Evol Biol*. 5:50.
- Reyes-Prieto A, Hackett JD, Soares MB, Bonaldo MF, Bhattacharya D. 2006. Cyanobacterial contribution to algal nuclear genomes is primarily limited to plastid functions. *Curr Biol*. 16:2320–2325.
- Reyes-Prieto A, Moustafa A, Bhattacharya D. 2008. Multiple genes of apparent algal origin suggest ciliates may once have been photosynthetic. *Curr Biol*. 18:956–962.
- Reyes-Prieto A, Weber APM, Bhattacharya D. 2007. The origin and establishment of the plastid in algae and plants. *Annu Rev Genet*. 41: 147–168.
- Rogers MB, Gilson PR, Su V, Mcfadden GI, Keeling PJ. 2007. The complete chloroplast genome of the chlorarachniophyte *Bigelowiella natans*: evidence for independent origins of chlorarachniophyte and euglenid secondary endosymbionts. *Mol Biol Evol*. 24:54–62.
- Stamatakis A. 2006. RAXML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.
- Stiller JW. 2011. Experimental design and statistical rigor in phylogenomics of horizontal and endosymbiotic gene transfer. *BMC Evol Biol*. 11:259.
- Stiller JW, Huang J, Ding Q, Tian J, Goodwillie C. 2009. Are algal genes in nonphotosynthetic protists evidence of historical plastid endosymbioses? *BMC Genomics* 10:484.
- Sun G, Yang Z, Ishwar A, Huang J. 2010. Algal genes in the closest relatives of animals. *Mol Biol Evol*. 27:2879–2889.
- Tyler BM. 2006. Phytophthora genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science* 313: 1261–1266.
- Woehle C, Dagan T, Martin WF, Gould SB. 2011. Red and problematic green phylogenetic signals among thousands of nuclear genes from the photosynthetic and apicomplexa-related *Chromera velia*. *Genome Biol Evol*. 3:1220–1230.

Associate editor: Bill Martin

Final conclusion

The heme pathway represents one of the most important biosynthetic processes in the vast majority of living organisms. It is responsible not only for the synthesis of heme but it also produces precursors for the synthesis of chlorophyll, which is particularly relevant for photosynthetic organisms. Yet, heme metabolism studies, until relatively recently, have been dedicated to a small portion of eukaryotes mostly belonging to metazoans and higher plants. However, along with the increase in the number of eukaryotic sequence data available, our understanding of heme metabolism in eukaryotes continuously grows and trends in the evolution of the heme pathway help us to gradually uncover the cloaks over their evolutionary mysteries.

During my postgraduate program, I focused mainly on studying the evolution of the heme pathway in various secondary or tertiary algae from different lineages with a particular emphasis on the spatial organization and evolutionary origins of individual enzymes. In order to elucidate the evolution of the heme pathway in a broader spectrum of phototrophic organisms we focused on *Bigeloviella natans* and *Guillardia theta*, algae containing a remnant endosymbiont nucleus within their plastids; on dinoflagellates containing tertiary endosymbionts derived from diatoms – called dinotoms; and on *Lepidodinium chlorophorum*, a dinoflagellate containing a secondary green plastid. This research resulted in many interesting findings. We have confirmed that *B. natans* shares an interesting feature of heme biosynthesis with *Euglena gracilis* (Kořený and Oborník, 2011; Genome Biol Evol. 3:359-64). Both algae possess two nearly complete redundant heme pathways, one representing the original heterotrophic pathway of the host (exosymbiont), and the other originates from the algal endosymbiont. This supports the proposal of both algae acquired their secondary green plastids relatively recently. Similarly, dinotoms possess two redundant heme pathways where one is located in the tertiary diatom endosymbiont, and the other represents the original pathway of the remnant peridinin-pigmented plastid. In *G. theta*, we discovered one additional ferrochelatase which is putatively targeted to the mitochondrion. This enzyme should be functionally linked with a single-copy protoporphyrinogen oxidase, which is probably dually targeted to both the plastid and the mitochondrion. This

indicates that *G. theta* is the only rhodophyte-derived secondary alga that retained, although only partially, the secondary host (exosymbiont) pathway. However, the most interesting discovery that came out from this research is that the heme pathway appears to be evolutionarily well conserved, even following serial endosymbioses, which is illustrated on the example of the dinoflagellate *Lepidodinium chlorophorum* and the chlorarachniophyte *Bigeloviella natans*. All the results were published by Cihlář et al., 2016 in PLoS ONE 11: e0166338.

I also participated in the genome analyses of chromerids (*Chromera velia* and *Vitrella brassicaformis*), where I focused mainly on finding and identifying genes of the heme pathway in these two free-living algae that are closely related to apicomplexan parasites. Although the heme pathway has been previously described in *C. velia* (Kořený et al., 2011; Plant Cell 23: 3454-3462), we further managed to identify one gene for ALA dehydrogenase and two genes for protoporphyrinogen oxidase, which had not been previously detected by rapid amplification of cDNA ends (RACE), within the *C. velia* genome. However, my primary goal was to describe the heme pathway of *V. brassicaformis*. We managed to find all the genes of interest and performed phylogenetic analyses and *in silico* targeting predictions to compare heme biosynthesis of both chromerids. These studies showed, that heme pathway in *V. brassicaformis* is homologous to that found in *C. velia*, and that both algae share one remarkable feature of heme biosynthesis, the use of the mitochondrial C4 pathway for the synthesis of the first precursor for the synthesis of both the heme and the chlorophyll. However, analyzes also revealed a difference in number and origins of the uroporphyrinogen synthase genes. This also supports the former proposal that chromerids form two independent lineages branching on the root of Apicomplexa. These findings were published by Woo et al., 2015 in eLife 4: e06974.

Further, we focused on the possible reasoning for the unusual ALA formation in both chromerid algae. I have found that both chromerids lack plastid-targeted enzymes for the synthesis of glutamate, which is the precursor of the plastid C5 pathway for tetrapyrrole synthesis. Both algae encode only the cytosolic NADH-dependent glutamate synthase. The loss of ferredoxin-dependent plastid glutamate synthase could

have therefore contributed to the loss of C5 pathway in these organisms. Our data imply that this loss happened probably already in the ancestor of apicomplexans, colpodellids, and chromerids. These findings are to be published within the scope of other experimental evidence our laboratory collected on the localization of heme pathway in chromerids.

Although my main topic was the evolution of the heme biosynthetic pathway in various eukaryotic phototrophs, at the very beginning of my postgraduate studies I participated in a project focused on evaluation of the impact of endosymbiotic gene transfer (EGT) on eukaryote genomes by re-analyzing an expressed sequence tag (EST) dataset for *Chromera velia*. We managed to substantially lower the previously published estimates of genes related to EGT. We also found indications of sequence contamination from land plants and accordingly corrected the number of genes acquired from red and green algae in *C. velia*, in the data presented by Woehle et al., 2011 (Genome Biol Evol. 3:1220-30). This work was published by Burki et al., 2012 in Genome Biol Evol. 4:626-35.

We also investigated the evolutionary origin of porphobilinogen deaminase, which is in most eukaryotic phototrophs related to α -proteobacterial/mitochondrial sequences. We have several explanation for this evolutionary distribution, and our data allowed us to propose a model of the PBGD starting from three genes (of eukaryotic/exosymbiont, mitochondrial and cyanobacterial origins) that were supposedly present in the ancestor of Archaeplastida. These findings are to be published.

In conclusion, I would like to express my impression that during my Ph.D. studies I gained insight into the unsolved questions about the heme metabolism. That allowed me to recast the current view of the evolution of this pathway in phototrophic eukaryotes into the form of an introductory review to my Ph.D. thesis. The review is not only building on previous knowledge but also contains novel data as my own scientific contribution.

Curriculum vitae

Name: Jaromír Cihlář
Nationality: Czech
Date of birth: 19 April 1984
Home address: Větrná 1454/72, 370 05 České Budějovice, Czech Republic

EDUCATION

2011 – present **Ph.D. student** of Molecular and Cell Biology and Genetics

Department of Molecular Biology and Genetics, Faculty of Science,
University of South Bohemia, České Budějovice

Institute of Parasitology, **Czech Academy of Sciences**, Czech
Republic

Thesis: Evolution of the heme biosynthetic pathway in eukaryotic
phototrophs

Supervisor: Miroslav Oborník

2008 – 2011 **M.S., Experimental Biology**

Department of Molecular Biology, Faculty of Science, **University of
South Bohemia**, Czech Republic.
Thesis: Molecular characterization of novel photosynthetic protozoan
strain from corals. Supervisor: Miroslav Oborník

2004 – 2008 **B.S., Biology**

Department of Molecular Biology, Faculty of Science, **University of
South Bohemia**, Czech Republic.
Thesis: Preparation of isotopic-labeled single-strand DNA
oligonucleotides for NMR spectroscopy. Supervisor: Lukáš Trantírek

EMPLOYMENT

2011 – present Graduate student

Institute of Parasitology, Biology Centre of the Czech Academy of
Sciences

INTERNSHIPS

2015 (IV.-VI., 8 weeks) Prof. Peter Kroth Lab, University of Konstanz, Germany

TEACHING ACTIVITIES

Faculty of Science, University of south Bohemia in České Budějovice, Czech Republic:
Methods in Molecular Biology, Laboratory instructor, Summer 2012, 2013

PUBLICATIONS

Cihlář J, Füssy Z, Horák A, Oborník M. (2016) Evolution of the Tetrapyrrole Biosynthetic Pathway in Secondary Algae: Conservation, Redundancy and Replacement. *PLOS One*. 11(11):e0166338.

Woo YH, Ansari H, Otto TD, Klinger CM, Kolisko M, Michálek J, Saxena A, Shanmugam D, Tayyrov A, Veluchamy A, Ali S, Bernal A, del Campo J, **Cihlář J**, Flegontov P, Gornik SG, Hajdušková E, Horák A, Janouškovec J, Katris NJ, Mast FD, Miranda-Saavedra D, Mourier T, Naeem R, Nair M, Panigrahi AK, Rawlings ND, Padron-Regalado E, Ramaprasad A, Samad N, Tomčala A, Wilkes J, Neafsey DE, Doerig C, Bowler C, Keeling PJ, Roos DS, Dacks JB, Templeton TJ, Waller RF, Lukeš J, Oborník M, Pain A. (2015) Chromerid genomes reveal the evolutionary path from photosynthetic algae to obligate intracellular parasites. *eLife*. 4:e06974.

Burki F, Flegontov P, Oborník M, **Cihlář J**, Pain A, Lukeš J, Keeling PJ. (2012) Re-evaluating the green versus red signal in eukaryotes with secondary plastid of red algal origin. *Genome Biology and Evolution*. 4(6):626-35.

Oborník M, Modrý D, Lukeš M, Cernotíková-Stříbrná E, **Cihlář J**, Tesařová M, Kotabová E, Vancová M, Prášil O, Lukeš J. (2012) Morphology, ultrastructure and life cycle of *Vitrella brassicaformis* n. sp., n. gen., a novel chromerid from the Great Barrier Reef. *Protist*. 163(2):306-23.

PRESENTATIONS AT CONFERENCES

- 2016 46th International Meeting of Czech Society for Protozoology
Bítov, Czech Republic
The evolution of heme pathway with regard to endosymbiotic gene transfer. Poster presentation.
- 2014 44th International Meeting of Czech Society for Protozoology
Visalaje, Czech Republic
Heme Pathway of *Vitrella brassicaformis*. Talk presentation.
- 2013 43th International Meeting of Czech Society for Protozoology
Nový Dvůr nad Vltavou, Czech Republic
Porphobilinogen Deaminase in Phototrophic Eukaryotes and its Mitochondrial Origin. Poster presentation.
- 2012 Endosymbiosis 2012
München, Germany
Novel isolate of *Chromera* sp. from the Great Barrier Reef. Poster presentation.
-