

Biology Centre, Institute of Parasitology, Czech Academy of Sciences, and
University of South Bohemia, Faculty of Sciences,
České Budějovice, Czech Republic

Ph.D. Thesis

Genomics of *Blastocrithidia*,
a trypanosomatid with all three stop
codons reassigned

Anna Nenarokova, MSc

Supervisor: Prof. RNDr. Julius Lukeš, CSc.

České Budějovice, 2020

This thesis should be cited as: Nenarokova A., 2019: Genomics of *Blastocrithidia*, a trypanosomatid with all three stop codons reassigned. Ph.D. thesis. University of South Bohemia, Faculty of Science, School of doctoral studies in biological sciences, České Budějovice, Czech Republic.

Annotation

This work describes genomic studies of *Blastocrithidia*, the trypanosomatid with all three stop codons encoding amino acids and the only known euglenozoan with non-standard nuclear genetic code. We describe unique genomic features of *Blastocrithidia* and the sister clade "*Jaculum*", and discuss possible mechanisms of translation termination with ambiguous stop codons and the evolutionary path that may have established this system.

Declaration

I hereby declare that I did all the work presented in this thesis by myself or in collaboration with the mentioned co-authors and only using the cited literature.

České Budějovice, Anna Nenarokova

Prohlášení

Prohlašuji, že svoji disertační práci jsem vypracovala 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 úpravě vzniklé vypuštěním vyznačených částí archivovaných Přírodovědeckou fakultou 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.

V Českých Budějovicích dne _____, Anna Nenarokova _____

This thesis originated from a partnership between the Faculty of Science, University of South Bohemia, and Institute of Parasitology, Biology Centre of the Czech Academy of Sciences, supporting doctoral studies in the Molecular and Cell biology and Genetics study program.

Financial support

This work was supported by the Czech Grant Agency (15–21974S and 16–18699S), the ERC CZ grant (LL1601), the Czech Ministry of Education (ERD Funds OPV VV16_019/ 0000759).

Acknowledgements

First of all, I wish to sincerely thank my supervisor Julius Lukeš III for believing in me, his patience and support through my Ph.D. studies. I admire his courage and boldness, endless life force and enthusiasm.

Next, I would like to thank my colleagues and collaborators on the *Blastocrithidia* project:

Kika Záhonová, who is the first author of the original *Blastocrithidia* paper and did a significant part of the bioinformatic work of the current project;

Serafim Nenarokov, who developed the annotation software and has been helping me in many other ways;

Ambar Kachale, who did most of the experimental work;

Zdeněk Paris, who is our advisor for the tRNA part of the project and performed some of the tRNA experiments;

Míša Svobodová, who prepared the RNAi cell lines of genuine and putative cytidine deaminases of *T. brucei*;

Eva Horáková, who is one of our advisors for experimental part of the project;

Jan Votýpka and Alexey Kostygov, who provided the cultures of *Blastocrithidia* sp. and “*Jaculum*”;

Juan Alfonzo for useful discussions;

Slava Yurchenko, who is a co-supervisor of the project and contributed a lot of high-quality sequence data;

Marek Eliáš, who is the senior author of the original *Blastocrithidia* paper and who shared with us many smart and useful ideas.

Besides, I want to thank Ambar Kachale and Míša Svobodová for their help with the description of the experimental part of the work.

I am very grateful to other members of the TriTryp labs and my other colleagues from the Institute of Parasitology who were very friendly and helped me in many ways.

Also, I want to express my gratitude to Drs. Tom Williams and Celine Petitjean from the University of Bristol for taking me to practice in their laboratory, being very friendly and welcoming to me, and sharing with me their experience and knowledge.

Last, but not least, I would like to thank all my friends and loved ones who stood on my side, who took care of me in my lowest moments, when I felt down and desperate. Without their help and support I would not get here.

List of publications and author's contribution

Publications listed chronologically:

1. Kostygov A.Y., Butenko A., **Nenarokova A.**, Tashyreva D., Flegontov P., Lukeš J. and Yurchenko V. (2017) Genome of *Ca. Pandoraea novymonadis*, an endosymbiotic bacterium of the trypanosomatid *Novymonas esmeraldas*. *Front. Microbiol.* 8:1940. doi:10.3389/fmicb.2017.01940
Anna Nenarokova assembled and annotated the genomes of Novymonas esmeraldas and its symbiont Ca. Pandoraea novymonadis and contributed to the interpretation of data.
2. Zíková A., Verner Z., **Nenarokova A.**, Michels P.A.M., Lukeš J. (2017). A paradigm shift: The mitoproteomes of procyclic and bloodstream *Trypanosoma brucei* are comparably complex. *PLoS Path.* 13 (12), e1006679. doi:10.1371/journal.ppat.1006679
Anna Nenarokova participated in the data preparation and analysis.
3. Schwelm A., Badstöber J., Bulman S., Desoignies N., Etemadi M., Falloon R. E., Gachon C. M. M., Legreve A., Lukeš J., Merz U., **Nenarokova A.**, Strittmatter M., Sullivan B. K. and Neuhauser S. (2018) Not in your usual Top 10: protists that infect plants and algae. *Molecular Plant Pathology.* doi:10.1111/mpp.12580
Anna Nenarokova wrote the chapter about the genus Phytomonas in this review.
4. Harmer J., Yurchenko V., **Nenarokova A.**, Lukeš J., Ginger M.L. (2018). Farming, slaving and enslavement: histories of endosymbioses during kinetoplastid evolution. *Parasitology* 145, 1311–1323. doi:10.1017/S0031182018000781
Anna Nenarokova participated in the data preparation and analysis.
5. Ebenezer T., Zoltner M., Burrell A., **Nenarokova A.**, Novák Vanclová A.M.G., Prasad B., Soukal P., Santana-Molina C., O'Neill E., Nankisoor N.N., Vadakedath N., Daiker V., Obado S., Jackson A.P., Devos D., Lukeš J., Lebert M., Vaughan S., Hampl V.,

Carrington M., Ginger M.L., Dacks J.B., Kelly S., Field M.C. (2018). Transcriptome, proteome and draft genome of *Euglena gracilis*. BMC Biology 17 (1), 11.
doi:10.1186/s12915-019-0626-8

Anna Nenarokova predicted in silico the mitochondrial proteome of Euglena gracilis, functionally annotated mitochondrial proteins and pathways, and wrote the chapter about mitochondrial proteins.

6. **Nenarokova A.**, Záhonová K., Krasilnikova M., Gahura O., McCulloch R., Zíková A., Yurchenko V., Lukeš J. (2019). Causes and effects of loss of classical nonhomologous end joining pathway in parasitic eukaryotes. mBio 10 (4), e01541-19.
doi:10.1128/mBio.01541-19.

Anna Nenarokova conceived and designed the study, performed most of the data analysis, and wrote most of the article.

7. Kosakyan A., Alama-Bermejo G., Bartošová-Sojtková P., Born-Torrijos A., Šíma R., **Nenarokova A.**, Eszterbauer E., Bartholomew J., Holzer A.S. (2019). Selection of suitable reference genes for gene expression studies in myxosporean (Myxozoa, Cnidaria) parasites. Scientific Reports. 9(1):15073. doi:10.1038/s41598-019-51479-0.

Anna Nenarokova helped with the transcriptome assembly, filtering and analysis.

8. Hammond M.J., **Nenarokova A.**, Butenko A., Zoltner M., Lacová Dobáková E., Field M.C., Lukeš J. A uniquely complex mitochondrial proteome from *Euglena gracilis*. Re-submitted to Molecular Biology and Evolution.

Anna Nenarokova made the initial bulk functional annotation of MS-MS determined proteins from the E. gracilis mitochondrial fraction, assigned them to functional modules and reconstructed mitochondrial pathways, and participated in data analysis and interpretation. Moreover, she wrote the chapter “RNA editing and processing”.

Prof. RNDr. Julius Lukeš, CSc. _____

Table of contents

Preface	11
My role in the <i>Blastocrithidia</i> project	12
Structure of the thesis	12
1 Introduction	13
1.1.1 <i>Blastocrithidia</i> phylogenetic position, evolution and systematics	13
1.2 Gene expression in kinetoplastids	16
1.2.1 Genome organization, transcription and <i>trans</i> -splicing	16
1.2.2 Post-transcriptional events and translation in kinetoplastids	19
1.3 Evolution and variations of genetic code	20
1.3.1 Standard genetic code and its origin	20
1.3.2 Non-standard genetic codes in nature	21
1.3.3 Codon reassignment theories	23
1.3.4 Translation termination and stop codon reassignment	25
1.3.5 Organisms with all three stop codons reassigned	27
1.3.6 Genetic code in kinetoplastids	28
2 Materials and Methods	30
2.1 Cell cultures	30
2.2 Poisoned primer extension	30
2.3 RNAi of <i>T. brucei</i> cytidine deaminase candidate genes	31
2.4 Genome and transcriptome sequencing, assembly and mapping	31
2.5 Mass spectrometry analysis	32
2.6 BLAST searches, alignments	33
2.7 Genome annotation	33
2.8 tRNA prediction and analysis	34

2.9	Codon usage	34
3	Results and Discussion	35
3.1	Genome statistics	35
3.2	Translation termination in <i>Blastocrithidia</i>	36
3.2.1	Defining the genuine stop codon	36
3.3	Potential mechanisms of translation termination in <i>Blastocrithidia</i>	38
3.4	Codon usage	40
3.4.1	Codon usage and GC content	40
3.4.2	Usage of the reassigned stop codons	41
3.5	tRNAs	45
3.6	NMD is absent in the <i>Blastocrithidia</i> clade	51
3.7	Insertions and DNA repair in <i>Blastocrithidia</i> and other trypanosomatids	53
4	Summary	54
4.1	Hypotheses about the evolutionary path of <i>Blastocrithidia</i>	54
4.2	Conclusions and perspectives	55
5	References	56
6	Supplementary information	73
	Figure S1. Proteomic proof of the reassigned stop codons in <i>Blastocrithidia</i>	73
	Figure S2. GC-content around genuine stop codon in trypanosomatids	75
	Table S1. Genes of <i>Blastocrithidia</i> sp. P57 without reassigned stop codons	78
	Table S2. <i>T. brucei</i> tRNA ^{Trp} _{CCA} -editing cytidine deaminases candidates	83
7	Full text of the publications included in the thesis	84
7.1	Genome of <i>Ca. Pandoraea novyimonadis</i> , an endosymbiotic bacterium of the trypanosomatid <i>Novyimonas esmeraldas</i>	84
7.2	A paradigm shift: The mitoproteomes of procyclic and bloodstream <i>Trypanosoma brucei</i> are comparably complex	97

7.3	Not in your usual Top 10: protists that infect plants and algae	106
7.4	Farming, slaving and enslavement: histories of endosymbioses during kinetoplastid evolution	122
7.5	Transcriptome, proteome and draft genome of <i>Euglena gracilis</i>	135
7.6	Causes and effects of loss of classical nonhomologous end joining pathway in parasitic eukaryotes	158
7.7	Selection of suitable reference genes for gene expression studies in myxosporean (Myxozoa, Cnidaria) parasites	170
7.8	A uniquely complex mitochondrial proteome from <i>Euglena gracilis</i>	184

Preface

Genetic code is a set of rules that living organisms use to translate the information from nucleic acids into amino acids. The canonical genetic code, which was deciphered in 1960s, contains 64 triplets, 61 of which code for amino acids and three are stop codons, denoting the end of translation. Although it is often referred to as “universal”, a more accurate term would be “quasi-universal”, because alterations of genetic code are lately found to be increasingly more widespread in extant organisms and organelles.

In 2016, trypanosomatid *Blastocrithidia* and ciliates *Condylostoma* and *Parduczia* were shown to use new genetic code variants with all three stop codons encoding amino acids. This finding challenges another key feature of the genetic code – its unambiguity. The mechanisms behind these reassignments and translation termination remain unclear. Arguably, *Blastocrithidia* represents an ideal model system for studying such a phenomenon for several reasons. It belongs to the order Trypanosomatida, a well-studied protist group, which includes model objects, such as *Trypanosoma* and *Leishmania* species, with available complete genomes, and a wide range of established methods and techniques. Unlike ciliates, which are well-known for altering their genetic code, all known trypanosomatids aside from the genus *Blastocrithidia* have canonical nuclear genetic code, so the reassignment must have happened in this group relatively late on the evolutionary scale. Thus, in trypanosomatids, we should be able to trace the main steps leading to the emergence of such a system.

Originally, this weird genetic code was described from an uncultured *Blastocrithidia* sp. from a heteropteran bug *Lygus hesperus*. Later on, we were able to grow two other species of *Blastocrithidia* in culture. We have sequenced the genomes, transcriptomes and proteomes of these two strains, and the genome and transcriptome of *Leptomonas jaculum* (“*Jaculum*”), a closely related species with a canonical genetic code. We have created new software for annotation of the *Blastocrithidia* genome, as existing annotation programs were not able to deal with ambiguous stop codons. This allowed us to look at the reassigned stop codons from a wider perspective to see the general trends in their features and distribution. Besides, we have performed mass spectrometry analysis of *Blastocrithidia* proteins to check experimentally the predicted meaning of the in-frame stop codons. Also, we have predicted and experimentally analyzed tRNA species that are responsible for stop codon decoding. We show the unique

prerequisites of the *Blastocrithidia* lineage, which made the reassignment of all stop codons possible and discuss how translation termination functions in this bizarre trypanosomatid flagellate.

My role in the *Blastocrithidia* project

I have been responsible for the bioinformatic part of the project. I have planned, designed and implemented most steps of the bioinformatic analysis, and coordinated the work of two other bioinformaticians, postdoctoral researcher Kristína Záhonová, the first author of the original 2016 paper, and Ph.D. student Serafim Nenarokov. I have invented the concept and the algorithm for our gene annotation software, capable to cope with the ambiguous stop codons of *Blastocrithidia*, and supervised Serafim, who developed the program. I also have been actively participating in planning, design and interpretation of the experimental part of the project.

Structure of the thesis

This thesis consists of three main parts: introduction, description of the unpublished work, and the full text of the publications that I have contributed to during my Ph.D. studies.

1 Introduction

1.1.1 *Blastocrithidia* phylogenetic position, evolution and systematics

Genus *Blastocrithidia* Laird, 1959 belongs to the family Trypanosomatidae, order Trypanosomatida, and class Kinetoplastea, which is part of the eukaryotic phylum Euglenozoa (Adl et al., 2019).

Euglenozoa is a monophyletic group of mono- or bi-flagellated protists characterized by a set of specific morphological features, most prominent of which are a closed mitosis with an intranuclear spindle, a single reticulated mitochondrion with discoidal cristae, and with one or two flagella inserted in an apical flagellar pocket with an emergent part containing heteromorphic paraflagellar rod (Adl et al., 2019; Cavalier-Smith, 1981; Simpson, 1997). This phylum is subdivided into three classes: Kinetoplastea, Diplonemea, Euglenoidea, and a group with an unresolved position – Symbiontida (Adl et al., 2012; Simpson, 1997).

The class Euglenoidea includes well-known photosynthetic representatives, such as *Euglena viridis* and a model species, *Euglena gracilis*. Euglenoids are characterized by two heterologous flagella and a rigid cover of parallel strips – a pellicula (Sommer, 1965). While it is the only euglenozoan group containing plastids, most euglenoids are non-photosynthetic (Leander et al., 2001). The photosynthetic group evolved from a predecessor that acquired a green alga, which was transformed into a secondary plastid (Keeling, 2009). Symbiontida, which is apparently a derived or a sister group of euglenids, contains only three known species found in the low-oxygen sea bottom and covered with a dense layer of ectosymbiotic bacteria (Yubuki et al., 2013).

Diplonemids (class Diplonemea) were thought to be a rare and thus ecologically insignificant group within Euglenozoa, but recent studies showed that this class includes a huge unknown clade of marine species, which surprisingly happen to be the most diverse and the 5th or 6th most abundant group of eukaryotic plankton in the World Ocean (Flegontova et al., 2016). While little is known about their morphology and cell biology, diplonemids carry two flagella of varying length during their life cycle (Prokopchuk et al., 2019), and contain huge amount of DNA in their mitochondrion equipped with giant cristae (Lukeš et al., 2018).

Finally, the class Kinetoplastea is characterized by the presence of a kinetoplast – an electron-dense structure within the mitochondrion, which consists of a large amount of

mitochondrial DNA, usually composed of catenated DNA circles (Lukeš et al., 2002). Traditionally, kinetoplastids are divided into two groups: the obligatory parasitic Trypanosomatida and the mostly free-living Bodonida, even though the former group is monophyletic and has originated inside the latter group, which is thereby paraphyletic (Deschamps et al., 2011).

Family Trypanosomatidae Doflein 1951 was defined as a taxon with invariably monoflagellar species that contain a disk-shaped kinetoplast composed of mutually interlocked mini- and maxicircles (Povelones, 2014). It is an extremely speciose group of obligatory parasites, invading invertebrates, vertebrates and plants (Maslov et al., 2013). Being the causative agents of severe diseases of people and livestock, members of the genera *Trypanosoma* and *Leishmania* are most important. African trypanosomiasis also known as sleeping sickness, is caused by *Trypanosoma brucei*, while *Trypanosoma cruzi* is responsible for American trypanosomiasis (Chagas disease), and leishmaniases are caused by various *Leishmania* species (Barrett et al., 2003; Myler and Fasel, 2008). All these pathogens have a dixenous life cycle, alternating between a blood-sucking invertebrate primary host (usually, an insect) and a secondary host – human or another vertebrate. Exceptionally, the invertebrate host was eliminated when transmission of the trypanosomes between two vertebrates is possible (Lai et al., 2010). Reflecting the evolutionary flexibility of trypanosomatids, the third known dixenous genus *Phytomonas* has adapted to sap-sucking rather than blood-sucking insects as primary hosts, while plants serve as secondary hosts instead of vertebrate animals (Jaskowska et al., 2015). However, most trypanosomatids have a monoxenous lifestyle and are thus confined to invertebrate hosts only (Maslov et al., 2013). Abundant evidence points to the monoxenous lifestyle being ancestral, from which the dixenous strategy evolved independently at least three times, namely in the genera *Trypanosoma*, *Leishmania*, and *Phytomonas* (Lukeš et al., 2014).

The genus *Blastocrithidia* Laird, 1959 was originally defined by a set of morphological features. However, already at an early stage it was shown by molecular methods that the original *Blastocrithidia* taxon was polyphyletic (Merzlyak et al., 2001), and some of the species were later transferred to other groups. For example, the endosymbiont-bearing *Blastocrithidia culicis* together with another endosymbiont-containing trypanosomatid *Crithidia oncopelti* were reclassified into a newly erected genus *Strigomonas* (Teixeira et al., 2011). Nevertheless, other *Blastocrithidia* species with available sequence data, namely *B. triatomae*, *B. leptocoridis*, *B.*

cyrtometri, *B. papi*, *B. largi*, *B. miridarum*, *Blastocrithidia* sp. 393M8 ex *Lygus* sp. and *Blastocrithidia* sp. ex *Lygus hesperus* form a well-supported monophyletic group, with the

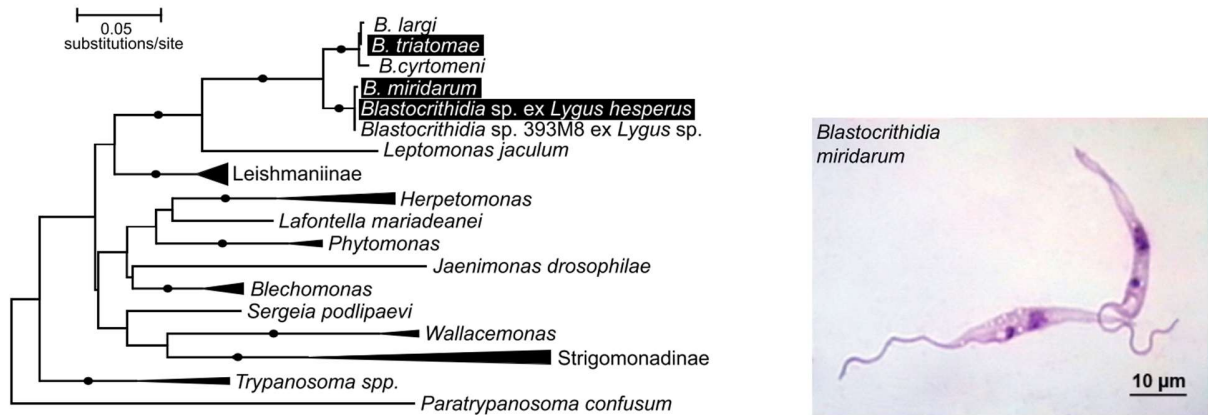


Figure 1. Phylogenetic position of *Blastocrithidia* and “*Jaculum*” (*Leptomonas jaculum*) (Záhonová et al., 2016)

“*Jaculum*” clade (*Leptomonas jaculum*-like species) constituting its closest outgroup (further in the text referred to as “*Jaculum*”) (Caicedo et al., 2011; Frolov et al., 2017; Votýpka et al., 2015; Záhonová et al., 2016) . This group is subdivided into two subgroups: the first one consists of *B. triatomae*, *B. cyrtometri*, *B. leptocoridis*, *B. papi*, and *B. largi*, and the second subgroup contains *B. miridarum*, *Blastocrithidia* sp. 393M8 ex *Lygus* sp. and *Blastocrithidia* sp. ex *Lygus hesperus* (Frolov et al., 2017; Záhonová et al., 2016) (Fig. 1). Furthermore, all *Blastocrithidia* species checked for their genetic code, namely *B. triatomae* from the first subgroup and *Blastocrithidia* sp. ex *L. hesperus* and *B. miridarum* from the second subgroup, use a genetic code modified in the same manner (Záhonová et al., 2016) (Fig. 1). Hence, this genetic code variant can be considered as a potential synapomorphy of the whole genus *Blastocrithidia*.

1.2 Gene expression in kinetoplastids

1.2.1 Genome organization, transcription and *trans*-splicing

Kinetoplastids lack introns (besides the exception of two genes (Mair et al., 2000; Preusser et al., 2014)), though both diplomemids and euglenids have many canonical and non-canonical introns (Gawryluk et al., 2016; Milanowski et al., 2016). Kinetoplastids and other euglenozoans have polycistronic transcription, which is the characteristic feature of prokaryotes, but not eukaryotes (Chung et al., 1990; Kozak, 1983; Worthey et al., 2003). Indeed, the kinetoplastid polycistronic transcription has little in common with the prokaryotic one (Nascimento et al., 2018). Bacterial genes are organized into small functional groups termed operons: proteins translated from a single operon usually act in distinct stages of the same process, as is the case of e.g. the *lac* operon (Salgado et al., 2000).

In kinetoplastids, functionally unrelated genes are organized into several large polycistronic transcriptional units, separated by strand-switch regions that act as transcription starts, initiating bidirectional transcription by polymerase II (Fig. 2) (Berriman et al., 2005; Martínez-Calvillo et al., 2004, 2003; Myler et al., 1999; Teixeira et al., 2012). Some of polycistronic transcriptional units also have additional internal transcription starts (Fig. 2) (Kolev et al., 2010). Both strand-switch regions and internal transcriptional starts do not have a conserved nucleotide sequence but are instead specified by histone variants and modifications (Anderson et al., 2013; Maree and Patterton, 2014; Siegel et al., 2005; Thomas et al., 2009). Transcription termination sites are also defined epigenetically: by the kinetoplastid-specific modified thymidine base J (Maree and Patterton, 2014), histone variants (Reynolds et al., 2014), and/or by the subsequent loci, transcribed by other RNA polymerases (such as tRNA and small RNA gene clusters, transcribed by RNA polymerase III, or pre-rRNA and surface proteins (VSGs and procyclins), transcribed by polymerase I) (Clayton, 2016; Martínez-Calvillo et al., 2004; Palenchar and Bellofatto, 2006) (Fig. 2). Notably, genes of trypanosomatids are located non-randomly in respect to the transcription initiation sites: it was shown that in *T. brucei* genes downregulated following a heat shock are located close to the transcription start sites, while genes upregulated following the same trigger are distant (Kelly et al., 2012). The authors explain this phenomenon by the fact that during heat shock transcription initiation is arrested, while transcription elongation still takes place, therefore only genes remote from transcription initiation

sites are expressed (Kelly et al., 2012). However, not only heat shock-related genes but many other gene categories demonstrate positional bias. Moreover, the distance of genes from their polycistronic transcription starts is related to the mRNA abundance in different phases of the cell division cycle, thus providing a mechanism for temporal regulation of gene expression (Kelly et al., 2012).

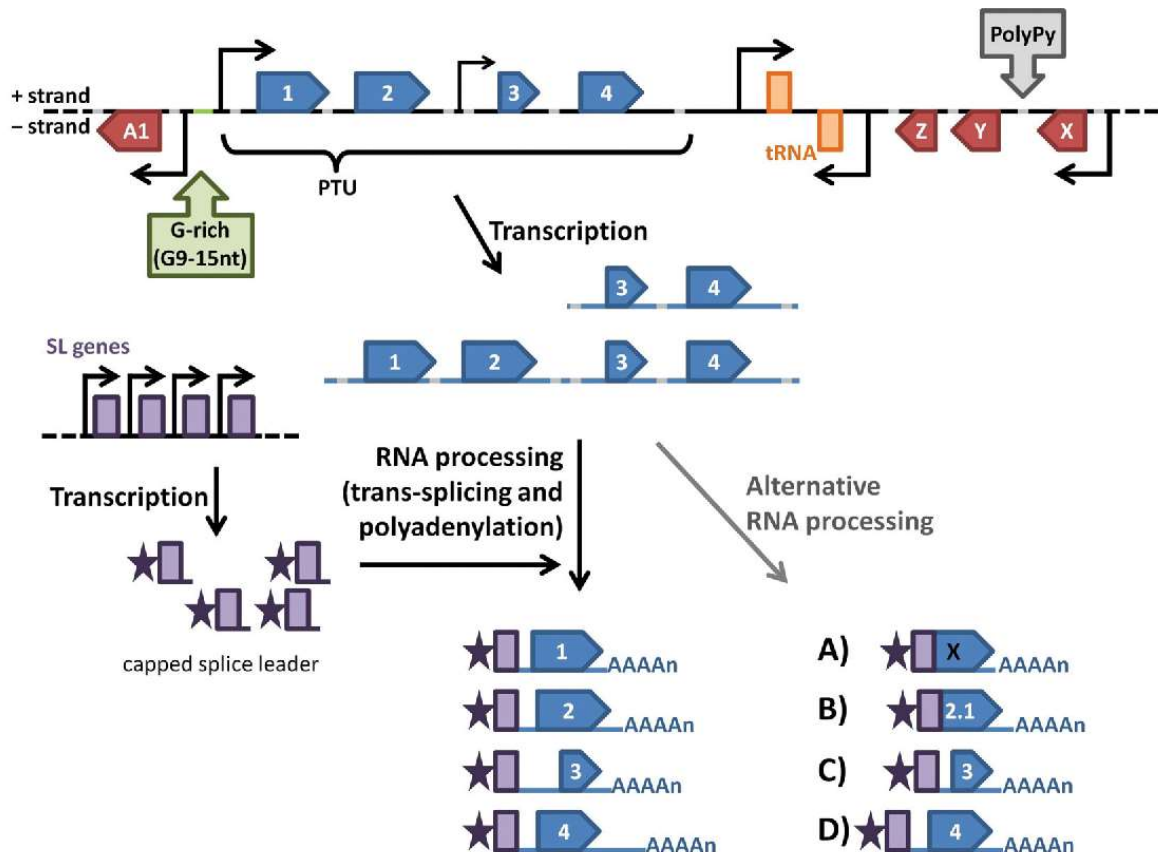


Figure 2. Gene expression in trypanosomatids (Teixera et al., 2012).

Arrow boxes represent genes, green line – strand-switch region, PTU – polycistronic transcriptional unit, grey lines (PolyPy) – polypyrimidine tracts, bend arrows – transcription starts.

Hence, the specific gene order in polycistronic transcription units and the length of them provide a mechanism of transcription control in kinetoplastids, which lack canonical eukaryotic differential regulation of transcription (Clayton, 2016; Kelly et al., 2012). The genomes of kinetoplastids are highly syntenic which is believed to be the consequence of their specific organization (Fig. 3) (Ghedini et al., 2004; Peacock et al., 2007).

The transcriptional units are subsequently processed into individual transcripts, to the 5' end of which a specific short RNA termed spliced leader (SL), transcribed from specific genome loci by polymerase II and capped, and attached by *trans*-splicing (Liang et al., 2003) (Fig. 2). *Trans*-splicing is performed in the nucleus by the spliceosome, a eukaryote-specific structure, which usually performs *cis*-splicing (Liang et al., 2003). Kinetoplastids also perform *cis*-splicing, but it is confined to only two genes with canonical *cis*-spliced introns: poly(A) polymerase and putative RNA helicase (Mair et al., 2000; Preusser et al., 2014). Polyadenylation of the transcripts is coupled with *trans*-splicing (LeBowitz et al., 1993).

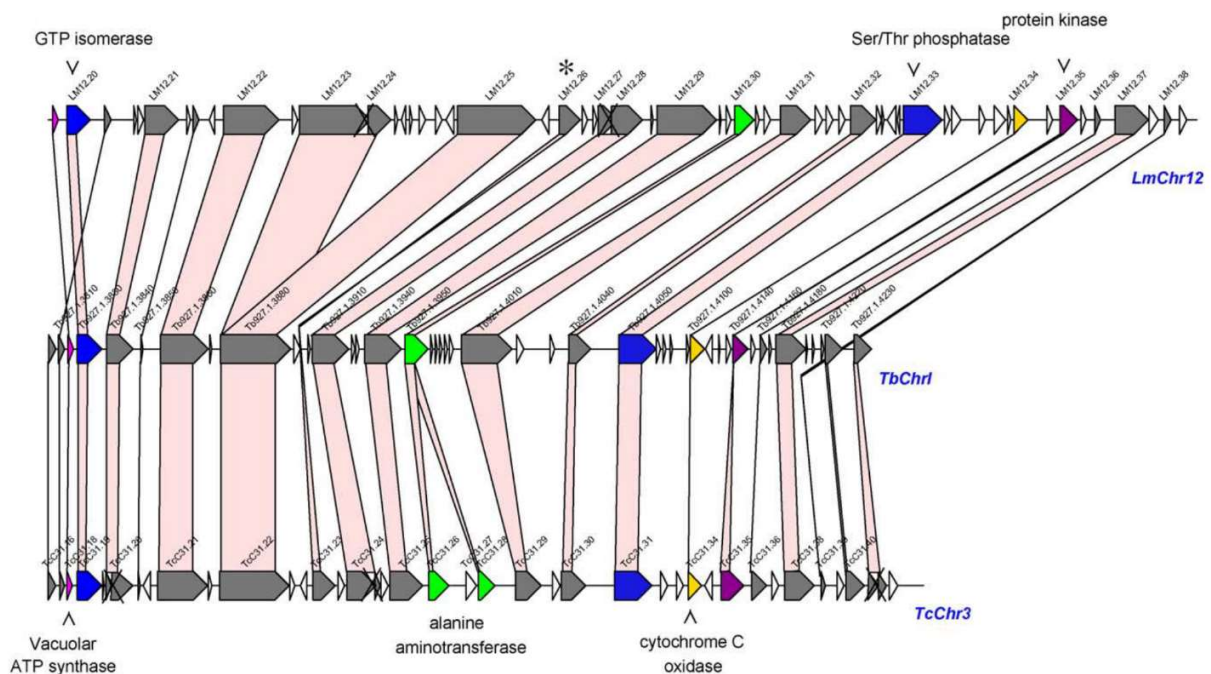


Figure 3. Syntenic region between *L. major*, *T. brucei* and *T. cruzi* (Ghedini et al., 2004).

Intriguingly, these two strange features of gene expression, namely SL RNA *trans*-splicing and polycistronic transcription, are not unique for euglenozoans, as both of them are found also in dinoflagellates and *Caenorhabditis elegans* (Krause et al., 1987; Lasda and Blumenthal, 2011). The emergence of the *trans*-splicing machinery is considered to be the premise of polycistronic transcription: *trans*-splicing allows the avoidance of a complex and labor-intensive process of transcription initiation (Lukeš et al., 2009). Indeed, this prevents fine-tuning of gene expression on the level of transcription but provides a faster and more efficient

process. Therefore, being incapable of a complex regulation on the level of transcription, which is common to almost all other eukaryotes, kinetoplastids have instead evolved many ways to regulate gene expression at the post-transcriptional level (Clayton and Shapira, 2007).

1.2.2 Post-transcriptional events and translation in kinetoplastids

Kinetoplastids are also well-known for their mitochondrial RNA editing machinery. Their mitochondrion employs dozens of essential proteins to ensure proper insertions or deletions of uridine residues in hundreds of sites in order to correct the disrupted open reading frame in most mitochondrial-encoded transcripts (Read et al., 2015). The complexity of this unique system is considered by some an outstanding example of constructive neutral evolution (Lukeš et al., 2011).

Translation process is divided into three main parts: initiation, elongation, and termination. In trypanosomatids, the most studied process among these is translation initiation (Zinoviev and Shapira, 2012). All studied trypanosomatids have five eIF4G, and six eIF4E, as well as two PABP paralogs (PABP1 and PABP2), while *Leishmania* spp. also have a third paralogue, PABP3 (da Costa Lima et al., 2010; Zinoviev and Shapira, 2012). In both *Trypanosoma* and *Leishmania* spp. PABP1 predominantly binds to poly(A) tails, while PABP2 is more promiscuous and binds other types of RNA apart from poly(A) sequences (da Costa Lima et al., 2010; Zoltner et al., 2018).

1.3 Evolution and variations of genetic code

1.3.1 Standard genetic code and its origin

First pointed out by Crick, the universality of the standard genetic code (Fig. 4) is arguably the strongest argument for the existence of Last Universal Common Ancestor (LUCA) (Crick, 1968; Koonin, 2009, 2017). Any variations of it, once established, would be deleterious to any living organism, as postulated by a theory called “Frozen accident” (Crick, 1968). However, why is there only one genetic code that has been “frozen”? Why do we not see more genetic codes, apart from the small variations of the standard one? A brilliant explanation was provided by Vetsigian and colleagues: the existence of multiple genetic codes makes horizontal gene transfer almost impossible, and this process, still being very important in extant prokaryotes and certain eukaryotes, was crucial in the early stages of life (Vetsigian et al., 2006). Indeed, organisms capable of effective gene sharing via horizontal gene transfer were probably evolutionally more successful than their “individualistic” competitors (Koonin, 2017; Vetsigian et al., 2006).

		Second Letter					
		U	C	A	G		
1st letter	U	UUU Phe UUC UUA Leu UUG	UCU UCC Ser UCA UCG	UAU Tyr UAC UAA Stop UAG Stop	UGU Cys UGC UGA Stop UGG Trp	U C A G	
	C	CUU CUC Leu CUA CUG	CCU CCC Pro CCA CCG	CAU His CAC CAA Gln CAG	CGU CGC Arg CGA CGG	U C A G	
	A	AUU AUC Ile AUA AUG Met	ACU ACC Thr ACA ACG	AAU Asn AAC AAA Lys AAG	AGU Ser AGC AGA Arg AGG	U C A G	
	G	GUU GUC Val GUA GUG	GCU GCC Ala GCA GCG	GAU Asp GAC GAA Glu GAG	GGU GGC Gly GGA GGG	U C A G	
						3rd letter	

Figure 4. The canonical genetic code

In any case, the standard genetic code is exceptionally robust: the chance of getting an amino acid with totally different biochemical properties with a single point mutation is low. The probability of reaching the same robustness with random permutations of codons is below 10^{-6} , yet there are many theoretical code variants that would be even more robust (Koonin and Novozhilov, 2009).

1.3.2 Non-standard genetic codes in nature

There is only one canonical genetic code, although, diverse small variations of it are spread across the tree of life. Until present, 32 genetic code variants have been entered into the NCBI Taxonomy database (Elzanowski and Ostell, 2020), with some of the genetic code reassignments happening many times independently, either in related or unrelated lineages (Cocquyt et al., 2010; de Koning et al., 2008; Inagaki et al., 1998; Ivanova et al., 2014; Lozupone et al., 2001; Pánek et al., 2017; Žihala and Eliáš, 2019). The derived genetic codes are neither equally spread in nature, nor are they equally spread functionally. We can observe hotspots of genetic code changes both in some codons, which have been reassigned many times independently, and in certain eukaryotic clades, where many departures from the canonical code occurred (Keeling, 2016; Knight et al., 2001a; Ling et al., 2015) (Fig. 5). Theories explaining unequal distribution of the codon reassignment are considered in the next chapter. Most genetic code changes occurred in the mitochondrial genomes (Bender et al., 2008; Sengupta et al., 2007; Sengupta and Higgs, 2015), since 13 of 32 known variants are confined to this organelle (Elzanowski and Ostell, 2020). It is worth noting that there is only a handful of cases of the genetic code change in the chloroplasts, which probably reflects a difference in gene expression and genome evolution of both organelles (Bender et al., 2008; Keeling, 2016).

In eukaryotes, most known changes in the nuclear genetic code happened in ciliates, all of which are stop codon reassignments (6 out of known 32 genetic code variants are confined to ciliates). Although hundreds of marine protozoa transcriptomes from Marine Microbial Transcriptome Sequencing Project (Keeling et al., 2014) have been systematically scanned, all three stop codon reassignment have been found only in two ciliate species with no case of genetic code change documented outside of ciliates (Heaphy et al., 2016; Swart et al., 2016).

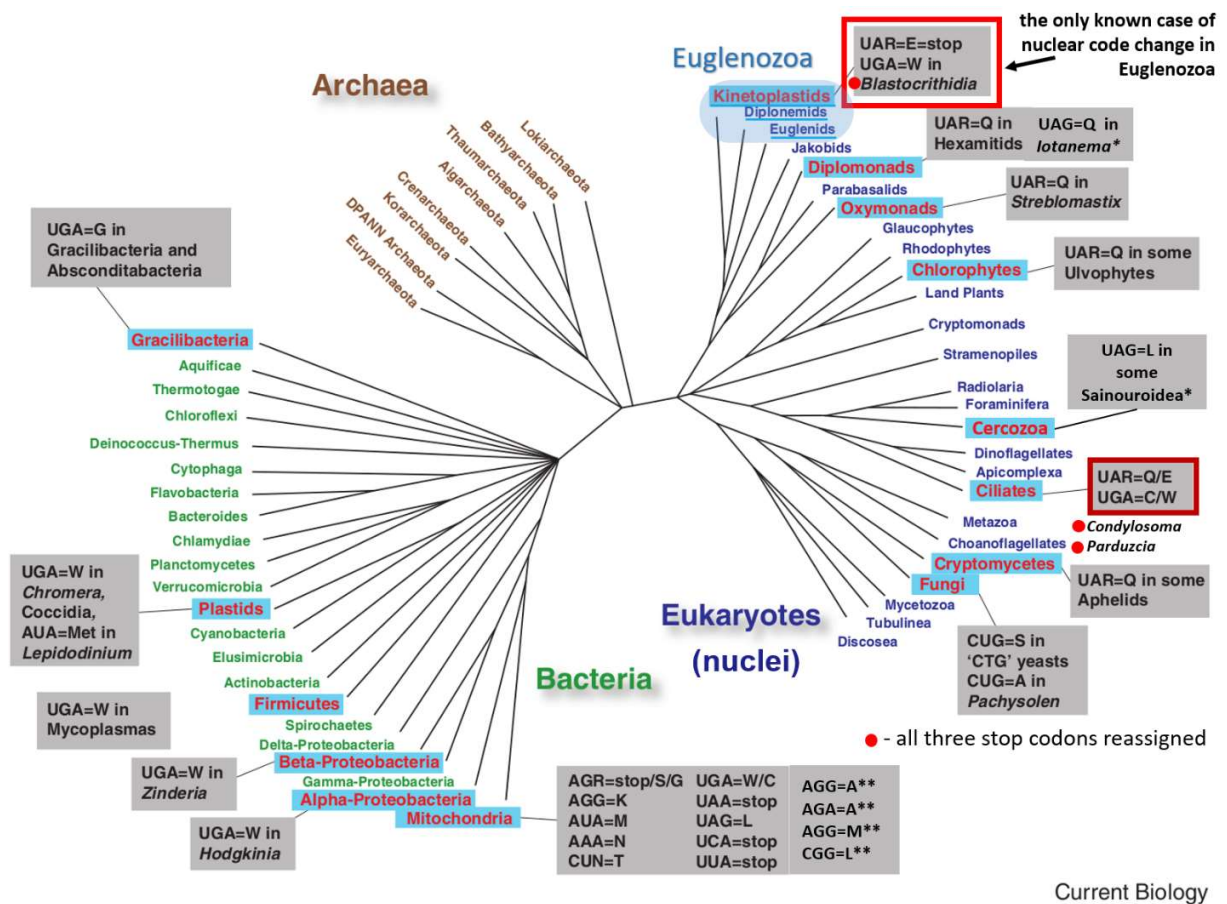


Figure 5. Genetic codes on the tree of life. Modified from Keeling (2016).

*, ** – new genetic codes, reported by Pánek et al., 2017 and Žihala and Eliáš, 2019, respectively.

A vast majority of changed genetic codes includes reassignment of one or more stop codons to amino acids (Elzanowski and Ostell, 2020). The most frequent change in mitochondria and bacteria alters the meaning of the UGA stop codon to tryptophan (Inagaki et al., 1998), while within eukaryotes the most frequent change is a reassignment of UAA and UAG (UAR) to glutamine (Keeling, 2016). Curiously, UAG and UAA are usually reassigned together to encode the same amino acid, while UGA is reassigned independently, though there are exceptions to this rule (Pánek et al., 2017). The UGA stop codon is also used to encode the non-canonical amino acid selenocysteine (Sec), the inclusion of which is guided by Sec-insertion sequence (SECIS) in an mRNA. This modification is known from bacteria, archaea, and eukaryotes, though only certain organisms within each group have it (Ling et al., 2015). Notably, the ciliate *Euplotes crassus* was proven to have both selenocysteine incorporation system and UGA reassigned to cysteine. The meaning of the UGA codon is determined by the presence or absence of SECIS in a particular mRNA (Turanov et al., 2009). Another example of a non-canonical amino acid

coded by a stop codon is the UAG-coded pyrrolysine (Pyl) found in archaea and bacteria (Rother and Krzycki, 2010). Unlike selenocysteine, incorporation of pyrrolysine is not guided by any specific sequence (Rother and Krzycki, 2010). Fascinatingly, some organisms use both Sec and Pyl, thus encoding all 22 natural amino acids (Ling et al., 2015).

1.3.3 Codon reassignment theories

There are two classical theories describing the mechanism of codon reassignment. First of them, the codon capture model, was formulated by Osawa and Jukes in 1989. In this model, due to a genetic drift or a GC-content pressure on the genome, some codons become rare and can even disappear. The next step is a change in the translation machinery, which loses the ability to recognize these specific codons. Subsequently, the lost codons reappear again due to a direction change of GC-content pressure or just by chance and are captured by near-cognate tRNA of another codon. Finally, natural selection moves the system towards a more efficient and specific recognition (Osawa and Jukes, 1989).

The tenets of the codon capture theory are supported by many documented cases when certain codons are strongly underrepresented or even absent from some genomes (Korkmaz et al., 2014; Nakabachi et al., 2006). Furthermore, this theory explains many known genetic code change cases. Why does a vast majority of genetic code changes involve the stop codons? It has been suggested that since when compared to codons specifying amino acid, stop codons are the rarest ones, they are more likely to disappear (Baranov et al., 2015; Knight et al., 2001a; Osawa et al., 1992).

Why do most genetic code changes happen in mitochondria (Knight et al., 2001b)? It is likely that the generally small size of the mitochondrial genomes which contain only a few genes is prone to the loss of some codons due the genetic drift (Baranov et al., 2015; Knight et al., 2001b; Swart et al., 2016). Moreover, many mitochondrial genomes are AT-rich or experienced AT-mutation pressure at some period of their evolution (Knight et al., 2001b; McCutcheon et al., 2009). These premises are particularly useful for explaining the main steps of codon reassignments exclusively by the mechanisms of neutral evolution (Jukes, 1996). However, each theory has weak points, and this applies also to the codon capture theory. It assumes that all the reassigned codons must somehow disappear (Osawa and Jukes, 1989), which is plausible for the stop codons in small mitochondrial or endosymbiont genomes but seems to be very unlikely in

the case of nuclear genetic code change and sense-to-sense reassignments (Baranov et al., 2015; Kollmar and Mühlhausen, 2017).

The second classical theory dealing with this question was formulated by Schultz and Yarus in 1994, who called it the theory of ambiguous intermediate. It postulates that mutations in tRNAs change their specificity and make them recognize other codons along with the main one. It is known that even the standard tRNAs can recognize not only their cognate codon, but also similar codons, albeit with lower efficiency. Modifications of the anticodon part or mutations/modifications of the other part of a tRNA can make recognition of the non-target codon(s) more efficient (Knight et al., 2001a; Schultz and Yarus, 1994). Counter-intuitively, this process does not necessarily lead to a lower fitness of the affected organism. Bender and colleagues showed that the reassignment of AUA from isoleucine to methionine, observed in most mitochondrial lineages, helps to cope with the oxidative stress due to the antioxidant function of methionine (Bender et al., 2008). Furthermore, some organisms are capable of regulated reassignment of certain codons, which allows them to alter their proteomes in response to changing conditions (Baranov et al., 2015). From the current perspective, both processes have their impact on codon reassignments and can represent two stages of this process in the same lineage. On one hand, the GC-content pressure can make some codons rare, and on the other hand, the mutations in corresponding tRNAs are less dreadful under such conditions.

Additionally, several other theories of genetic code change have been proposed. The theory of genome streamlining assumes that selection pressure on the size and complexity of a genome may remove redundant components, such as certain codons and corresponding tRNAs, or one of the two release factors in the case of stop codon loss in a prokaryote or prokaryote-derived organelle (Knight et al., 2001a). The theory of tRNA loss-driven codon reassignment suggests that tRNA or translation termination factor loss-of-function mutations or gene loss events may be the driving force for genetic code changes (Kollmar and Mühlhausen, 2017; Mühlhausen et al., 2016). Finally, the hypothesis of error resistance explains genetic code changes from the perspective of making it more error-prone (Knight et al., 2001a). Theories specifically explaining the stop codon reassignment are detailed in the next chapters.

1.3.4 Translation termination and stop codon reassignment

A vast majority of genetic code changes is represented by the reassignment of one or more stop codons (Ivanova et al., 2014; Keeling, 2016). Stop codons are believed to be an easy target for reassignment because they are the rarest codons, as there is only one stop codon for one gene (Baranov et al., 2015). Therefore, their reassignment is not so dramatic for the genome, as are the reassignments of the amino acid codons. It is probable that due to the GC-content bias or due to the genetic drift, to which small genomes are particularly susceptible, a certain kind of stop codons could become very rare or disappear, and then be captured by a near-cognate tRNA (Blanchet et al., 2014). Moreover, stop codons outside of the open reading frames are frequent, so most likely the gene which has lost its stop codon will find another one in a close proximity and the affected protein will be only several amino acids longer than the normal version, which may likely not dramatically affect its function (Baranov et al., 2015). However, there is an alternative point of view that the stop codon reassignment may be not the most frequent one, but the most visible change, as open reading frames disrupted by several in-frame stop codons are easy to notice, while for finding a sense-to-sense change a thorough bioinformatic analysis is required (Keeling, 2016).

To understand how stop codon reassignment happens, we need to consider what is known about the termination of protein synthesis in prokaryotes and eukaryotes. The current view postulates that the termination happens when the ribosome finishes translation of the coding sequence on a transcript and a stop codon enters the ribosomal A-site in the large ribosomal subunit (Dever and Green, 2012). In both prokaryotes and eukaryotes, this process is performed by two classes of polypeptide chain release factors (RFs): class I release factors (RF1 and RF2 in bacteria, aRF1/eRF1 in archaea/eukaryotes), which recognize the stop codon and release a nascent peptide chain, and class II release factors (RF3 in bacteria, aRF3/eRF3 in archaea/eukaryotes), which remove the class I RF from the ribosome after the peptide release (Dever and Green, 2012; Duarte et al., 2012).

Bacteria and bacteria-derived organelles have two canonical class I release factors: RF1, specific to UAA and UAG codons, and RF2, specific to UAA and UGA (Duarte et al., 2012; Scolnick et al., 1968). Archaea and eukaryotes have only one RF of class I, recognizing all three stop codons, labeled aRF1 or eRF1, correspondingly. The eukaryotic/archaeal and bacterial class I RFs are not homologous and have dissimilar protein structure (Kisselev, 2002). Furthermore,

eRF1 acts in a complex with eRF3 both *in vivo* and *in vitro*, while neither RF1 nor RF2 binds to RF3 in bacteria (Nakamura et al., 1996; Pel et al., 1998). All this led some authors to conclude that the LUCA used another mechanism of translation termination, maybe including special RNA instead of proteinaceous RFs (Baranov et al., 2015).

Mutations in RFs affect their efficiency and specificity for certain stop codons. This was shown both by site-directed mutagenesis experiments in model organisms, as well as by using naturally occurring RFs found in organisms with reassigned stop codons. In these experiments, several important conserved motifs and amino acid residues in eRF1 were identified, namely GTS, NIKS, YxCxxF, E55, and S70 (Blanchet et al., 2015; Conard et al., 2012; Eliseev et al., 2011; Frolova et al., 2002; Kolosov et al., 2005; Kryuchkova et al., 2013; Lekomtsev et al., 2007; Lozupone et al., 2001; Saito and Ito, 2015; Salas-Marco et al., 2006). Subsequent investigations of the structure of eukaryotic termination complex have confirmed that these motifs are indeed involved in the stop recognition centers of eRF1 (Brown et al., 2015; Matheisl et al., 2015).

When compared to elongation, translation termination is a slower and less efficient process (Baranov et al., 2015). Hence, especially the mutated RFs have a high chance to be dislodged by competing cognate or near-cognate aminoacylated tRNAs, which forces the system to read stop codons as sense. This so-called eRF1 mutation hypothesis has been proposed by Lozupone and colleagues (Lozupone et al., 2001) and reviewed by Alkalaeva and Mikhailova (Alkalaeva and Mikhailova, 2016). Moreover, the efficiency of termination strongly depends on the nucleotide context of a given stop codon, which differs in the cases of in-frame and genuine stop codons (Bonetti et al., 1995; Brown et al., 1990; Namy et al., 2001; Pavlov et al., 1998). This also explains some of the known cases of stop codon readthrough in organisms with standard stop codons, mostly used by viruses to produce different protein isoforms and as a regulatory system (Baranov et al., 2015; Dreher and Miller, 2006; Firth and Brierley, 2012).

It is important to mention that eukaryotes have a conserved nonsense-mediated decay (NMD) machinery, the function of which is to remove transcripts with in-frame stop codons resulting from nonsense mutations (Brognia and Wen, 2009). The presence of such a system seemingly contradicts the scenario, in which in-frame stop codons specify amino acids. Nevertheless, conserved NMD components are present in some species with reassigned stop codons, although it is not known whether they are active (Swart et al., 2016).

1.3.5 Organisms with all three stop codons reassigned

In 2016, three eukaryotic genera, namely the trypanosomatid *Blastocrithidia* and the ciliates *Parduzia* and *Condylostoma* were shown to bear genetic code with all three stop codons reassigned (Heaphy et al., 2016; Swart et al., 2016; Záhonová et al., 2016). In *Blastocrithidia* sp. UAR were predicted to encode glutamate, and UGA tryptophan, while UAR are also used as genuine stop codons. However, UAA is used more frequently as a genuine stop codon than UAG, although this conclusion was based on a rather small number of genes (Záhonová et al., 2016). In *Parduzia* sp. UAR are used to encode glutamate, while UGA encodes tryptophan, and UGA is used as a genuine stop codon (Swart et al., 2016). In *Condylostoma magnum* UAA and UAG encode glutamine, UGA codes for tryptophan, and all stop codons are used as genuine ones in a context-dependent manner (Heaphy et al., 2016; Swart et al., 2016). This finding was confirmed experimentally using mass spectrometry and ribosomal profiling (Swart et al., 2016). Both ciliates bear ambiguous genetic codes, which are read in a context-dependent manner: in-frame stop codons are read as amino acids, while in the end of the genes they are used as genuine stop codons. However, it is not the first known case of ambiguous stop codons: another ciliate *Blepharisma japonicum* (Eliseev et al., 2011) and archaeon *Methanosarcina barkeri* (James et al., 2001) are also known to read stop codons ambiguously.

All this raised a question about how translation termination works in such systems: they must have gained the ability to efficiently distinguish between in-frame and genuine stop codons, which have an identical triplet, but completely different sense. Two of the three above-described cases were found in ciliates, which had been already prominent for their stop codon reassignment. It was noticed that ciliates as a group have extremely short 3'-UTRs (20-40 nt), which means that their poly(A) tails start almost right after their genuine stop codons (Heaphy et al., 2016; Swart et al., 2016). Moreover, it was also noted that in the ciliate genes in-frame upstream stop codons are being avoided in close proximity to the genuine ones (Swart et al., 2016). Previously, it was shown that the poly(A)-binding protein (PABP) facilitates translation termination in eukaryotes (Ivanov et al., 2016). Therefore, it was hypothesized that the efficiency of translation termination in the considered species, and maybe in ciliates in general, is determined by the interaction of release factors with PABP (Heaphy et al., 2016; Swart et al., 2016) (Fig. 6). Other authors hypothesized that in these organisms most of the RF complexes are bound with PABP, thus the whole translation termination complex waits until the ribosome

arrives to the translation termination site, and there is consequently little chance that the RF will ever meet in-frame stop codons (Alkalaeva and Mikhailova, 2016).

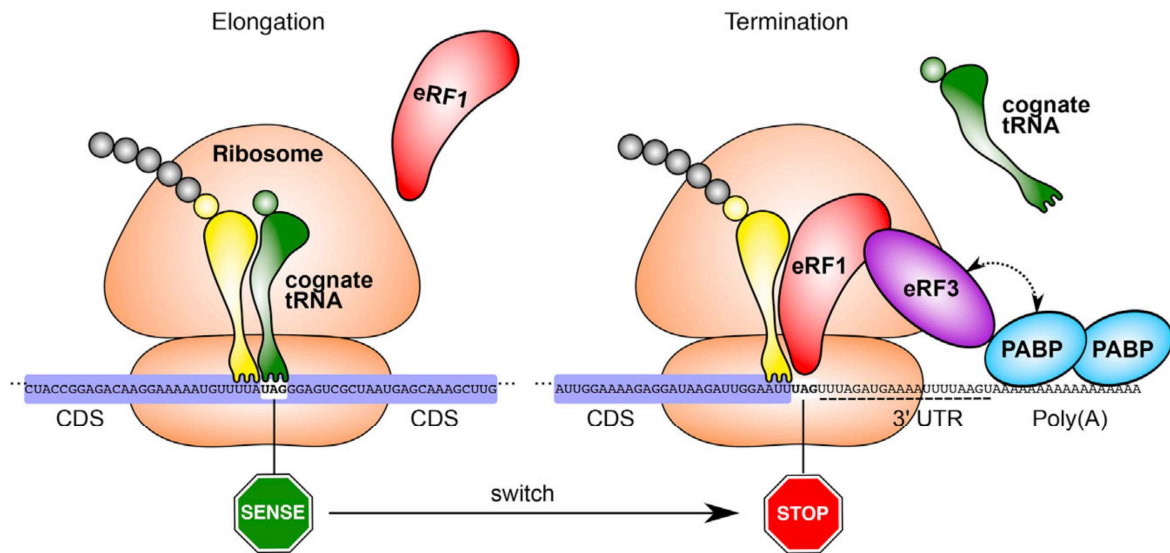


Figure 6. Model of context-dependent termination (Swart et al., 2016).

1.3.6 Genetic code in kinetoplastids

Blastocrithidia is the first known case of nuclear code reassignment among both Kinetoplastida and the whole Euglenozoa clade (Záhonová et al., 2016). However, trypanosomatid flagellates are known to use UGA to encode tryptophan in their mitochondrial (= kinetoplast) genome (De la Cruz et al., 1984). In trypanosomatids, all tRNAs are encoded exclusively in the nuclear genome, with a subset of tRNAs being imported into the mitochondrion (Hancock and Jahduk, 1990; Simpson et al., 1989; Tan et al., 2002). In the trypanosomatid nuclear genome, there is only one tRNA^{T_{rp}} with a CCA anticodon, which recognizes the standard tryptophan UGG codon but cannot decode UGA (Alfonzo et al., 1999; Shi et al., 1994). Therefore, tRNA^{T_{rp}}_{CCA} must be modified to recognize UGA, but this process must be confined to the mitochondrion, since it would otherwise result in the UGA stop codon readthrough all over the nuclear genome. This is achieved by C to U editing of the first position of tRNA^{T_{rp}}_{CCA} anticodon (CCA to UCA) in the mitochondrion by a deaminase (Alfonzo et al., 1999) (Fig. 7). It is worth noting that despite significant efforts this enzyme has not yet been identified. Notably, trypanosomatids use two distinct tryptophanyl-tRNA synthetases to charge tRNA^{T_{rp}} in the cytosol and the mitochondrion (Charrière et al., 2006).

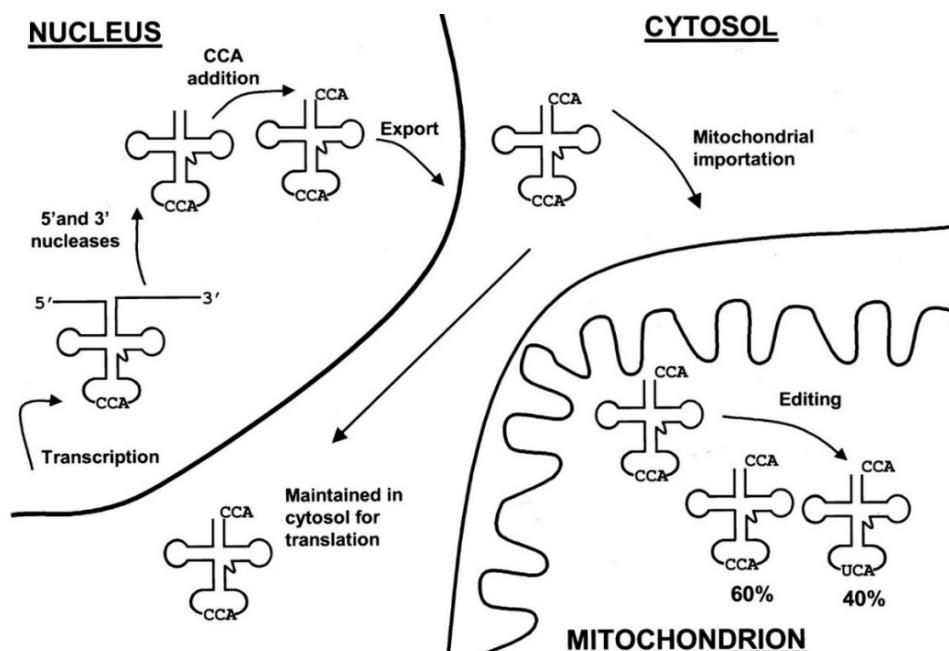


Figure 7. Trypanosomatid tRNA^{Trp} transcription, processing, and editing (Kapushoc et al., 2000).

Blastocrithidia uses UGA to encode tryptophan in the nucleus and bears corresponding mutation in the S70 position of eRF1, which is associated with UGA recognition (Eliseev et al., 2011; Záhonová et al., 2016). Remarkably, Swart and colleagues found specific tRNA genes for both reassigned UAR codons in the genome of *Condylostoma magnum*, but failed to find a tRNA-Trp gene, recognizing UGA stop codon, so tRNA^{Trp}_{CCA} in *C. magnum* may potentially undergo similar modification. However, RT-PCR of tRNA^{Trp}_{CCA} and consequential sequencing of the product did not show edited variants (Swart et al., 2016).

Nonetheless, nothing is known about the UAR reassignment in *Blastocrithidia*, and how its translation termination works. In the ciliates with ambiguous stop codons, the most plausible hypothesis explains the differential translation of in-frame and genuine stop codons by the proximity of the latter to poly(A) tails. As mentioned above, the ciliates are apparently exceptionally prone to stop codon reassignment, likely due to their very short poly(A) tails, which allows direct mutual interaction among the genuine stop codon, eRF complex and poly(A)-binding protein (Alkalaeva and Mikhailova, 2016; Heaphy et al., 2016; Swart et al., 2016). However, trypanosomatids are not generally known to have short 3'-UTRs:

Trypanosoma brucei has 3'-UTRs of mean length 587 nt (Benz et al., 2005), while their size in *T. cruzi* is 264 nt on average (Campos et al., 2008).

Pages 30-55 contain classified information (unpublished data) and are available only in the archived original of the graduation thesis deposited at the Faculty of Sciences, University of South Bohemia.

Pages 30-55 contain classified information (unpublished data) and are available only in the archived original of the graduation thesis deposited at the Faculty of Sciences, University of South Bohemia.

5 References

- Adl, S.M., Bass, D., Lane, C.E., Lukeš, J., Schoch, C.L., Smirnov, A., Agatha, S., Berney, C., Brown, M.W., Burki, F., Cárdenas, P., Čepička, I., Chistyakova, L., del Campo, J., Dunthorn, M., Edvardsen, B., Eglit, Y., Guillou, L., Hampl, V., Heiss, A.A., Hoppenrath, M., James, T.Y., Karnkowska, A., Karpov, S., Kim, E., Kolisko, M., Kudryavtsev, A., Lahr, D.J.G., Lara, E., Le Gall, L., Lynn, D.H., Mann, D.G., Massana, R., Mitchell, E.A.D., Morrow, C., Park, J.S., Pawlowski, J.W., Powell, M.J., Richter, D.J., Rueckert, S., Shadwick, L., Shimano, S., Spiegel, F.W., Torruella, G., Youssef, N., Zlatogursky, V., Zhang, Q., 2019. Revisions to the classification, nomenclature, and diversity of eukaryotes. *J. Eukaryot. Microbiol.* 66, 4–119.
- Adl, S.M., Simpson, A.G.B., Lane, C.E., Lukeš, J., Bass, D., Bowser, S.S., Brown, M.W., Burki, F., Dunthorn, M., Hampl, V., Heiss, A., Hoppenrath, M., Lara, E., le Gall, L., Lynn, D.H., McManus, H., Mitchell, E.A.D., Mozley-Stanridge, S.E., Parfrey, L.W., Pawlowski, J., Rueckert, S., Shadwick, L., Schoch, C.L., Smirnov, A., Spiegel, F.W., Shadwick, R.S., Shadwick, L., Schoch, C.L., Smirnov, A., Spiegel, F.W., Shadwick, R.S., Shadwick, L., Schoch, C.L., Smirnov, A., Spiegel, F.W., 2012. The revised classification of eukaryotes. *J. Eukaryot. Microbiol.* 59, 429–493.
- Alfonzo, J.D., Blanc, V., Estévez, A.M., Rubio, M.A., Simpson, L., 1999. C to U editing of the anticodon of imported mitochondrial tRNA(Trp) allows decoding of the UGA stop codon in *Leishmania tarentolae*. *EMBO J.* 18, 7056–7062.
- Alkalaeva, E., Mikhailova, T., 2016. Reassigning stop codons via translation termination: How a few eukaryotes broke the dogma. *BioEssays* 39, 1–6.
- Anderson, B.A., Wong, I.L.K., Baugh, L., Ramasamy, G., Myler, P.J., Beverley, S.M., 2013. Kinetoplastid-specific histone variant functions are conserved in *Leishmania major*. *Mol. Biochem. Parasitol.* 191, 53–57.
- Andrews, S., 2010. FastQC: a quality control tool for high throughput sequence data. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
- Ayub, M.J., Atwood, J., Nuccio, A., Tarleton, R., Levin, M.J., 2009. Proteomic analysis of the *Trypanosoma cruzi* ribosomal proteins. *Biochem. Biophys. Res. Commun.* 382, 30–34.
- Baejen, C., Torkler, P., Gressel, S., Essig, K., Söding, J., Cramer, P., 2014. Transcriptome Maps

- of mRNP Biogenesis Factors Define Pre-mRNA Recognition. *Mol. Cell* 55, 745–757.
- Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M.A., Pevzner, P.A., 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477.
- Baranov, P. V., Atkins, J.F., Yordanova, M.M., 2015. Augmented genetic decoding: global, local and temporal alterations of decoding processes and codon meaning. *Nat. Rev. Genet.* 16, 517–529.
- Barrett, M.P., Burchmore, R.J.S., Stich, A., Lazzari, J.O., Frasch, A.C., Cazzulo, J.J., Krishna, S., 2003. The trypanosomiasis. In: *Lancet*. Elsevier Limited, pp. 1469–1480.
- Bender, A., Hajieva, P., Moosmann, B., 2008. Adaptive antioxidant methionine accumulation in respiratory chain complexes explains the use of a deviant genetic code in mitochondria. *Proc. Natl. Acad. Sci. U. S. A.* 105, 16496–16501.
- Benz, C., Nilsson, D., Andersson, B., Clayton, C., Guilbride, D.L., 2005. Messenger RNA processing sites in *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* 143, 125–134.
- Berriman, M., Ghedin, E., Hertz-Fowler, C., Blandin, G., Renauld, H., Bartholomeu, D.C., Lennard, N.J., Caler, E., Hamlin, N.E., Haas, B., Böhme, U., Hannick, L., Aslett, M.A., Shallom, J., Marcello, L., Hou, L., Wickstead, B., Alsmark, U.C.M., Arrowsmith, C., Atkin, R.J., Barron, A.J., Bringaud, F., Brooks, K., Carrington, M., Cherevach, I., Chillingworth, T.-J., Churcher, C., Clark, L.N., Corton, C.H., Cronin, A., Davies, R.M., Doggett, J., Djikeng, A., Feldblyum, T., Field, M.C., Fraser, A., Goodhead, I., Hance, Z., Harper, D., Harris, B.R., Hauser, H., Hostetler, J., Ivens, A., Jagels, K., Johnson, D., Johnson, J., Jones, K., Kerhornou, A.X., Koo, H., Larke, N., Landfear, S., Larkin, C., Leech, V., Line, A., Lord, A., Macleod, A., Mooney, P.J., Moule, S., Martin, D.M.A., Morgan, G.W., Mungall, K., Norbertczak, H., Ormond, D., Pai, G., Peacock, C.S., Peterson, J., Quail, M.A., Rabbinowitsch, E., Rajandream, M.-A., Reitter, C., Salzberg, S.L., Sanders, M., Schobel, S., Sharp, S., Simmonds, M., Simpson, A.J., Tallon, L., Turner, C.M.R., Tait, A., Tivey, A.R., Van Aken, S., Walker, D., Wanless, D., Wang, S., White, B., White, O., Whitehead, S., Woodward, J., Wortman, J., Adams, M.D., Embley, T.M., Gull, K., Ullu, E., Barry, J.D., Fairlamb, A.H., Opperdoes, F., Barrell, B.G., Donelson, J.E., Hall, N., Fraser, C.M., Melville, S.E., El-Sayed, N.M., 2005. The genome of the African trypanosome

- Trypanosoma brucei*. Science 309, 416–422.
- Bidou, L., Hatin, I., Perez, N., Allamand, V., Panthier, J.-J., Rousset, J.-P., 2004. Premature stop codons involved in muscular dystrophies show a broad spectrum of readthrough efficiencies in response to gentamicin treatment. Gene Ther. 11, 619–27.
- Blanchet, S., Cornu, D., Argentini, M., Namy, O., 2014. New insights into the incorporation of natural suppressor tRNAs at stop codons in *Saccharomyces cerevisiae*. Nucleic Acids Res. 42, 10061–10072.
- Blanchet, S., Rowe, M., Von der Haar, T., Fabret, C., Demais, S., Howard, M.J., Namy, O., 2015. New insights into stop codon recognition by eRF1. Nucleic Acids Res. 43, 3298–3308.
- Bonetti, B., Fu, L., Moon, J., Bedwell, D.M., 1995. The efficiency of translation termination is determined by a synergistic interplay between upstream and downstream sequences in *Saccharomyces cerevisiae*. J. Mol. Biol. 251, 334–345.
- Brogna, S., Wen, J., 2009. Nonsense-mediated mRNA decay (NMD) mechanisms. Nat. Struct. Mol. Biol. 16, 107–113.
- Brown, A., Shao, S., Murray, J., Hegde, R.S., Ramakrishnan, V., 2015. Structural basis for stop codon recognition in eukaryotes. Nature 524, 493–496.
- Brown, C.M., Stockwell, P.A., Trotman, C.N., Tate, W.P., 1990. Sequence analysis suggests that tetra-nucleotides signal the termination of protein synthesis in eukaryotes. Nucleic Acids Res. 18, 6339–6345.
- Bushnell, B., 2017. BBMap. <https://sourceforge.net/projects/bbmap/>.
- Caicedo, A.M., Gallego, G., Muñoz, J.E., Suárez, H., Torres, G.A., Carvajal, H., Carvajal, F.C. De, Posso, A.M., Maslov, D., Montoya-Lerma, J., 2011. Morphological and molecular description of *Blastocrithidia cyrtomeni* sp. nov. (Kinetoplastea: Trypanosomatidae) associated with *Cyrtomenus bergi* Froeschner (Hemiptera: Cydnidae) from Colombia. Mem. Inst. Oswaldo Cruz 106, 301–307.
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: Architecture and applications. BMC Bioinformatics 10, 421.
- Campos, P.C., Bartholomeu, D.C., DaRocha, W.D., Cerqueira, G.C., Teixeira, S.M.R., 2008. Sequences involved in mRNA processing in *Trypanosoma cruzi*. Int. J. Parasitol. 38, 1383–1389.

- Cavalier-Smith, T., 1981. Eukaryote kingdoms: seven or nine? *Biosystems*. 14, 461–481.
- Charrière, F., Helgadóttir, S., Horn, E.K., Söll, D., Schneider, A., 2006. Dual targeting of a single tRNA^{Trp} requires two different tryptophanyl-tRNA synthesis in *Trypanosoma brucei*. *Proc. Natl. Acad. Sci. U. S. A.* 103, 6847–6852.
- Chung, H.M., Shea, C., Fields, S., Taub, R.N., Van der Ploeg, L.H., Tse, D.B., 1990. Architectural organization in the interphase nucleus of the protozoan *Trypanosoma brucei*: location of telomeres and mini-chromosomes. *EMBO J.* 9, 2611–2619.
- Clayton, C., Shapira, M., 2007. Post-transcriptional regulation of gene expression in trypanosomes and leishmanias. *Mol. Biochem. Parasitol.* 156, 93–101.
- Clayton, C.E., 2016. Gene expression in Kinetoplastids. *Curr. Opin. Microbiol.* 32, 46–51.
- Cocquyt, E., Gile, G.H., Leliaert, F., Verbruggen, H., Keeling, P.J., De Clerck, O., 2010. Complex phylogenetic distribution of a non-canonical genetic code in green algae. *BMC Evol. Biol.* 10, 327.
- Conard, S.E., Buckley, J., Dang, M., Bedwell, G.J., Carter, R.L., Khass, M., Bedwell, D.M., 2012. Identification of eRF1 residues that play critical and complementary roles in stop codon recognition. *RNA* 18, 1210–1221.
- Crick, F.H.C., 1968. The Origin of the Genetic Code. *J. Mol. Biol.* 38, 367–379.
- da Costa Lima, T.D., Moura, D.M.N., Reis, C.R.S., Vasconcelos, J.R.C., Ellis, L., Carrington, M., Figueiredo, R.C.B.Q., de Melo Neto, O.P., 2010. Functional characterization of three *Leishmania* poly(A) binding protein homologues with distinct binding properties to RNA and protein partners. *Eukaryot. Cell* 9, 1484–1494.
- Darty, K., Denise, A., Ponty, Y., 2009. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics* 25, 1974–1975.
- de Koning, A.P., Noble, G.P., Heiss, A.A., Wong, J., Keeling, P.J., 2008. Environmental PCR survey to determine the distribution of a non-canonical genetic code in uncultivable oxymonads. *Environ. Microbiol.* 10, 65–74.
- De la Cruz, V.F., Neckelmann, N., Simpson, L., 1984. Sequences of six genes and several open reading frames in the kinetoplast maxicircle DNA of *Leishmania tarentolae*. *J. Biol. Chem.* 259, 15136–15147.
- Delhi, P., Queiroz, R., Inchaustegui, D., Carrington, M., Clayton, C., 2011. Is there a classical nonsense-mediated decay pathway in trypanosomes? *PLoS One* 6, e25112.

- Deschamps, P., Lara, E., Marande, W., López-García, P., Ekelund, F., Moreira, D., 2011. Phylogenomic analysis of kinetoplastids supports that trypanosomatids arose from within bodonids. *Mol. Biol. Evol.* 28, 53–58.
- Dever, T.E., Green, R., 2012. The elongation, termination, and recycling phases of translation in eukaryotes. *Cold Spring Harb. Perspect. Biol.* 4, 1–16.
- Dreher, T.W., Miller, W.A., 2006. Translational control in positive strand RNA plant viruses. *Virology* 344, 185–197.
- Duarte, I., Nabuurs, S.B., Magno, R., Huynen, M., 2012. Evolution and diversification of the organellar release factor family. *Mol. Biol. Evol.* 29, 3497–3512.
- Edgar, R.C., 2004. MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5, 113.
- Ehara, M., Inagaki, Y., Watanabe, K.I., Ohama, T., 2000. Phylogenetic analysis of diatom *coxI* genes and implications of a fluctuating GC content on mitochondrial genetic code evolution. *Curr. Genet.* 37, 29–33.
- Eliseev, B., Kryuchkova, P., Alkalaeva, E., Frolova, L., 2011. A single amino acid change of translation termination factor eRF1 switches between bipotent and omnipotent stop-codon specificity. *Nucleic Acids Res.* 39, 599–608.
- Elzanowski, A., Ostell, J., 2020. The genetic codes [WWW Document]. URL <https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi> (accessed 1.22.20).
- Fiebig, M., Gluenz, E., Carrington, M., Kelly, S., 2014. SLaP mapper: a webserver for identifying and quantifying spliced-leader addition and polyadenylation site usage in kinetoplastid genomes. *Mol. Biochem. Parasitol.* 196, 71–4.
- Firth, A.E., Brierley, I., 2012. Non-canonical translation in RNA viruses. *J. Gen. Virol.* 93, 1385–1409.
- Flegontova, O., Flegontov, P., Malviya, S., Audic, S., Wincker, P., de Vargas, C., Bowler, C., Lukeš, J., Horák, A., 2016. Extreme diversity of diplomemid eukaryotes in the ocean. *Curr. Biol.* 26, 3060–3065.
- Frolov, A.O., Malysheva, M.N., Ganyukova, A.I., Yurchenko, V., Kostygov, A.Y., 2017. Life cycle of *Blastocrithidia papi* sp. n. (Kinetoplastea, Trypanosomatidae) in *Pyrrhocoris apterus* (Hemiptera, Pyrrhocoridae). *Eur. J. Protistol.* 57, 85–98.
- Frolova, L., Seit-Nebi, A., Kisselev, L., 2002. Highly conserved NIKS tetrapeptide is

- functionally essential in eukaryotic translation termination factor eRF1. *RNA* 8, 129–36.
- Gao, H., Ayub, M.J., Levin, M.J., Frank, J., 2005. The structure of the 80S ribosome from *Trypanosoma cruzi* reveals unique rRNA components. *Proc. Natl. Acad. Sci. U. S. A.* 102, 10206–10211.
- Gawryluk, R.M.R., del Campo, J., Okamoto, N., Strassert, J.F.H., Lukeš, J., Richards, T.A., Worden, A.Z., Santoro, A.E., Keeling, P.J., 2016. Morphological identification and single-cell genomics of marine diplomonads. *Curr. Biol.* 26, 3053–2059.
- Ghedini, E., Bringaud, F., Peterson, J., Myler, P., Berriman, M., Ivens, A., Andersson, B., Bontempi, E., Eisen, J., Angiuoli, S., Wanless, D., Von Arx, A., Murphy, L., Lennard, N., Salzberg, S., Adams, M.D., White, O., Hall, N., Stuart, K., Fraser, C.M., El-Sayed, N.M., 2004. Gene synteny and evolution of genome architecture in trypanosomatids. *Mol. Biochem. Parasitol.* 134, 183–191.
- Gurevich, A., Saveliev, V., Vyahhi, N., Tesler, G., 2013. QUASt: Quality assessment tool for genome assemblies. *Bioinformatics* 29, 1072–1075.
- Hancock, K., Jahduk, S.L., 1990. The mitochondrial tRNAs of *Trypanosoma brucei* are nuclear encoded. *J. Biol. Chem.* 265, 19208–19215.
- Hashimi, H., Čičová, Z., Novotná, L., Yan-Zi, W.E.N., Lukeš, J., 2009. Kinetoplastid guide RNA biogenesis is dependent on subunits of the mitochondrial RNA binding complex 1 and mitochondrial RNA polymerase. *RNA* 15, 588–599.
- Heaphy, S.M., Mariotti, M., Gladyshev, V.N., Atkins, J.F., Baranov, P. V., 2016. Novel ciliate genetic code variants including the reassignment of all three stop codons to sense codons in *Condylostoma magnum*. *Mol. Biol. Evol.* 33, 2885–2889.
- Hirsh, D., 1971. Tryptophan transfer RNA as the UGA suppressor. *J. Mol. Biol.* 58, 439–458.
- Hirsh, D., Gold, L., 1971. Translation of the UGA triplet in vitro by tryptophan transfer RNA's. *J. Mol. Biol.* 58, 459–468.
- Howard, M.T., Shirts, B.H., Petros, L.M., Flanigan, K.M., Gesteland, R.F., Atkins, J.F., 2000. Sequence specificity of aminoglycoside-induced stop codon readthrough: Potential implications for treatment of Duchenne muscular dystrophy. *Ann. Neurol.* 48, 164–169.
- Inagaki, Y., Ehara, M., Watanabe, K.I., Hayashi-Ishimaru, Y., Ohama, T., 1998. Directionally evolving genetic code: The UGA codon from stop to tryptophan in mitochondria. *J. Mol. Evol.* 47, 378–384.

- Ivanov, A., Mikhailova, T., Eliseev, B., Yeramala, L., Sokolova, E., Susorov, D., Shuvalov, A., Schaffitzel, C., Alkalaeva, E., 2016. PABP enhances release factor recruitment and stop codon recognition during translation termination. *Nucleic Acids Res.* 44, 7766–7776.
- Ivanova, N.N., Schwientek, P., Tripp, H.J., Rinke, C., Pati, A., Huntemann, M., Visel, A., Woyke, T., Kyrpides, N.C., Rubin, E.M., 2014. Stop codon reassignments in the wild. *Science* 344, 909–913.
- James, C.M., Ferguson, T.K., Leykam, J.F., Krzycki, J.A., 2001. The amber codon in the gene encoding the monomethylamine methyltransferase isolated from *Methanosarcina barkeri* is translated as a sense codon. *J. Biol. Chem.* 276, 34252–34258.
- Jaskowska, E., Butler, C., Preston, G., Kelly, S., 2015. *Phytomonas*: trypanosomatids adapted to plant environments. *PLoS Pathog.* 11, 1–17.
- Jukes, T.H., 1996. Neutral changes and modifications of the genetic code. *Theor. Popul. Biol.* 49, 143–145.
- Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780.
- Keeling, P.J., 2009. Chromalveolates and the evolution of plastids by secondary endosymbiosis. *J. Eukaryot. Microbiol.* 56, 1–8.
- Keeling, P.J., 2016. Genomics: evolution of the genetic code. *Curr. Biol.* 26, R851–R853.
- Keeling, P.J., Burki, F., Wilcox, H.M., Allam, B., Allen, E.E., Amaral-Zettler, L.A., Armbrust, E.V., Archibald, J.M., Bharti, A.K., Bell, C.J., Beszteri, B., Bidle, K.D., Cameron, C.T., Campbell, L., Caron, D.A., Cattolico, R.A., Collier, J.L., Coyne, K., Davy, S.K., Deschamps, P., Dyhrman, S.T., Edvardsen, B., Gates, R.D., Gobler, C.J., Greenwood, S.J., Guida, S.M., Jacobi, J.L., Jakobsen, K.S., James, E.R., Jenkins, B., John, U., Johnson, M.D., Juhl, A.R., Kamp, A., Katz, L.A., Kiene, R., Kudryavtsev, A., Leander, B.S., Lin, S., Lovejoy, C., Lynn, D., Marchetti, A., McManus, G., Nedelcu, A.M., Menden-Deuer, S., Miceli, C., Mock, T., Montresor, M., Moran, M.A., Murray, S., Nadathur, G., Nagai, S., Ngam, P.B., Palenik, B., Pawlowski, J., Petroni, G., Piganeau, G., Posewitz, M.C., Rengefors, K., Romano, G., Rumpho, M.E., Ryneerson, T., Schilling, K.B., Schroeder, D.C., Simpson, A.G.B., Slamovits, C.H., Smith, D.R., Smith, G.J., Smith, S.R., Sosik, H.M., Stief, P., Theriot, E., Twary, S.N., Umale, P.E., Vaultot, D., Wawrik, B., Wheeler, G.L., Wilson, W.H., Xu, Y., Zingone, A., Worden, A.Z., 2014. The Marine Microbial

- Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol.* 12, e1001889.
- Kelly, S., Kramer, S., Schwede, A., Maini, P.K., Gull, K., Carrington, M., 2012. Genome organization is a major component of gene expression control in response to stress and during the cell division cycle in trypanosomes. *Open Biol.* 2, 120033.
- Kini, H.K., Silverman, I.M., Ji, X., Gregory, B.D., Liebhaber, S.A., 2016. Cytoplasmic poly(A) binding protein-1 binds to genomically encoded sequences within mammalian mRNAs. *RNA* 22, 61–74.
- Kisselev, L., 2002. Polypeptide release factors in prokaryotes and eukaryotes: same function, different structure. *Structure* 10, 8–9.
- Knight, R.D., Freeland, S.J., Landweber, L.F., 2001a. Rewiring the keyboard: evolvability of the genetic code. *Nat. Rev. Genet.* 2, 49–58.
- Knight, R.D., Landweber, L.F., Yarus, M., 2001b. How mitochondria redefine the code. *J. Mol. Evol.* 53, 299–313.
- Kolev, N.G., Franklin, J.B., Carmi, S., Shi, H., Michaeli, S., Tschudi, C., 2010. The transcriptome of the human pathogen *Trypanosoma brucei* at single-nucleotide resolution. *PLoS Pathog.* 6, e1001090.
- Kollmar, M., Mühlhausen, S., 2017. Nuclear codon reassignments in the genomics era and mechanisms behind their evolution. *BioEssays* 39, 1–12.
- Koloso, P., Frolova, L., Seit-Nebi, A., Dubovaya, V., Kononenko, A., Oparina, N., Justesen, J., Efimov, A., Kisselev, L., 2005. Invariant amino acids essential for decoding function of polypeptide release factor eRF1. *Nucleic Acids Res.* 33, 6418–25.
- Koonin, E. V., 2009. The Origin at 150: is a new evolutionary synthesis in sight? *Trends Genet.* 25, 473–475.
- Koonin, E. V., 2017. Frozen accident pushing 50: stereochemistry, expansion, and chance in the evolution of the genetic code. *Life* 7, e22.
- Koonin, E. V., Novozhilov, A.S., 2009. Origin and evolution of the genetic code: The universal enigma. *IUBMB Life* 61, 99–111.
- Korkmaz, G., Holm, M., Wiens, T., Sanyal, S., 2014. Comprehensive analysis of stop codon usage in bacteria and its correlation with release factor abundance. *J. Biol. Chem.* 289,

30334–30342.

- Kozak, M., 1983. Comparison of initiation of protein synthesis in procaryotes, eucaryotes, and organelles. *Microbiol. Rev.* 47, 1–45.
- Krause, M., Hirsh, D., Cech, T.R., Borst, P., Blumenthal, T., Agabian, N., 1987. A trans-spliced leader sequence on actin mRNA in *C. elegans*. *Cell* 49, 753–761.
- Kryuchkova, P., Grishin, A., Eliseev, B., Karyagina, A., Frolova, L., Alkalaeva, E., 2013. Two-step model of stop codon recognition by eukaryotic release factor eRF1. *Nucleic Acids Res.* 41, 4573–4586.
- Lai, D.H., Wang, Q.P., Li, Z., Luckins, A.G., Reid, S.A., Lun, Z.R., 2010. Investigations into human serum sensitivity expressed by stocks of *Trypanosoma brucei evansi*. *Int. J. Parasitol.* 40, 705–710.
- Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359.
- Larsson, A., 2014. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* 30, 3276–8.
- Lasda, E.L., Blumenthal, T., 2011. Trans-splicing. *Wiley Interdiscip. Rev. RNA* 2, 417–434.
- Laslett, D., Canback, B., 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* 32, 11–16.
- Leander, B.S., Triemer, R.E., Farmer, M.A., 2001. Character evolution in heterotrophic euglenids. *Eur. J. Protistol.* 37, 337–356.
- LeBowitz, J.H., Smith, H.Q., Rusche, L., Beverley, S.M., 1993. Coupling of poly(A) site selection and trans-splicing in *Leishmania*. *Genes Dev.* 7, 996–1007.
- Lekomtsev, S., Kolosov, P., Bidou, L., Frolova, L., Rousset, J.-P., Kisselev, L., 2007. Different modes of stop codon restriction by the *Stylonychia* and *Paramecium* eRF1 translation termination factors. *Proc. Natl. Acad. Sci. U. S. A.* 104, 10824–10829.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079.
- Liang, X., Haritan, A., Uliel, S., Michaeli, S., 2003. *trans* and *cis* splicing in trypanosomatids: mechanism, factors, and regulation. *Eukaryot. Cell* 2, 830–40.
- Liebermeister, W., Noor, E., Flamholz, A., Davidi, D., Bernhardt, J., Milo, R., 2014. Visual

- account of protein investment in cellular functions. *Proc. Natl. Acad. Sci. U. S. A.* 111, 8488–8493.
- Ling, J., O’Donoghue, P., Söll, D., 2015. Genetic code flexibility in microorganisms: novel mechanisms and impact on physiology. *Nat. Rev. Microbiol.* 13, 707–721.
- Loughran, G., Chou, M.Y., Ivanov, I.P., Jungreis, I., Kellis, M., Kiran, A.M., Baranov, P. V., Atkins, J.F., 2014. Evidence of efficient stop codon readthrough in four mammalian genes. *Nucleic Acids Res.* 42, 8928–8938.
- Lowe, T.M., Chan, P.P., 2016. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res.* 44, W54–W57.
- Lozupone, C.A., Knight, R.D., Landweber, L.F., 2001. The molecular basis of nuclear genetic code change in ciliates. *Curr. Biol.* 11, 65–74.
- Lukeš, J., Archibald, J.M., Keeling, P.J., Doolittle, W.F., Gray, M.W., 2011. How a neutral evolutionary ratchet can build cellular complexity. *IUBMB Life* 63, 528–537.
- Lukeš, J., Guilbride, D.L., Votýpka, J., Zíková, A., Benne, R., Englund, P.T., 2002. Kinetoplast DNA network: evolution of an improbable structure. *Eukaryot. Cell* 1, 495–502.
- Lukeš, J., Leander, B.S., Keeling, P.J., 2009. Cascades of convergent evolution: the corresponding evolutionary histories of euglenozoans and dinoflagellates. *Proc. Natl. Acad. Sci. U. S. A.* 106 Suppl, 9963–9970.
- Lukeš, J., Skalický, T., Týč, J., Votýpka, J., Yurchenko, V., 2014. Evolution of parasitism in kinetoplastid flagellates. *Mol. Biochem. Parasitol.* 195, 115–122.
- Lukeš, J., Wheeler, R., Jirsová, D., David, V., Archibald, J.M., 2018. Massive mitochondrial DNA content in diplomonid and kinetoplastid protists. *IUBMB Life* 70, 1267–1274.
- Mair, G., Shi, H., Li, H., Djikeng, A., Aviles, H.O., Bishop, J.R., Falcone, F.H., Gavrilescu, C., Montgomery, J.L., Santori, M.I., Stern, L.S., Wang, Z., Ullu, E., Tschudi, C., 2000. A new twist in trypanosome RNA metabolism: *cis*-splicing of pre-mRNA. *RNA* 6, 163–169.
- Manuvakhova, M., Keeling, K., Bedwell, D.M., 2000. Aminoglycoside antibiotics mediate context-dependent suppression of termination codons in a mammalian translation system. *RNA* 6, 1044–1055.
- Maree, J.P., Patterton, H.G., 2014. The epigenome of *Trypanosoma brucei*: A regulatory interface to an unconventional transcriptional machine. *Biochim. Biophys. Acta* 1839, 743–750.

- Martínez-Calvillo, S., Nguyen, D., Stuart, K., Myler, P.J., 2004. Transcription initiation and termination on *Leishmania major* chromosome 3. *Eukaryot. Cell* 3, 506–517.
- Martínez-Calvillo, S., Yan, S., Nguyen, D., Fox, M., Stuart, K., Myler, P.J., 2003. Transcription of *Leishmania major* Friedlin chromosome 1 initiates in both directions within a single region. *Mol. Cell* 11, 1291–1299.
- Maslov, D.A., Votýpka, J., Yurchenko, V., Lukeš, J., 2013. Diversity and phylogeny of insect trypanosomatids: All that is hidden shall be revealed. *Trends Parasitol.* 29, 43–52.
- Matheisl, S., Berninghausen, O., Becker, T., Beckmann, R., 2015. Structure of a human translation termination complex. *Nucleic Acids Res.* 43, 8615–8626.
- McCutcheon, J.P., McDonald, B.R., Moran, N.A., Sato, T., Nagasu, T., 2009. Origin of an alternative genetic code in the extremely small and GC-rich genome of a bacterial symbiont. *PLoS Genet.* 5, e1000565.
- Merzlyak, E., Yurchenko, V., Kolesnikov, A.A., Alexandrov, K., Podlipaev, S.A., Maslov, D.A., 2001. Diversity and phylogeny of insect trypanosomatids based on small subunit rRNA genes: polyphyly of *Leptomonas* and *Blastocrithidia*. *J. Eukaryot. Microbiol.* 48, 161–169.
- Milanowski, R., Gumińska, N., Karnkowska, A., Ishikawa, T., Zakryś, B., 2016. Intermediate introns in nuclear genes of euglenids – are they a distinct type? *BMC Evol. Biol.* 16, 49.
- Mirarab, S., Nguyen, N., Guo, S., Wang, L.S., Kim, J., Warnow, T., 2015. PASTA: Ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. *J. Comput. Biol.* 22, 377–386.
- Mühlhausen, S., Findeisen, P., Plessmann, U., Urlaub, H., Kollmar, M., 2016. A novel nuclear genetic code alteration in yeasts and the evolution of codon reassignment in eukaryotes. *Genome Res.* 26, 945–55.
- Myler, P.J., Audleman, L., Devos, T., Hixson, G., Kiser, P., Lemley, C., Magness, C., Rickel, E., Sisk, E., Sunkin, S., Swartzell, S., Westlake, T., Bastien, P., Fu, G., Ivens, A., Stuart, K., 1999. *Leishmania major* Friedlin chromosome 1 has an unusual distribution of protein-coding genes. *Proc. Natl. Acad. Sci. U. S. A.* 96, 2902–2906.
- Myler, P.J., Fasel, N. (Eds.), 2008. *Leishmania: after the genome*. Caister Academic Press.
- Nakabachi, A., Yamashita, A., Toh, H., Ishikawa, H., Dunbar, H.E., Moran, N.A., Hattori, M., 2006. The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. *Science* 314, 267.

- Nakamura, Y., Ito, K., Isaksson, L.A., 1996. Emerging understanding of translation termination. *Cell* 87, 147–150.
- Namy, O., Hatin, I., Rousset, J.P., 2001. Impact of the six nucleotides downstream of the stop codon on translation termination. *EMBO Rep.* 2, 787–793.
- Nascimento, J. de F., Kelly, S., Sunter, J., Carrington, M., 2018. Codon choice directs constitutive mRNA levels in trypanosomes. *Elife* 7, e32467.
- Nguyen, L.T., Schmidt, H.A., Von Haeseler, A., Minh, B.Q., 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274.
- Osawa, S., Jukes, T.H., 1989. Codon reassignment (codon capture) in evolution. *J. Mol. Evol.* 28, 271–278.
- Osawa, S., Jukes, T.H., Watanabe, K., Muto, A., 1992. Recent evidence for evolution of the genetic code. *Microbiol. Rev.* 56, 229–264.
- Palenchar, J.B., Bellofatto, V., 2006. Gene transcription in trypanosomes. *Mol. Biochem. Parasitol.* 146, 135–141.
- Pánek, T., Žihala, D., Sokol, M., Derelle, R., Klimeš, V., Hradilová, M., Zadrobílková, E., Susko, E., Roger, A.J., Čepička, I., Eliáš, M., 2017. Nuclear genetic codes with a different meaning of the UAG and the UAA codon. *BMC Biol.* 15, 8.
- Panigrahi, A.K., Ogata, Y., Zíková, A., Anupama, A., Dalley, R.A., Acestor, N., Myler, P.J., Stuart, K.D., 2009. A comprehensive analysis of *Trypanosoma brucei* mitochondrial proteome. *Proteomics* 9, 434–450.
- Pavlov, M.Y., Freistoffer, D. V., Dinibas, V., MacDougall, J., Buckingham, R.H., Ehrenberg, M., 1998. A direct estimation of the context effect on the efficiency of termination. *J. Mol. Biol.* 284, 579–590.
- Peacock, C.S., Seeger, K., Harris, D., Murphy, L., Ruiz, J.C., Quail, M.A., Peters, N., Adlem, E., Tivey, A., Aslett, M., Kerhornou, A., Ivens, A., Fraser, A., Rajandream, M.A., Carver, T., Norbertczak, H., Chillingworth, T., Hance, Z., Jagels, K., Moule, S., Ormond, D., Rutter, S., Squares, R., Whitehead, S., Rabinowitsch, E., Arrowsmith, C., White, B., Thurston, S., Bringaud, F., Baldauf, S.L., Faulconbridge, A., Jeffares, D., Depledge, D.P., Oyola, S.O., Hilley, J.D., Brito, L.O., Tosi, L.R.O., Barrell, B., Cruz, A.K., Mottram, J.C., Smith, D.F., Berriman, M., 2007. Comparative genomic analysis of three *Leishmania* species that cause

- diverse human disease. *Nat. Genet.* 39, 839–847.
- Pechmann, S., Frydman, J., 2013. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nat. Struct. Mol. Biol.* 20, 237–43.
- Peikert, C.D., Mani, J., Morgenstern, M., Käser, S., Knapp, B., Wenger, C., Harsman, A., Oeljeklaus, S., Schneider, A., Warscheid, B., 2017. Charting organellar importomes by quantitative mass spectrometry. *Nat. Commun.* 8, 15272.
- Pel, H.J., Moffat, J.G., Ito, K., Nakamura, Y., Tate, W.P., 1998. *Escherichia coli* release factor 3: resolving the paradox of a typical G protein structure and atypical function with guanine nucleotides. *RNA* 4, 47–54.
- Povelones, M.L., 2014. Beyond replication: division and segregation of mitochondrial DNA in kinetoplastids. *Mol. Biochem. Parasitol.* 196, 53–60.
- Preusser, C., Rossbach, O., Hung, L.-H., Li, D., Bindereif, A., 2014. Genome-wide RNA-binding analysis of the trypanosome U1 snRNP proteins U1C and U1-70K reveals cis/trans-spliceosomal network. *Nucleic Acids Res.* 42, 6603–6615.
- Prokopchuk, G., Tashyreva, D., Yabuki, A., Horák, A., Masařová, P., Lukeš, J., 2019. Morphological, ultrastructural, motility and evolutionary characterization of two new Hemistasiidae species. *Protist* 170, 259–282.
- Ramrath, D.J.F., Niemann, M., Leibundgut, M., Bieri, P., Prange, C., Horn, E.K., Leitner, A., Boehringer, D., Schneider, A., Ban, N., 2018. Evolutionary shift toward protein-based architecture in trypanosomal mitochondrial ribosomes. *Science* 362, eaau7735.
- Read, L.K., Lukeš, J., Hashimi, H., 2015. Trypanosome RNA editing: the complexity of getting U in and taking U out. *Wiley Interdiscip. Rev. RNA* 7, 33–51.
- Reuveni, S., Ehrenberg, M., Paulsson, J., 2017. Ribosomes are optimized for autocatalytic production. *Nature* 547, 293–297.
- Reynolds, D., Cliffe, L., Förstner, K.U., Hon, C.C., Siegel, T.N., Sabatini, R., 2014. Regulation of transcription termination by glucosylated hydroxymethyluracil, base J, in *Leishmania major* and *Trypanosoma brucei*. *Nucleic Acids Res.* 42, 9717–9729.
- Rother, M., Krzycki, J.A., 2010. Selenocysteine, pyrrolysine, and the unique energy metabolism of methanogenic archaea. *Archaea* 2010, 453642.
- Saito, K., Ito, K., 2015. Genetic analysis of L123 of the tRNA-mimicking eukaryote release factor eRF1, an amino acid residue critical for discrimination of stop codons. *Nucleic Acids*

- Res. 43, 4591–4601.
- Salas-Marco, J., Fan-Minogue, H., Kallmeyer, A.K., Klobutcher, L.A., Farabaugh, P.J., Bedwell, D.M., 2006. Distinct paths to stop codon reassignment by the variant-code organisms *Tetrahymena* and *Euplotes*. *Mol. Cell. Biol.* 26, 438–447.
- Salgado, H., Moreno-Hagelsieb, G., Smith, T.F., Collado-Vides, J., 2000. Operons in *Escherichia coli*: Genomic analyses and predictions. *Proc. Natl. Acad. Sci. U. S. A.* 97, 6652–6657.
- Saurer, M., Ramrath, D.J.F., Niemann, M., Calderaro, S., Prange, C., Mattei, S., Scaiola, A., Leitner, A., Bieri, P., Horn, E.K., Leibundgut, M., Boehringer, D., Schneider, A., Ban, N., 2019. Mitochondrial small subunit biogenesis in trypanosomes involves an extensive assembly machinery. *Science* 365, 1144–1149.
- Schatz, G., Mason, T.L., 1974. The biosynthesis of mitochondrial proteins. *Annu. Rev. Biochem.* 43, 51–87.
- Schmeing, T.M., Voorhees, R.M., Kelley, A.C., Ramakrishnan, V., 2011. How mutations in tRNA distant from the anticodon affect the fidelity of decoding. *Nat. Struct. Mol. Biol.* 18, 432–437.
- Schultz, D.W., Yarus, M., 1994. Transfer RNA mutation and the malleability of the genetic code. *J. Mol. Biol.* 235, 1377–1380.
- Scolnick, E., Tompkins, R., Caskey, T., Nirenberg, M., 1968. Release factors differing in specificity for terminator codons. *Proc. Natl. Acad. Sci. U. S. A.* 61, 768–774.
- Sekine, S.I., Nureki, O., Sakamoto, K., Niimi, T., Tateno, M., Go, M., Kohno, T., Brisson, A., Lapointe, J., Yokoyama, S., 1996. Major identity determinants in the “augmented D helix” of tRNA^{Glu} from *Escherichia coli*. *J. Mol. Biol.* 256, 685–700.
- Sengupta, S., Higgs, P.G., 2015. Pathways of genetic code evolution in ancient and modern organisms. *J. Mol. Evol.* 80, 229–243.
- Sengupta, S., Yang, X., Higgs, P.G., 2007. The mechanisms of codon reassignments in mitochondrial genetic codes. *J. Mol. Evol.* 64, 662–688.
- Sharma, M.R., Booth, T.M., Simpson, L., Maslov, D.A., Agrawal, R.K., 2009. Structure of a mitochondrial ribosome with minimal RNA. *Proc. Natl. Acad. Sci. U. S. A.* 106, 9637–9642.
- Shi, X., Chen, D.H.T., Suyama, Y., 1994. A nuclear tRNA gene cluster in the protozoan

- Leishmania tarentolae* and differential distribution of nuclear-encoded tRNAs between the cytosol and mitochondria. *Mol. Biochem. Parasitol.* 65, 23–37.
- Siegel, T.N., Hekstra, D.R., Kemp, L.E., Figueiredo, L.M., Lowell, J.E., Fenyo, D., Wang, X., Dewell, S., Cross, G.A.M.M., 2005. Four histone variants mark the boundaries of polycistronic transcription units in *Trypanosoma brucei*. *Genes Dev.* 23, 1063–1076.
- Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E. V., 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Genome Anal.* 9–10.
- Simpson, A.G.B., 1997. The identity and composition of the Euglenozoa. *Arch. für Protistenkd.* 148, 318–328.
- Simpson, A.M., Suyama, Y., Dewes, H., Campbell, D.A., Simpson, L., 1989. Kinetoplastid mitochondria contain functional tRNAs which are encoded in nuclear DNA and also contain small minicircle and maxicircle transcripts of unknown function. *Nucleic Acids Res.* 17, 5427–5445.
- Sladic, R.T., Lagnado, C.A., Bagley, C.J., Goodall, G.J., 2004. Human PABP binds AU-rich RNA via RNA-binding domains 3 and 4. *Eur. J. Biochem.* 271, 450–457.
- Smit, A., Hubley, R., Green, P., 2015. RepeatMasker. <http://www.repeatmasker.org>.
- Sommer, J.R., 1965. The ultrastructure of the pellicle complex of *Euglena gracilis*. *J. Cell Biol.* 24, 253–257.
- Steinbiss, S., Silva-Franco, F., Brunk, B., Foth, B., Hertz-Fowler, C., Berriman, M., Otto, T.D., 2016. *Companion* : a web server for annotation and analysis of parasite genomes. *Nucleic Acids Res.* 44, W29–W34.
- Swart, E.C., Serra, V., Petroni, G., Nowacki, M., 2016. Genetic codes with no dedicated stop codon context-dependent translation termination. *Cell* 166, 691–702.
- Tan, T.H.P., Pach, R., Crausaz, A., Ivens, A., Schneider, A., 2002. tRNAs in *Trypanosoma brucei*: genomic organization, expression, and mitochondrial import. *Mol. Cell. Biol.* 22, 3707–3717.
- Teixeira, M.M.G., Borghesan, T.C., Ferreira, R.C., Santos, M.A., Takata, C.S.A., Campaner, M., Nunes, V.L.B., Milder, R. V., de Souza, W., Camargo, E.P., 2011. Phylogenetic validation of the genera *Angomonas* and *Strigomonas* of trypanosomatids harboring bacterial endosymbionts with the description of new species of trypanosomatids and of

- proteobacterial symbionts. *Protist* 162, 503–524.
- Teixeira, S.M., de Paiva, R.M.C., Kangussu-Marcolino, M.M., DaRocha, W.D., Paiva, R.M.C. de, Kangussu-Marcolino, M.M., DaRocha, W.D., 2012. Trypanosomatid comparative genomics: Contributions to the study of parasite biology and different parasitic diseases. *Genet. Mol. Biol.* 35, 1–17.
- Thomas, S., Green, A., Sturm, N.R., Campbell, D.A., Myler, P.J., 2009. Histone acetylations mark origins of polycistronic transcription in *Leishmania major*. *BMC Genomics* 10, 152.
- Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., Pachter, L., 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7, 562–578.
- Turanov, A.A., Lobanov, A. V., Fomenko, D.E., Morrison, H.G., Sogin, M.L., Klobutcher, L.A., Hatfield, D.L., Gladyshev, V.N., 2009. Genetic code supports targeted insertion of two amino acids by one codon. *Science* 323, 259–261.
- Tzagoloff, A., Macino, G., 1979. Mitochondrial Genes and Translation Products. *Annu. Rev. Biochem.* 48, 419–439.
- Ulmasov, B., Topin, A., Chen, Z., He, S.H., Folk, W.R., 1998. Identity elements and aminoacylation of plant tRNA^{Trp}. *Nucleic Acids Res.* 26, 5139–5141.
- Vaser, R., Šikić, M., 2019. Yet another de novo genome assembler. *bioRxiv* 656306.
- Vetsigian, K., Woese, C., Goldenfeld, N., 2006. Collective evolution and the genetic code. *Proc. Natl. Acad. Sci.* 103, 10696–10701.
- Votýpka, J., D’Avila-Levy, C.M., Grellier, P., Maslov, D.A., Lukeš, J., Yurchenko, V., 2015. New approaches to systematics of trypanosomatidae: criteria for taxonomic (re)description. *Trends Parasitol.* 31, 460–469.
- Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., Earl, A.M., 2014. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9, e112963.
- Worthey, E.A., Martinez-Calvillo, S., Schnauffer, A., Aggarwal, G., Cawthra, J., Fazelinia, G., Fong, C., Fu, G., Hassebrock, M., Hixson, G., Ivens, A.C., Kiser, P., Marsolini, F., Rickell, E., Salavati, R., Sisk, E., Sunkin, S.M., Stuart, K., Myler, P.J., 2003. *Leishmania major* chromosome 3 contains two long convergent polycistronic gene clusters separated by a

- tRNA gene. *Nucleic Acids Res.* 31, 4201–4210.
- Xue, H., Shen, W., Giege, R., Wong, J.T.F., 1993. Identity elements of tRNA(Trp). Identification and evolutionary conservation. *J. Biol. Chem.* 268, 9316–9322.
- Yesland, K.D., Nelson, A.W., Six Feathers, D.M., Johnson, J.D., 1993. Identity of *Saccharomyces cerevisiae* tRNA(Trp) is not changed by an anticodon mutation that creates an amber suppressor. *J. Biol. Chem.* 268, 217–220.
- Yona, A.H., Bloom-Ackermann, Z., Frumkin, I., Hanson-Smith, V., Charpak-Amikam, Y., Feng, Q., Boeke, J.D., Dahan, O., Pilpel, Y., 2013. tRNA genes rapidly change in evolution to meet novel translational demands. *Elife* 2, e01339.
- Yubuki, N., Simpson, A.G.B., Leander, B.S., 2013. Reconstruction of the feeding apparatus in *Postgaardia mariagerensis* provides evidence for character evolution within the Symbiontida (Euglenozoa). *Eur. J. Protistol.* 49, 32–39.
- Záhonová, K., Kostygov, A.Y., Ševčíková, T., Yurchenko, V., Eliáš, M., 2016. An unprecedented non-canonical nuclear genetic code with all three termination codons reassigned as sense codons. *Curr. Biol.* 26, 2364–2369.
- Žihala, D., Eliáš, M., 2019. Evolution and unprecedented variants of the mitochondrial genetic code in a lineage of green algae. *Genome Biol. Evol.* 11, 2992–3007.
- Zíková, A., Panigrahi, A.K., Dalley, R.A., Acestor, N., Anupama, A., Ogata, Y., Myler, P.J., Stuart, K., 2008. *Trypanosoma brucei* mitochondrial ribosomes: affinity purification and component identification by mass spectrometry. *Mol. Cell. Proteomics* 7, 1286–1296.
- Zinoviev, A., Shapira, M., 2012. Evolutionary conservation and diversification of the translation initiation apparatus in trypanosomatids. *Comp. Funct. Genomics* 2012, 813718.
- Zoltner, M., Krienitz, N., Field, M.C., Kramer, S., 2018. Comparative proteomics of the two *T. brucei* PABPs suggests that PABP2 controls bulk mRNA. *PLoS Negl. Trop. Dis.* 12, e0006679.

Pages 73-83 contain classified information (unpublished data) and are available only in the archived original of the graduation thesis deposited at the Faculty of Sciences, University of South Bohemia.



Genome of *Ca. Pandoraea novymonadis*, an Endosymbiotic Bacterium of the Trypanosomatid *Novymonas esmeraldas*

Alexei Y. Kostygov^{1,2†}, Anzhelika Butenko^{1,3†}, Anna Nenarokova^{3,4}, Daria Tashyreva³, Pavel Flegontov^{1,3,5}, Julius Lukeš^{3,4} and Vyacheslav Yurchenko^{1,3,6*}

¹ Life Science Research Centre, Faculty of Science, University of Ostrava, Ostrava, Czechia, ² Zoological Institute of the Russian Academy of Sciences, St. Petersburg, Russia, ³ Biology Centre, Institute of Parasitology, Czech Academy of Sciences, České Budějovice, Czechia, ⁴ Faculty of Sciences, University of South Bohemia, České Budějovice, Czechia, ⁵ Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia, ⁶ Institute of Environmental Technologies, Faculty of Science, University of Ostrava, Ostrava, Czechia

OPEN ACCESS

Edited by:

João Marcelo Pereira Alves,
University of São Paulo, Brazil

Reviewed by:

Zhao-Rong Lun,
Sun Yat-sen University, China
Vera Tai,
University of Western Ontario, Canada

*Correspondence:

Vyacheslav Yurchenko
vyacheslav.yurchenko@osu.cz

† These authors have contributed
equally to this work.

Specialty section:

This article was submitted to
Microbial Symbioses,
a section of the journal
Frontiers in Microbiology

Received: 27 June 2017

Accepted: 21 September 2017

Published: 04 October 2017

Citation:

Kostygov AY, Butenko A,
Nenarokova A, Tashyreva D,
Flegontov P, Lukeš J and
Yurchenko V (2017) Genome of *Ca.*
Pandoraea novymonadis, an
Endosymbiotic Bacterium of the
Trypanosomatid *Novymonas*
esmeraldas. *Front. Microbiol.* 8:1940.
doi: 10.3389/fmicb.2017.01940

We have sequenced, annotated, and analyzed the genome of *Ca. Pandoraea novymonadis*, a recently described bacterial endosymbiont of the trypanosomatid *Novymonas esmeraldas*. When compared with genomes of its free-living relatives, it has all the hallmarks of the endosymbionts' genomes, such as significantly reduced size, extensive gene loss, low GC content, numerous gene rearrangements, and low codon usage bias. In addition, *Ca. P. novymonadis* lacks mobile elements, has a strikingly low number of pseudogenes, and almost all genes are single copied. This suggests that it already passed the intensive period of host adaptation, which still can be observed in the genome of *Polynucleobacter necessarius*, a certainly recent endosymbiont. Phylogenetically, *Ca. P. novymonadis* is more related to *P. necessarius*, an intracytoplasmic bacterium of free-living ciliates, than to *Ca. Kinetoplastibacterium* spp., the only other known endosymbionts of trypanosomatid flagellates. As judged by the extent of the overall genome reduction and the loss of particular metabolic abilities correlating with the increasing dependence of the symbiont on its host, *Ca. P. novymonadis* occupies an intermediate position *P. necessarius* and *Ca. Kinetoplastibacterium* spp. We conclude that the relationships between *Ca. P. novymonadis* and *N. esmeraldas* are well-established, although not as fine-tuned as in the case of Strigomonadinae and their endosymbionts.

Keywords: bacterial endosymbiont, *Pandoraea*, phylogenomics, metabolism, Trypanosomatidae

INTRODUCTION

Pandoraea is a genus of Gram-negative rod-shaped β -proteobacteria belonging to the family Burkholderiaceae of the order Burkholderiales. Members of this genus are phenotypically diverse, reflecting a wide spectrum of life strategies. Several species of these microorganisms were documented as opportunistic pathogens in cystic fibrosis patients or in individuals after lung transplantation (Coenye et al., 2000; Stryjewski et al., 2003). Besides, a number of *Pandoraea* spp. (including some pathogenic ones) were isolated from environmental samples such as soils, hen

ding, and oxic water layer above a sulfide-containing sediment (Coenye et al., 2000; Anandham et al., 2010; Sahin et al., 2011). These free-living species participate in the biodegradation of various organic substances (including important pollutants) or perform chemosynthesis by oxidation of heterotrophic sulfur (Okeke et al., 2002; Graff and Stubner, 2003; Ozaki et al., 2007; Liz et al., 2009; Kumar et al., 2015; Jeong et al., 2016).

Previously, we discovered a new species of *Pandoraea*, which, in contrast to its relatives, is an intracellular symbiont of the flagellate *Novymonas esmeraldas* (Kinetoplastea: Trypanosomatidae) (Flegontov et al., 2016). This endosymbiosis appears to have been established relatively recently as judged by the fact that neither of the two participants has close relatives involved in similar relationships. In addition, the division of *Ca. Pandoraea novymonadis* is not synchronized with that of the host cell. As a result, the number of endosymbionts per *Novymonas* cell is unstable and bacteria-free trypanosomatids appear at a relatively high frequency of ~6%. We hypothesized that the endosymbiosis is favorable for *N. esmeraldas*, since large-scale cloning experiments did not reveal any aposymbiotic clone (Flegontov et al., 2016).

All other studied endosymbioses in trypanosomatids are restricted to flagellates of the subfamily Strigomonadinae (Votýpka et al., 2014) and bacteria *Ca. Kinetoplastibacterium* spp. (Burkholderiales: Alcaligenaceae). These relationships seem to have been established earlier in evolution. As judged from the phylogenies of the prokaryotic and eukaryotic partners, the origin of this endosymbiosis was a single event (Du et al., 1994; Teixeira et al., 2011). The long evolution of *Kinetoplastibacterium* resulted in “one bacterium per host cell” arrangement with fine-tuned mechanisms synchronizing their division (Motta et al., 2010; Catta-Preta et al., 2015). The bacterium provides its host with essential nutrients and is remunerated with a direct access to the ATP-producing glycosomes (Motta et al., 1997; de Souza and Motta, 1999; Alves et al., 2011, 2013a).

The free-living hypotrichous ciliates (*Euplotes aediculatus* and related species) with their intracytoplasmic bacterium *Ca. Polynucleobacter necessarius* (hereafter conventionally called *P. necessarius*) represent yet another endosymbiotic association in a protist, reminiscent of the *Novymonas/Pandoraea* system. Although the ciliates are evolutionary extremely distant from trypanosomatids and represent a different eukaryotic supergroup (SAR versus Excavata), their endosymbiont *P. necessarius* belongs to the same β -proteobacterial family *Burkholderiaceae*. This endosymbiosis seems to be quite recent, since there is a very closely related free-living bacterium formally attributed to a separate species *P. asymbioticus*, but showing 99% identity with *P. necessarius* in their 16S rRNA gene sequences (Vannini et al., 2007). Another sign of the relatively nascent nature of these relationships is that *P. necessarius* is apparently a substitute for a more ancient symbiont (*Ca. Protistobacter heckmanni*), another representative of the family *Burkholderiaceae*, which can be found in some *Euplotes* isolates (Vannini et al., 2012, 2013).

Obligate intracellular bacterial symbionts demonstrate similar patterns of genome evolution: reduction of its size, decrease in GC content, elevated evolutionary rate, loss of genes from certain

functional groups (transcriptional regulation, DNA repair, etc.), shrinkage of the repertoire of metabolic capabilities, gene transfer to host's nucleus, and others (Moya et al., 2008; Nowack and Melkonian, 2010; McCutcheon and Moran, 2011). At early phase of endosymbiosis these changes are accompanied by the expansion of mobile genetic elements, pseudogenization, and multiple genomic rearrangements (Ochman and Davalos, 2006; Toh et al., 2006; Burke and Moran, 2011). In the case of Strigomonadinae/*Ca. Kinetoplastibacterium*, all the above-mentioned traits typical of ancient endosymbiotic associations can be observed (Alves et al., 2013b). The comparison of genomes of *P. necessarius* and *P. asymbioticus* revealed only a limited genome size reduction (~28% on DNA and ~34% on the protein level) with a substantial pseudogenization (~18%), but without any mobile elements (Meincke et al., 2012; Boscaro et al., 2013).

While the host of *Ca. P. novymonadis* is closely related to that of *Ca. Kinetoplastibacterium* spp., the bacterium itself is phylogenetically closer to *Polynucleobacter*. In order to understand the nature of endosymbiotic relationships, their underlying mechanisms and routes of adaptation in the *Novymonas/Pandoraea* system, we analyzed the genome of *Ca. P. novymonadis* and compared it with both endosymbiotic systems discussed above.

MATERIALS AND METHODS

Establishing Aposymbiotic Strain of *Novymonas esmeraldas*

The strain E262AT.01 of *N. esmeraldas* was cultivated at 27°C in RPMI-1640 medium (Sigma—Aldrich, St. Louis, MO, United States) supplemented with heat-inactivated 10% fetal bovine serum (FBS; Thermo Fisher Scientific, Waltham, MA, United States). At the logarithmic phase of growth, cells from the 10 ml culture aliquots were pelleted by centrifugation at 1,500 × g for 10 min and re-suspended in the fresh RPMI-1640 medium containing 10, 50, 125, 250, or 500 µg/ml of azithromycin (Barry et al., 2004). This macrolide antibiotic was chosen because of its ability to cross eukaryotic plasma membrane, accumulate in the cytoplasm at high concentration, and retain its activity under these conditions (Maurin and Raoult, 2001; Carryn et al., 2003). The presence/absence of bacterial endosymbionts was monitored after 7 and 14 days of incubation by fluorescent *in situ* hybridization with universal bacteria-specific probe Eub338 (5'-GCTGCCTCCCGTAGGAGT-3') labeled with 5'-Cy3 fluorescent dye, as described previously (Kostygov et al., 2016). After 14 days incubation with 10 and 50 µg/ml of azithromycin, all observed *N. esmeraldas* cells were free of endosymbionts, while at the higher concentrations of the antibiotic trypanosomatid cells died. The bacteria-free cultures were pelleted and transferred to a fresh azithromycin-free medium. The strain obtained with 10 µg/ml of azithromycin (hereafter named E262-AZI) displayed better growth and was used for all the subsequent experiments. The absence of bacteria in the culture was also confirmed by PCR with universal eubacterial 16S rRNA primers P1seq and 1486R, with the original bacteria-containing strain (hereafter named E262-wt) used as a positive control (Teixeira et al., 2011).

Given the significant deceleration of E262-AZI growth as compared to the E262-wt, for the subsequent work we switched from RPMI to a more nutrient-rich medium, M199 (Sigma—Aldrich, St. Louis, MO, United States) supplemented with 10% FBS, 2 $\mu\text{g/ml}$ hemin (Jena Bioscience, Jena, Germany), 2 $\mu\text{g/ml}$ biotin, 100 units/ml of penicillin, and 100 $\mu\text{g/ml}$ of streptomycin (all from Thermo Fisher Scientific, Waltham, MA, United States). In these conditions, E262-AZI was able to propagate at higher rate, comparable to that of E262-wt.

Genomic DNA Isolation and Sequencing

Total genomic DNA was isolated from $\sim 10^9$ cells of the strains E262-wt and E262-AZI of *N. esmeraldas* using the DNeasy Blood and Tissue Kit (Qiagen, Hilden, Germany) according to the manufacturer's protocol. The genome of the wild-type *N. esmeraldas* was sequenced using a combination of Illumina Technologies: HiSeq 2000 (Macrogen Inc., Seoul, South Korea) and MiSeq (Palacký University, Olomouc, Czechia), yielding 47,024,780 reads with 145 \times average coverage and 21,715,370 reads with 160 \times average coverage, respectively. The genome of the aposymbiotic *N. esmeraldas* was sequenced solely with the Illumina MiSeq technology, resulting into 17,834,848 of reads with 136 \times average coverage. The lengths of the obtained paired-end reads were 100 nt for the HiSeq and 300 nt for the MiSeq sequences.

Genome Assembly and Annotation

DNA sequencing reads were processed using BBtools package v.36.02¹. The reads were merged and quality-trimmed using BBmerge with the quality threshold of 20. Non-merged reads were quality-trimmed using BBduk with the same parameters. The quality of raw and trimmed reads was assessed using FASTQC program v.0.11.5².

The genome assembly for both strains was performed using Spades Genome assembler v.3.9.0 with recommended options (Bankevich et al., 2012). Genomic reads of E262-wt were mapped onto the contigs of the aposymbiotic E262-AZI and the remaining reads were used for assembling the endosymbiont genome. However, a read mapping rate was low ($\sim 50\%$) and the obtained assembly contained both endosymbiont and host contigs. Hence, we decided to use other methods for identification of the bacterial contigs. Firstly, each of E262-wt contigs was used as a query in BLAST searches against the custom database composed of *Pandoraea* spp. and trypanosomatid genomes. The BLASTN program from the BLAST package v.2.2.31+ (Camacho et al., 2009) was used with an *E*-value cut-off of 10^{-5} and other settings left as default. The total length of a BLAST alignment per contig was calculated using custom Ruby script. For every contig, the query coverage with *Pandoraea* hits was divided by that with trypanosomatid hits. The values above 1 were considered as evidencing the bacterial origin. Secondly, we checked the absence of the putative endosymbiont contigs in the E262-AZI assembly using the BLASTN program as above. The best hits for the presumed bacterial contigs were

those with low coverage ($\sim 1\times$), probably representing technical contamination during sequencing of the E262-AZI sample. Thirdly, the read coverage was considered for distinguishing contigs of *N. esmeraldas* and *Ca. P. novyimonadis*. Typically, a cell of this trypanosomatid bears several endosymbionts in the cytoplasm (Kostygov et al., 2016), and each of them might have multiple copies of the bacterial genome. Therefore, the read coverage of *Ca. P. novyimonadis* contigs is expected to be higher than that of *N. esmeraldas* contigs. Indeed, the mean coverage per position in the putative bacterial contigs was $\sim 874\times$ while the remaining ones had only $\sim 25\times$ read coverage. In addition, the contigs of different origin could be discriminated by their GC content. Trypanosomatid contigs had $\sim 65\%$ GC, while those of the endosymbiont were only 43–49% GC-rich. This is in agreement with the observation that endosymbiotic genomes usually have lower GC content than the genomes of their hosts (Moran et al., 2008). It should be noted that GC-rich sequences are generally harder to sequence than AT-rich, and this effect may impact the coverage difference and result in overestimation of the bacterial load. Lastly, the Bandage software³, analyzing assembly using a BLAST-based approach, was used. The program created contig graphs, showing that all the putative endosymbiont contigs may compose a single circular chromosome (under assumption that the two shortest bacterial contigs having double coverage, as compared to longer ones are duplicated). This also evidenced that our assembly was complete. Despite the results of the Bandage analysis, we were unable to assemble the bacterial contigs into one chromosome due to some ambiguities. Genome completeness analysis was performed using BUSCO software (Simão et al., 2015) with bacteria, proteobacteria, and betaproteobacteria universal gene datasets and the predicted *Ca. P. novyimonadis* proteins.

Parameters of the genome assemblies were estimated using QUAST v.4.3 (Gurevich et al., 2013). DNA reads were mapped on the assemblies using Bowtie2 v.2.2.9 (Langmead and Salzberg, 2012), with the “–very-fast” option. The structural and functional annotation of the *Ca. P. novyimonadis* genome was obtained using Prokka package v.1.12-beta (Seemann, 2014), signal peptides were predicted using SignalP v.4.1 (Petersen et al., 2011).

This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession MUHY00000000. The version described in this paper is version MUHY01000000. The raw reads are available at the NCBI Sequence Read Archive under the accession no. SRR5280512.

Gene Family Inference and Analysis

The inference of protein orthologous groups (OGs) was performed with OrthoFinder v.1.1.3 (Emms and Kelly, 2015) using a dataset of 23 bacterial genomes, including *Ca. P. novyimonadis* sequenced in this study, 13 other *Pandoraea* spp., 5 *Ca. Kinetoplastibacterium* spp., 2 *Polynucleobacter* spp., *Cupriavidus basilensis*, and *Burkholderia cepacia* available in GenBank (Supplementary Table S1). Gene family gains and losses were mapped on the reference species tree using the

¹<https://sourceforge.net/projects/bbtools/>

²<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

³<https://academic.oup.com/bioinformatics/article/31/20/3350/196114/Bandage-interactive-visualization-of-de-novo>

COUNT software with the Dollo parsimony algorithm (Csuros, 2010) as described elsewhere (Flegontov et al., 2016). Using UpSetR package for R⁴ and a custom Python script we found OGs exclusively shared between *Ca. P. novymonadis* and the following groups of species: (i) *C. basilensis* and *B. cepacia*, (ii) *Polynucleobacter* spp., (iii) *Pandoraea* spp., and (iv) *Ca. Kinetoplastibacterium* spp. Putative annotations for the *Ca. P. novymonadis*-specific proteins were inferred using HHpred v.2.0.16 against Pfam-A database and *E*-value cut-off set to 1 (Soding et al., 2005).

Phylogenomic Analysis

In the 16 bacterial strains selected for phylogenetic inference (**Supplementary Table S1**) 556 shared OGs contained only one gene. The amino acid sequences of each single gene were aligned using L-INS-i algorithm in MAFFT v. 7.310 (Katoh and Standley, 2013). The resulting alignments were trimmed in Gblocks v.0.91b with relaxed parameters ($-b3 = 8$, $-b4 = 2$, $-b5 = h$) and then used for phylogenetic reconstruction in IQ-TREE v.1.5.3 with LG + I + G4 + F model and 1,000 ultrafast bootstrap replicates (Minh et al., 2013; Nguyen et al., 2015). The amino acid substitution model had been selected in the same program using the supermatrix concatenated from the individual alignments of all 556 genes (Kalyanamoorthy et al., 2017). To estimate the resolution power of single genes, for each of the reconstructed trees the average bootstrap support was calculated. Setting 70% as a threshold, we selected 119 genes, which constituted the final dataset. The alignments of these genes were concatenated, producing a supermatrix with 54,345 characters. Maximum-likelihood tree was reconstructed using IQ-Tree with LG + I + G4 + F model and 1,000 standard bootstrap replicates. Bayesian inference of phylogeny was performed in MrBayes v. 3.2.6 (Ronquist et al., 2012) under mixed model prior, empirical amino acid frequencies, and heterogeneity of rates across sites assessed using Γ -distribution and proportion of invariant sites. The analysis was run for 100,000 generations with sampling every 10th of them. The chains demonstrated efficient mixing and the two runs converged at the early phase of the analysis (after 2,500 generations). As set by default, 25% samples were discarded as burn-in.

Metabolic Pathways Analysis

For the comparative metabolic study an automatic assignment of KEGG Orthology (KO) identifiers to the proteins of 19 bacterial strains including *Ca. P. novymonadis* (**Supplementary Table S1**) was completed using BlastKOALA v.2.1 (Kanehisa et al., 2016). The search was performed against a non-redundant pangenomic databases of prokaryotes on the genus level and of eukaryotes on the family level. KEGG Mapper v.2.8 was used for the reconstruction of metabolic pathways and their comparison (Kanehisa, 2017).

The search for lipolytic enzymes was performed using BLASTP with an *E*-value of 10^{-20} with the lipase and esterase sequences from the study of Arpigny and Jaeger as a query and annotated proteins of *Ca. P. novymonadis* and other bacteria

as a database (Arpigny and Jaeger, 1999). In the case of *Ca. P. novymonadis* the *E*-value threshold was relaxed to 10^{-10} .

Synteny Analysis

The overall level of synteny in *Ca. P. novymonadis* as compared to other species of interest was studied using the reference dataset of 11 bacteria (**Supplementary Table S1**). Syntenic regions were inferred and visualized using SyMAP v.4.2 (Soderlund et al., 2011). The settings were as follows: minimum number of anchors required to define a synteny block, 7; overlapping (or nearby) synteny blocks were automatically merged into larger blocks, and only the larger block was kept if two synteny blocks overlapped on a chromosome.

Search for Pseudogenes, Phages, and Mobile Elements

Pseudogenes in *Ca. P. novymonadis* genome were identified using BLASTX with an *E*-value cut-off of 1 against the dataset of annotated proteins of *C. basilensis*, *B. cepacia*, and *Pandoraea* spp. (**Supplementary Table S1**). Prior to homology searches, *Ca. P. novymonadis* genes were masked with Maskfasta script from BEDTools package v. 2.25.0 (Quinlan and Hall, 2010). Genomic regions with BLAST hits were manually inspected and the coordinates of the BLAST hits were used for annotation of pseudogenes. We also checked the presence of pseudogenes among the features annotated with Prokka package by analyzing the annotations of the adjacent genes and concluded that all of them were functional.

The search for mobile elements and phages in the genome of *Ca. P. novymonadis* was performed algorithmically in Phispy v. 2.3 (Akhter et al., 2012), as well as using database searches on the online web servers Phaster (Arndt et al., 2016) and IS Finder⁵ using *E*-value cut-off of 10^{-2} .

Analyses of Genome Sequence Properties

Files with the genome sequences and corresponding annotations for the species of interest were downloaded from the NCBI Genome database (12.12.2016). Pseudogene sequences were excluded from further analyses. Lengths of genes and intergenic regions were calculated based on the gene coordinates within GFF files containing annotation data.

For the analysis of GC content, nucleotide sequences of all genes were extracted using Artemis genome browser release v. 16.0.0 (Rutherford et al., 2000). GC content was calculated with Infoseq script from EMBOSS package v. 6.6.0.0 (Rice et al., 2000). Statistical significance of the differences in GC content, lengths of genes, and intergenic regions was tested using one-way analysis of variance (ANOVA) combined with Tukey's honest significance test in R with *p*-value < 0.05.

Nucleotide composition by codon position, amino acid composition, and codon usage bias of protein-coding genes were analyzed using MEGA 7.0 software (Kumar et al., 2016) on the concatenated sequences of all these genes within a genome.

⁴<https://cran.r-project.org/web/packages/UpSetR/>

⁵<http://phaster.ca> and <https://www-is.biotoul.fr>

Standard deviation of relative synonymous codon usage (RSCU) values (Sharp et al., 1986) was calculated as an integral measure of codon usage bias in a particular species. Stop codons and the two amino acids coded by only one codon (methionine and tryptophan) were excluded.

RESULTS AND DISCUSSION

General Characterization of *Ca. P. novyimonadis* Genome

The genome of *Ca. P. novyimonadis* was assembled into six contigs with a total length of approximately 1.16 Mb (**Supplementary Table S1**), which is smaller than in free-living *Pandoraea* spp. (4.46–6.5 Mb) or in both *Polynucleobacter* spp. (1.56–2.16 Mb), but larger than in *Ca. Kinetoplastibacterium* spp. (~0.8 Mb). The average coverage with the paired-end 100 nt Illumina HiSeq and 300 nt MiSeq reads was ~874× and the largest contig had the length of 844,906 nt. The two shortest contigs (5,920 and 1,318 bp), containing genes for ribosomal RNA, translation factor Tu 1 and tRNAs for alanine, isoleucine, and tryptophan had approximately doubled coverage (1,555× and 1,864×, respectively) pointing to the probable duplication of these fragments in the genome. The assessment of genome assembly and annotation completeness with single-copy orthologs using BUSCO demonstrated that 147/148 (99.3%) universal genes from bacteria dataset, 216/221 (97.7%) from proteobacteria, and 529/582 (90.9%) from betaproteobacteria were present. This indicates that our assembly was complete.

Currently, there are 1,015 annotated genes, 968 of which are protein-coding. For comparison, free-living *Pandoraea* spp. have 3,960–5,342, *Polynucleobacter* spp. – 1,401 and 2,115, while *Ca. Kinetoplastibacterium* spp. only 690–732 protein-coding genes (**Supplementary Table S1**). The number of identified pseudogenes in *Ca. P. novyimonadis* (13) is significantly smaller than in other species of the genus *Pandoraea* (76–361) but is comparable to that in *Ca. Kinetoplastibacterium* spp. (2–19) (**Supplementary Table S1**). Interestingly, *P. necessarius* possesses a high number of pseudogenes (269), which is apparently indicative of intense process of genome evolution and is in agreement with a recent origin of endosymbiosis in this species (Vannini et al., 2007).

No mobile elements were found in the genome of *Ca. P. novyimonadis* with any of the used tools. This appears to be a consequence of genome minimization. The genome of this species has lost ~80% of its length and protein-coding capacity compared to the genomes of its free-living *Pandoraea* spp. (**Supplementary Table S1**). We did not find statistically significant differences between the lengths of genes and intergenic regions of *Ca. P. novyimonadis* compared to other *Pandoraea* spp., *Ca. Kinetoplastibacterium* spp., *Polynucleobacter* spp., *C. basilensis*, and *B. cepacia* (**Supplementary Figure S1**).

The comparison of GC content in *Ca. P. novyimonadis* with that of *P. apista*, *P. necessarius*, and *Ca. Kinetoplastibacterium* crithidii genomes revealed significant differences both in genes and intergenic regions between *Ca. P. novyimonadis* and the other analyzed species (**Supplementary Figure S2**). Interestingly,

these differences were most pronounced in the genomes of trypanosomatid endosymbionts, *Ca. P. novyimonadis*, and *Ca. K. crithidii*. The average GC content of the *Ca. P. novyimonadis* genome (43.8%) is intermediate between that of the free-living *Pandoraea* spp. (62–65%) and *Ca. Kinetoplastibacterium* spp. (30–33%). However, it is similar to that of both endosymbiotic and free-living *Polynucleobacter* spp. (45.6 and 44.8, respectively). This pattern is also conspicuous when considering nucleotide composition in protein coding genes by codon position, with the most pronounced differences at the third position (**Supplementary Figure S3**). We found 35 genes in the *Ca. P. novyimonadis* genome with the GC content higher than 56% and all these genes encode tRNAs. This is in agreement with an earlier observation that in prokaryotes the GC content of such genes does not correlate with that of the whole genome (Kawai and Maeda, 2009).

The amino acid frequencies in *Ca. P. novyimonadis* differ from those in its close relatives. The most discordant ones are for alanine, isoleucine, and lysine (**Supplementary Figure S4**). As with the nucleotide composition, the amino acids frequencies in this species are intermediate between those of other *Pandoraea* spp. and *Ca. Kinetoplastibacterium* spp. and appear most similar to those in *Polynucleobacter* spp.

In agreement with the previously described general trend, the codon usage bias in analyzed species correlated with genomic GC content (Sharp et al., 2005). This relationship was represented by a sideways parabola with the vertex (i.e., lowest value of RSCU standard deviation) situated at about 50% GC: further from the equilibrium nucleotide frequencies, the more pronounced was the bias. Most of the Alcaligenaceae and Burkholderiaceae species fitted this parabolic curve. Three notable exceptions were *Ca. P. novyimonadis* (possessing the least prominent codon usage bias) and the two *Polynucleobacter* spp. (**Supplementary Figure S5**). It was previously proposed that species under selection for rapid growth have stronger codon usage bias (Sharp et al., 2005, 2010). However, this is not the case here. In terms of growth rate, the outliers *Ca. P. novyimonadis* and *P. necessarius* do not differ much from *Ca. Kinetoplastibacterium* spp. fitting to the trend, since all these bacteria are endosymbionts. An alternative explanation appears to be more plausible: the bacteria that have to switch gene expression from time to time (usually owing to the changing environment) have a stronger bias as compared to those living in stable conditions (Botzman and Margalit, 2011). Although *Ca. Kinetoplastibacterium* spp. are endosymbionts, their close interactions with the host, reflected by a tight coordination of their cell divisions, may lead to similar switches. As for *Ca. P. novyimonadis*, its relationship with the host cell seems to be more relaxed (Kostygov et al., 2016) and apparently does not require complex gene expression.

Synteny analysis with free-living *Pandoraea* spp. demonstrated that 62–69% of “anchors” (pairwise alignments) in *Ca. P. novyimonadis* genome are located within synteny blocks with maximal values observed for *P. faecigallinarum* and *P. vervacti* (**Supplementary Table S2**). The fact that the majority of the synteny blocks are inverted (15/24 and 11/21 for *P. faecigallinarum* and *P. vervacti*, respectively), reflects a relatively long evolutionary distances between these species

and *Ca. P. novymonadis*. The pairwise synteny between *Ca. P. novymonadis* and the genomes of other *Pandoraea* spp. available in GenBank is presented in **Supplementary Figure S6**. This analysis demonstrated the reduction of the *Ca. P. novymonadis* genome compared to those of free-living *Pandoraea* spp. and a high number of genome rearrangements occurring in the evolution of this endosymbiotic bacterium.

Thus, sequencing and annotation of the *Ca. P. novymonadis* genome revealed several features characteristic for other endosymbiotic bacteria: reduced size, massive gene losses, and decrease in GC content as compared to the genomes of its free-living relatives (Boscaro et al., 2017). Taken together, *Ca. P. novymonadis* is closer to *P. necessarius* than to *Ca. Kinetoplastibacterium* spp.

Phylogenomic Analysis

The maximum likelihood and Bayesian trees inferred using the supermatrix containing 119 genes displayed identical topologies with all branches having maximal bootstrap percentage and posterior probabilities. Previous reconstruction, based on the 16S rRNA gene sequences, placed *Ca. P. novymonadis* in the very crown of the *Pandoraea* clade, though with a low support (Kostygov et al., 2016). However, the results presented here, which are based on much more extensive dataset, demonstrate this species to be an early branch diverged next to *P. thiooxidans* (**Figure 1**). The same position of *Ca. P. novymonadis* could be observed in analyses using either 556 genes supermatrix, or concatenated 16S rRNA and 23S rRNA genes, or a popular bacterial marker, *gyrB* (data not shown). As compared to other *Pandoraea* spp., the species under study demonstrated significantly longer branch (**Figure 1**). This is related to multiple amino acid substitutions in conserved sites and may be explained by fast adaptive evolution of this species. However, in comparison with the outgroups *B. cepacia* and *C. basilensis*, the branch of *Ca. P. novymonadis* does not appear to be uniquely long (**Figure 1**).

Analysis of Protein Orthologous Groups

We performed OrthoFinder analysis on a dataset of 23 annotated bacterial genomes (**Supplementary Table S1**). This resulted in 12,248 OGs, of which 5,437 contained only one protein. Similarly to the *Ca. Kinetoplastibacterium* spp. (Alves et al., 2013b), the genome of *Ca. P. novymonadis* is minimized and the vast majority of genes are single-copy: we found only five OGs containing two proteins with the sequence identity varying from 36 to 96%. These proteins were annotated as ATP-dependent RNA helicase, NADP⁺ reductase, BolA family transcriptional regulator, alanine-tRNA ligase, and threonine synthase. According to our analysis, ATP-dependent RNA helicase and NADP⁺ reductase were also duplicated in the genomes of several *Ca. Kinetoplastibacterium* spp. This situation is drastically different from that observed in the free-living *Pandoraea* spp., which have a substantially higher number of OGs containing two or more genes (e.g., 338 OGs in *P. apista* and 324 in *P. pnomensusa*).

We mapped gene family gains and losses on the phylogenomic tree (**Figure 1**). Gene loss is a predominant trend for all the leaves and most of the nodes within the *Pandoraea* clade. It is especially

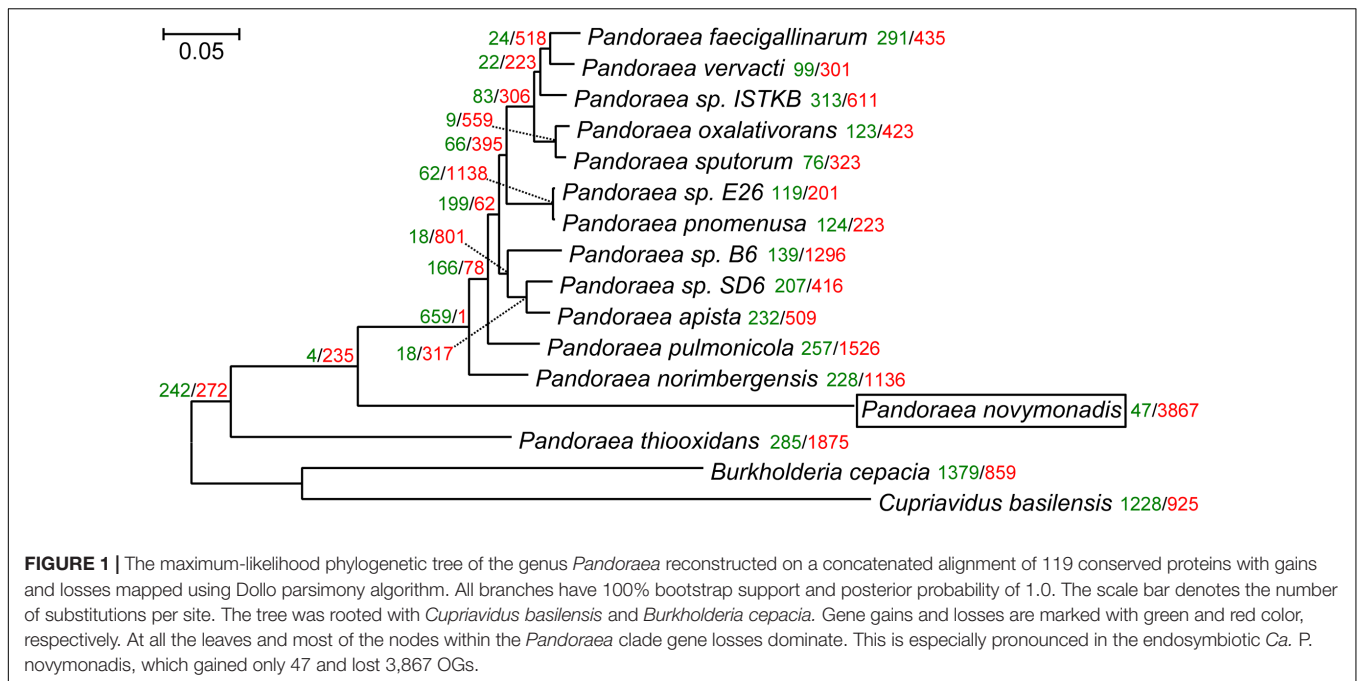
pronounced in the endosymbiotic *Ca. P. novymonadis*, which gained only 47 and lost 3,867 OGs. We used a sensitive HHpred tool attempting to illuminate functions of the proteins within OGs specific for *Ca. P. novymonadis* (Supplementary Table S3). Only 9 out of 47 proteins could be annotated using an *E*-value cut-off of 1. The following putative domains were identified: histidine kinase-like ATPase, cytoplasmic E component of the type III secretion system needle, myristoyl-CoA:protein *N*-myristoyltransferase, and carbohydrate binding domain.

We grouped gene annotations for the 3,867 OGs lost in *Ca. P. novymonadis* according to the KO system. Most of them belong to the following categories: “environmental information processing,” “amino acid metabolism,” “carbohydrate metabolism,” “genetic information processing,” “xenobiotics biodegradation,” and “energy metabolism” (**Supplementary Figure S7**). Out of 3,867 OGs, 1,273 were uniquely lost in *Ca. P. novymonadis*. The composition of functional categories assigned to the proteins within these OGs according to the KO system is similar to that assigned to all 3,867 OGs lost in *Ca. P. novymonadis*. However, the proportion of proteins belonging to the categories “genetic information processing,” “energy metabolism,” and “lipid metabolism” is increased in case of the annotations of OGs uniquely lost in *Ca. P. novymonadis*. The largest portion of OGs lost in *Ca. P. novymonadis* belong to the functional category “environmental information processing,” and more specifically “ATP-binding cassette transporters (ABC transporters).” *Ca. P. novymonadis* has lost many members of this protein family as compared to free-living *Pandoraea* spp.: mineral and organic ion transporters (e.g., for sulfate, nitrate, taurine, molybdate), monosaccharide transporters (e.g., for glycerol-3-phosphate), phosphate and amino acid transporters (e.g., for phosphate, phosphonate, glutamate, aspartate, cystine, urea, D-methionine), and transporters for glutathione and lipooligosaccharides.

Interestingly, there were no OGs uniquely shared between *Ca. P. novymonadis* and either of the endosymbiotic bacterial species investigated here (**Supplementary Figure S8**).

Lipid Metabolism

We identified a full set of enzymes essential for the type-II fatty acid synthesis (FAS) in *Ca. P. novymonadis* and other *Pandoraea* spp., *Ca. Kinetoplastibacterium* spp., *C. basilensis*, *B. cepacia*, and *Polynucleobacter* spp. (Supplementary Table S4). Acetyl-CoA carboxylase, the starting enzyme of the type-II FAS, in bacteria is composed of several polypeptides encoded by four distinct genes: *accA*, *accB*, *accC*, and *accD*. The *accB* and *accC* genes in *Ca. P. novymonadis* are located adjacent to each other and belong to the same operon, similarly to the situation observed in *Escherichia coli* (Janssen and Steinbuchel, 2014). FabF and FabH, 3-ketoacyl-acyl-carrier-protein (ACP) synthases II and III, which catalyze the formation of 3-ketoacyl-ACP by condensation of fatty acyl-ACP with malonyl-ACP, are present, while 3-ketoacyl-ACP synthase II (FabB) is absent in all the analyzed genomes, except for *P. oxalativorans* and *P. vervacti*. FabB participates in the synthesis of unsaturated fatty acids (FAs), catalyzing the condensation of *cis*-3-decenoyl-ACP (formed by the FabA catalyzed reaction), *cis*-5-dodecenoyl-ACP,



and *cis*-7-tetradecenoyl-ACP with malonyl-ACP (Feng and Cronan, 2009). 3-Hydroxydecanoyl-ACP dehydratase/isomerase (FabA), another key player in the synthesis of unsaturated FA is also missing from the analyzed genomes. Interestingly, *C. basilensis* possesses three different enoyl-ACP reductases, catalyzing the last step of the elongation cycle in the synthesis of FA: FabI, FabK, and FabV (Massengo-Tiassé and Cronan, 2009). *Ca. P. novymonadis*, *Ca. Kinetoplastibacterium* spp., and *Polynucleobacter* spp. have only FabI-encoding gene. The majority of the free-living *Pandoraea* spp. retain only FabV, while *B. cepacia*, *P. norimbergensis*, *P. oxalativorans*, *P. pulmonicola*, and *P. thiooxidans* retain FabK along with FabV. The physiological rationale for the presence of multiple enoyl-ACP reductases is poorly understood (Zhu et al., 2013).

All *Pandoraea* spp., *Polynucleobacter* spp., *C. basilensis*, and *B. cepacia* are able to synthesize cardiolipin, phosphatidylethanolamine, and phosphatidyl-L-serine, important components of the bacterial membranes (Supplementary Table S5). In all bacteria analyzed, the end product of the FA biosynthesis, acyl-ACP, can be activated with an inorganic phosphate group by the action of the PlsX component of the PlsX/PlsY/PlsC acyltransferase system, leading to acyl-phosphate, which is subsequently added to glycerol-3-phosphate by the action of the PlsY component (Janssen and Steinbuchel, 2014). The next steps to synthesize diacylglycerol-3-phosphate and cytosine diphosphate diacylglycerol (CDP-diacylglycerol) are performed by 1-acyl-sn-glycerol-3-phosphate acyltransferase (PlsC) and phosphatidate cytidyltransferase (CdsA). CDP-diacylglycerol is the intermediate which is then used for the formation of cardiolipin, phosphatidyl-L-serine, and phosphatidylethanolamine by cardiolipin synthase, CDP-diacylglycerol-serine O-phosphatidyltransferase, and

phosphatidylserine decarboxylase, respectively. All *Ca. Kinetoplastibacterium* spp. lack the capacity to synthesize cardiolipin, while *Ca. K. galatii*, *Ca. K. oncopeltii*, and *Ca. K. blastocrithidii* are not able to produce any of the membrane lipids mentioned above.

Interestingly, no lipases and esterases could be detected in the genome of *Ca. P. novymonadis* even with the *E*-value cutoff of 10^{-10} . We found proteins belonging to the family VI of bacterial lipolytic enzymes in all *Ca. Kinetoplastibacterium* spp. and in *P. necessarius* (Arpigny and Jaeger, 1999). The lipases and esterases belonging to the families I, IV, V, and VI are readily identifiable in the genomes of *C. basilensis* and *B. cepacia*, as well in the free-living *Pandoraea* spp., which in addition possess proteins belonging to the family VII of the lipolytic enzymes.

Importantly, all endosymbionts of trypanosomatids, including *Ca. P. novymonadis*, are unable to oxidize FAs since all the enzymes required for β -oxidation are missing, similarly to the situation observed in bacterial endosymbionts of insects (Zientz et al., 2004).

Carbon Metabolism

All species analyzed in this work preserve enzymes for glycolysis and the central (non-oxidative) part of the pentose phosphate pathway (Supplementary Figure S9). However, only the free-living *Pandoraea* spp. have hexokinase and, thus, are able to utilize glucose. In contrast to the endosymbiotic bacteria, they also can use classic and alternative (i.e., non-phosphorylated) variants of the Entner–Doudoroff pathway. Interestingly, only *P. thiooxidans* possesses phosphofructokinase converting fructose-6-phosphate into fructose 1,6-bisphosphate. Other species must use a bypass through the pentose phosphate pathway for hexose catabolism. Fructose 1,6-bisphosphatase, the

enzyme catalyzing the reverse reaction, is present in all studied species suggesting its importance for anabolic processes, in particular, gluconeogenesis.

We were unable to trace the carbon source that *Ca. Kinetoplastibacterium* spp. utilize instead of glucose. However, for *Ca. P. novyimonadis* and *P. necessarius* this appears to be fructose. Similarly to the situation with glucose, there is no typical phosphorylating enzyme, i.e., fructokinase (it is also absent from all other *Pandoraea* spp.). In all these species we identified the three cytoplasmic components of phosphotransferase system (PTS), namely phosphoenolpyruvate (PEP)-protein phosphotransferase (PTS-EI), histidine phosphocarrier protein (HPr), and PTS system fructose-specific EIIA component (PTS-EIIA^{Fru}). The main function of PTS is a concomitant transfer of sugars inside the cell and their phosphorylation (Saier, 2015). In addition to the three proteins mentioned above, the fully functional PTS must also contain juxtamembrane permease PTS-EIIB and transmembrane PTS-EIIC (sometimes along with PTS-EIID). The phosphate from PEP is successively transferred to PTS-EI, then to HPr, PTS-EIIA, PTS-EIIB, and then to sugar (Saier, 2015). Numerous proteobacteria possess incomplete PTS-lacking EIIB and EIIC components. Such PTSs were proposed to have only regulatory functions (Deutscher et al., 2014). We hypothesize that the incomplete fructose-specific PTS may be used for phosphorylation of fructose. Indeed, in addition to the abovementioned lack of fructokinase, *Ca. P. novyimonadis* also does not have pyruvate kinase, the key enzyme for the production of ATP from PEP at the end of glycolysis. Meanwhile, PTS using PEP as a phosphate donor could substitute this missing link. The lack of hexokinase and fructokinase along with the presence of PTS was also documented in obligate intracellular bacteria of insects (Zientz et al., 2004).

The complete tricarboxylic acid (TCA) cycle is present in all considered bacteria except *Ca. Kinetoplastibacterium* spp., which possess enzymes for two consecutive steps of this cycle: transformation of 2-oxoglutarate to succinyl-CoA and then to succinate. These steps may be preserved because succinyl-CoA is required for lysine biosynthesis. In addition, these bacteria possess malate dehydrogenase interconverting malate and oxaloacetate.

In addition to the TCA cycle, the free-living *Pandoraea* spp. also have the complete glyoxylate pathway, enabling usage of short-chain compounds as a carbon source. Endosymbiotic bacteria in their stable environment do not need such capability. Intriguingly, *P. necessarius* has malate synthase interconverting glyoxylate and malate, whereas other enzymes of this cycle are absent from its genome.

Amino Acid Metabolism

The free-living *Pandoraea* spp. are able to synthesize all 20 amino acids. Meanwhile, the three groups of endosymbionts considered here (*Ca. P. novyimonadis*, *P. necessarius*, and *Kinetoplastibacterium* spp.) demonstrate different phases of gradual loss of those capabilities (Figure 2 and Supplementary Table S6). This process starts with the loss of the pathways for the synthesis of the non-essential amino acids such as alanine, asparagine, and aspartate, a situation observed in

the evolutionary young endosymbiont *P. necessarius*. *Ca. P. novyimonadis* is unable to synthesize three additional amino acids: cysteine, methionine, and proline. *Ca. Kinetoplastibacterium* spp. exhibit the most advanced state, lacking enzymes for the synthesis of 13 amino acids (Figure 2). As judged from previous studies, the metabolic pathways of these endosymbionts and their hosts are interlaced and, for most of the amino acids, the enzymes missing in the bacterium can be substituted by those of the trypanosomatid (Alves et al., 2013a; Alves, 2017). Although the metabolism of *N. esmeraldas* has not been studied yet, it is likely similar to that of its relatives – trypanosomatids of the subfamily Leishmaniinae. This group of flagellates is auxotrophic for arginine, histidine, isoleucine, leucine, phenylalanine, serine, tryptophan, tyrosine, and valine (Opperdoes et al., 2016). Therefore, it is not surprising that *Ca. P. novyimonadis* retained the ability to synthesize them. In return, *N. esmeraldas* may provide the six amino acids, which its symbiont is unable to produce.

In addition to losing the ability to synthesize particular amino acids, the endosymbionts are devoid of some biochemical bypasses. Thus, phenylalanine-4-hydroxylase, converting phenylalanine to tyrosine, is present in free-living *Pandoraea* spp., but absent in all endosymbionts analyzed here. The same concerns arginase, the enzyme transforming arginine to ornithine (Figure 2).

Histidinol-phosphate phosphatase (HPPase), responsible for the penultimate step of histidine biosynthesis, was not found by BlastKOALA in any of the analyzed genomes. Nevertheless, HPPases are present in GenBank genome annotations of all free-living *Pandoraea* spp. Homologous proteins in *Polynucleobacter* spp. and *Ca. Kinetoplastibacterium* spp. are annotated as inositol monophosphatases. The same result was obtained for *Ca. P. novyimonadis* in Prokka annotation. It is known that inositol-monophosphatase-like enzymes may exhibit histidinol-phosphatase activity (Mormann et al., 2006; Petersen et al., 2010; Nourbakhsh et al., 2014). Of note, none of the bacteria analyzed here has other enzymes of inositol metabolism, so it is unlikely that the protein in question is an inositol monophosphatase. Thus, we argue that all analyzed species possess divergent histidinol-phosphatases.

Urea Cycle/Polyamine Synthesis

All free-living *Pandoraea* spp. have complete set of enzymes for the urea cycle and synthesis of important polyamines. *Ca. P. novyimonadis* and *P. necessarius* lack arginase, while preserving ornithine carbamoyltransferase, argininosuccinate synthase, and argininosuccinate lyase (Figure 3). They also possess arginine decarboxylase converting arginine to agmatine, the first intermediate in the synthesis of polyamines, but for the rest of this pathway these bacteria apparently rely on their respective hosts. *Ca. Kinetoplastibacterium* spp. showed the most reduced state with only two enzymes remaining in their arsenal: carbamoyltransferase and arginine decarboxylase (Figure 3).

Vitamins and Cofactors

All bacteria analyzed here are able to synthesize a number of porphyrins, including heme, an essential compound for

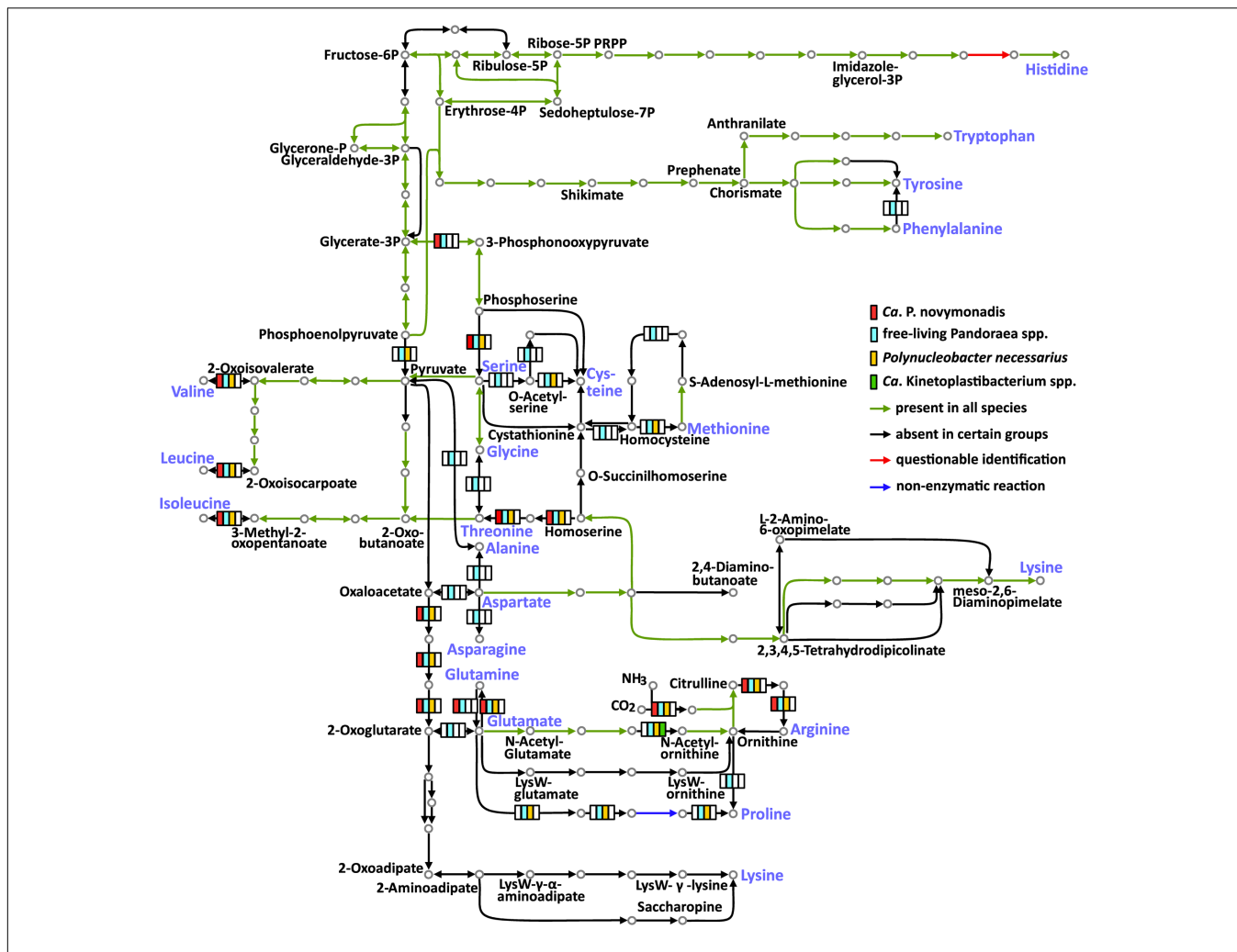


FIGURE 2 | Amino acid metabolism in *Ca. P. novymonadis*, free-living *Pandoraea* spp., *Polynucleobacter necessarius*, and *Ca. Kinetoplastibacterium* spp. The 20 standard amino acids are shown in violet.

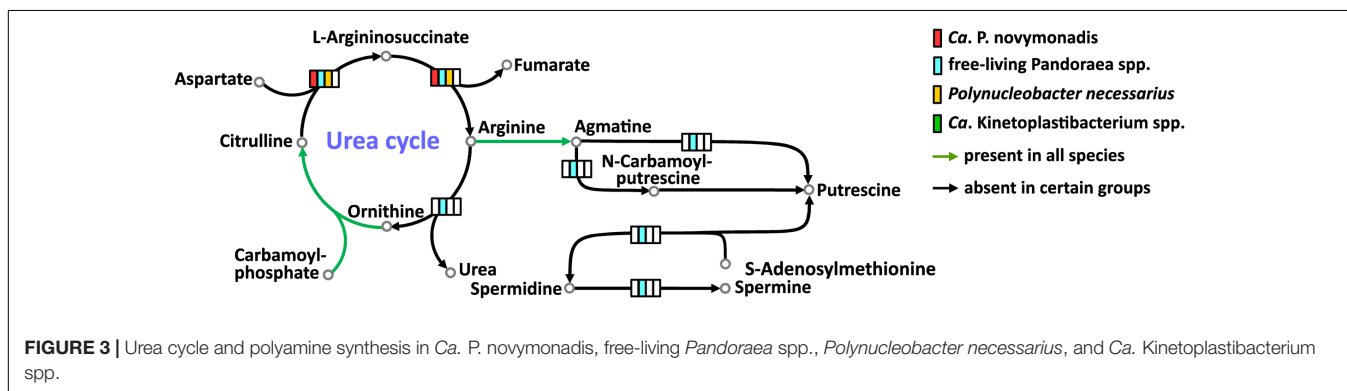


FIGURE 3 | Urea cycle and polyamine synthesis in *Ca. P. novymonadis*, free-living *Pandoraea* spp., *Polynucleobacter necessarius*, and *Ca. Kinetoplastibacterium* spp.

most trypanosomatids (Kořený et al., 2012). The free-living *Pandoraea* spp., *Ca. P. novymonadis* and *P. necessarius*, are prototrophic for all vitamins. As for *Ca. Kinetoplastibacterium* spp., their metabolism of vitamins was previously scrutinized by others (Klein et al., 2013). It has been demonstrated that in

contrast to the rest of bacteria considered here, they are unable to synthesize thiamine, nicotinic acid, and biotin, which are apparently acquired by the trypanosomatid host from the insect's gut content. All enzymes needed to produce folic acid, vitamin B6, and riboflavin essential for the trypanosomatid host are

encoded in the genomes of *Ca. Kinetoplastibacterium* spp., but the pathway of pantothenic acid biosynthesis is interrupted at the very end (Klein et al., 2013). The missing enzyme (ketopantoate reductase) is encoded in the genome of the trypanosomatid host, thus representing an example of deep integration of metabolic pathways in this symbiotic association.

CONCLUSION

Here, we sequenced and analyzed the genome of *Ca. P. novymonadis*, the bacterial endosymbiont of the trypanosomatid *N. esmeraldas*. To better understand the evolution and biology of this bacterium, we compared its genome to those of related prokaryotes, namely the free-living *Pandoraea* spp., two sister *Polynucleobacter* spp., from which one is free-living and the other is endosymbiotic, as well as with *Ca. Kinetoplastibacterium* spp., which are the only other known endosymbionts of trypanosomatids. The genome of *Ca. P. novymonadis* revealed all hallmarks of an endosymbiont genome: size reduction, massive gene losses, decreased GC content, and lowered codon usage bias. At the same time, this genome preserves main metabolic pathways, including biosynthesis of vitamins and heme, essential for the trypanosomatid host. The bacterium does not produce some amino acids, which are likely provided by the host, but retains the ability to synthesize those, for which the trypanosomatid is auxotrophic.

Our data allow first comparative analysis of the endosymbionts of trypanosomatids and strongly indicate that their evolution followed different scenarios, reflected by the fact that they do not have uniquely shared traits. Importantly, from the perspective of both its general genomic features and metabolism, *Ca. P. novymonadis* is closer to the ciliate-dwelling *P. necessarius*, which belongs to the same family Burkholderiaceae, than to *Ca. Kinetoplastibacterium* spp., the only other known endosymbionts of trypanosomatids.

Previously, we proposed that the endosymbiosis between *Ca. P. novymonadis* and *N. esmeraldas* was established relatively recently (Kostygov et al., 2016). This opinion was based on the phylogenetic position of the bacterium and seemingly unsophisticated relationships in this symbiotic association. However, the phylogenomic analysis presented here demonstrates that the endosymbiont diverged earlier than as inferred from its 16S rRNA gene. As judged from its genomic characteristics, *Ca. P. novymonadis* has already passed the intensive period of host adaptation, which can still be observed in *P. necessarius*, the best candidate for a recent endosymbiosis. As judged by the extent of the overall genome reduction and the loss of particular metabolic abilities correlating with the increasing dependence of the symbiont on its host, *Ca. P. novymonadis* occupies an intermediate position *P. necessarius* and *Ca. Kinetoplastibacterium* spp. We conclude that the relationship between *Ca. P. novymonadis* and *N. esmeraldas* is already well-established, although not as fine-tuned as in the case of related flagellates of the family Strigomonadinae and their endosymbionts.

AUTHOR CONTRIBUTIONS

VY and JL jointly conceived the study. AK and AB contributed equally to this work: participated in the design of the study, the analysis and interpretation of data, and the manuscript writing. AN and PF conducted genome assembly and curated annotation, and contributed to the interpretation of data. DT established and analyzed the aposymbiotic strain of *N. esmeraldas*. VY, AK, and JL revised and corrected the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the Grant Agency of Czech Republic awards 17-10656S to VY, 16-18699S to JL and VY, Moravskoslezský kraj research initiative DT01-021358 to VY and AK, the COST action CM1307, and the European Research Council CZ LL1601 to JL. Work in VY lab is financially supported by the Ministry of Education, Youth and Sports of the Czech in the “National Feasibility Program I,” project LO1208 “TEWEP.” AB was funded by grant from the University of Ostrava SGS16/PRF/2017. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://journal.frontiersin.org/article/10.3389/fmicb.2017.01940/full#supplementary-material>

FIGURE S1 | Graphs showing length distribution for genomic features in *Ca. P. novymonadis* and other bacteria with various lifestyles. (A,C) Distribution of gene lengths; (B,D) distribution of the lengths of intergenic regions. Pnov, *Ca. P. novymonadis*; Bcep, *B. cepacia*; Cbas, *C. basillensis*; Papi, *Pandoraea apista*; Pfae, *Pandoraea faecigallinarum*; Pnor, *Pandoraea norimbergensis*; Poxa, *Pandoraea oxalativorans*; Ppno, *Pandoraea pnomenusa*; Ppul, *Pandoraea pulmonicola*; Pspu, *Pandoraea sputorum*; Pthi, *Pandoraea thiooxidans*; Pver, *Pandoraea vervacti*; Kbla, *Ca. Kinetoplastibacterium blastocrithidii*; Kcri, *Ca. Kinetoplastibacterium crithidii*; Kdes, *Ca. Kinetoplastibacterium desouzai*; Kgal, *Ca. Kinetoplastibacterium galatii*; Konc, *Ca. Kinetoplastibacterium oncopeltii*; Pasy, *Polynucleobacter asymbioticus*; and Pnec, *Polynucleobacter necessarius*.

FIGURE S2 | Graphs showing GC content distribution for genomic features in *Ca. P. novymonadis* and other bacteria with various lifestyles: solid lines for genes and dotted lines for intergenic regions. Abbreviations of bacterial species names are as in **Supplementary Figure S1**; “ir” stands for intergenic regions.

FIGURE S3 | Nucleotide composition of protein-coding genes by codon positions. Abbreviations of bacterial species names are as in **Supplementary Figure S1**; nucleotides in a particular codon position are denoted below the figure.

FIGURE S4 | Amino acid frequencies. Abbreviations of bacterial species names are as in **Supplementary Figure S1**.

FIGURE S5 | Relationship of codon usage bias and GC content in protein coding genes averaged over the whole genome. The dotted line represents a sideways parabola fitting the distribution of values. Akas, *Advenella kashmirensis*; Tequ, *Taylorella equigenitalis*; Axyl, *Achromobacter xylosoxidans*;

Aspa, *Achromobacter spanius*; Ebro, *Bordetella bronchiseptica*; Bhol, *Bordetella holmesii*; Pnoe, *Fusillimonas noertemanni*; Pind, *Pelistega indica*; Oure, *Oligella urethralis*; Lsip, *Limnobacter* sp. CACIAM 66H1; Ggig, *Ca. Glomeribacter gigasporarum*; Csor, *Caballeronia sordidicola*. All other abbreviations of bacterial species names are as in **Supplementary Figure S1**.

FIGURE S6 | Schematic representation of two-way synteny between *Ca. P. novymonadis* and other *Pandoraea* spp. with sequenced genomes. The four longest *Ca. P. novymonadis* contigs are colored according to the legend. Only scaffolds with synteny blocks are shown. Direct synteny blocks are displayed in red, inverted ones – in green. The contigs are drawn proportionately to their actual length. The genomes of *Pandoraea* spp. shown on the figure are fully assembled to the level of circular chromosomes depicted as the longest colored bars. For some species the shorter colored bars representing plasmids are shown in addition to the chromosomal scaffolds.

REFERENCES

- Akhter, S., Aziz, R. K., and Edwards, R. A. (2012). PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. *Nucleic Acids Res.* 40, e126. doi: 10.1093/nar/gks406
- Alves, J. M., Klein, C. C., da Silva, F. M., Costa-Martins, A. G., Serrano, M. G., Buck, G. A., et al. (2013a). Endosymbiosis in trypanosomatids: the genomic cooperation between bacterium and host in the synthesis of essential amino acids is heavily influenced by multiple horizontal gene transfers. *BMC Evol. Biol.* 13:190. doi: 10.1186/1471-2148-13-190
- Alves, J. M., Serrano, M. G., Maia, da Silva, F., Voegtly, L. J., Matveyev, A. V., et al. (2013b). Genome evolution and phylogenomic analysis of *Candidatus* Kinetoplastibacterium, the betaproteobacterial endosymbionts of *Strigomonas* and *Angomonas*. *Genome Biol. Evol.* 5, 338–350. doi: 10.1093/gbe/evt012
- Alves, J. M., Voegtly, L., Matveyev, A. V., Lara, A. M., da Silva, F. M., Serrano, M. G., et al. (2011). Identification and phylogenetic analysis of heme synthesis genes in trypanosomatids and their bacterial endosymbionts. *PLOS ONE* 6:e23518. doi: 10.1371/journal.pone.0023518
- Alves, J. M. (2017). “Amino acid biosynthesis in endosymbiont-harboring Trypanosomatidae,” in *The Handbook of Microbial Metabolism of Amino Acids*, ed. J. P. F. D’Mello (Oxfordshire: CAB International), 371–383.
- Anandham, R., Indiragandhi, P., Kwon, S. W., Sa, T. M., Jeon, C. O., Kim, Y. K., et al. (2010). *Pandoraea thiooxydans* sp. nov., a facultatively chemolithotrophic, thiosulfate-oxidizing bacterium isolated from rhizosphere soils of sesame (*Sesamum indicum* L.). *Int. J. Syst. Evol. Microbiol.* 60(Pt 1), 21–26. doi: 10.1099/ijs.0.012823-0
- Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y., et al. (2016). PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* 44, W16–W21. doi: 10.1093/nar/gkw387
- Arpigny, J. L., and Jaeger, K. E. (1999). Bacterial lipolytic enzymes: classification and properties. *Biochem. J.* 343(Pt 1), 177–183. doi: 10.1042/bj3430177
- Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021
- Barry, A., Bryskier, A., Traczewski, M., and Brown, S. (2004). Preparation of stock solutions of macrolide and ketolide compounds for antimicrobial susceptibility tests. *Clin. Microbiol. Infect.* 10, 78–83. doi: 10.1111/j.1469-0691.2004.00759.x
- Boscaro, V., Felletti, M., Vannini, C., Ackerman, M. S., Chain, P. S., Malfatti, S., et al. (2013). *Polynucleobacter necessarius*, a model for genome reduction in both free-living and symbiotic bacteria. *Proc. Natl. Acad. Sci. U.S.A.* 110, 18590–18595. doi: 10.1073/pnas.1316687110
- Boscaro, V., Kolisko, M., Felletti, M., Vannini, C., Lynn, D. H., and Keeling, P. J. (2017). Parallel genome reduction in symbionts descended from closely related free-living bacteria. *Nat. Ecol. Evol.* 1, 1160–1167. doi: 10.1038/s41559-017-0237-0
- Botzman, M., and Margalit, H. (2011). Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles. *Genome Biol.* 12:R109. doi: 10.1186/gb-2011-12-10-r109
- FIGURE S7** | A pie chart showing gene annotations for 3,867 OGs lost in *Ca. P. novymonadis* (Pnov) categorized according to the KEGG Orthology system. *Pandoraea pnomenus* (Ppno), *Pandoraea norimbergensis* (Pnor), and *Pandoraea vervacti* (Pver) are shown for comparison.
- FIGURE S8** | Analysis of OGs sharing between free-living *Pandoraea* spp., *Ca. Pandoraea novymonadis*, *Ca. Kinetoplastibacterium* spp., *Polynucleobacter* spp., *Burkholderia cepacia*, and *Cupriavidus basilensis*. OGs were categorized according to their presence in the analyzed species. Depicted bars indicate number of OGs that are unique or shared among the genomes of the organisms listed, as indicated by the black dots.
- FIGURE S9** | Carbon metabolism in *Ca. P. novymonadis*, free-living *Pandoraea* spp., *Polynucleobacter necessarius*, and *Ca. Kinetoplastibacterium* spp.
- TABLE S1** | Genomic characteristics of species used in analyses.
- Burke, G. R., and Moran, N. A. (2011). Massive genomic decay in *Serratia symbiotica*, a recently evolved symbiont of aphids. *Genome Biol. Evol.* 3, 195–208. doi: 10.1093/gbe/evr002
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. doi: 10.1186/1471-2105-10-421
- Carryn, S., Chanteux, H., Seral, C., Mingeot-Leclercq, M. P., Van Bambeke, F., and Tulkens, P. M. (2003). Intracellular pharmacodynamics of antibiotics. *Infect. Dis. Clin. North Am.* 17, 615–634. doi: 10.1016/S0891-5520(03)00066-7
- Catta-Preta, C. M., Brum, F. L., da Silva, C. C., Zuma, A. A., Elias, M. C., de Souza, W., et al. (2015). Endosymbiosis in trypanosomatid protozoa: the bacterium division is controlled during the host cell cycle. *Front. Microbiol.* 6:520. doi: 10.3389/fmicb.2015.00520
- Coenye, T., Falsen, E., Hoste, B., Ohlen, M., Goris, J., Govan, J. R., et al. (2000). Description of *Pandoraea* gen. nov. with *Pandoraea apista* sp. nov., *Pandoraea pulmonicola* sp. nov., *Pandoraea pnomenus* sp. nov., *Pandoraea sputorum* sp. nov. and *Pandoraea norimbergensis* comb. nov. *Int. J. Syst. Evol. Microbiol.* 50(Pt 2), 887–899. doi: 10.1099/00207713-50-2-887
- Csuros, M. (2010). Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* 26, 1910–1912. doi: 10.1093/bioinformatics/btq315
- de Souza, W., and Motta, M. C. (1999). Endosymbiosis in protozoa of the Trypanosomatidae family. *FEMS Microbiol. Lett.* 173, 1–8. doi: 10.1111/j.1574-6968.1999.tb13477.x
- Deutscher, J., Ake, F. M., Derkaoui, M., Zebre, A. C., Cao, T. N., Bouraoui, H., et al. (2014). The bacterial phosphoenolpyruvate:carbohydrate phosphotransferase system: regulation by protein phosphorylation and phosphorylation-dependent protein-protein interactions. *Microbiol. Mol. Biol. Rev.* 78, 231–256. doi: 10.1128/MMBR.00001-14
- Du, Y., Maslov, D. A., and Chang, K. P. (1994). Monophyletic origin of beta-division proteobacterial endosymbionts and their coevolution with insect trypanosomatid protozoa *Blastocrithidia culicis* and *Crithidia* spp. *Proc. Natl. Acad. Sci. U.S.A.* 91, 8437–8441. doi: 10.1073/pnas.91.18.8437
- Emms, D. M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 16, 157. doi: 10.1186/s13059-015-0721-2
- Feng, Y., and Cronan, J. E. (2009). *Escherichia coli* unsaturated fatty acid synthesis: complex transcription of the *fabA* gene and *in vivo* identification of the essential reaction catalyzed by FabB. *J. Biol. Chem.* 284, 29526–29535. doi: 10.1074/jbc.M109.023440
- Flegontov, P., Butenko, A., Firsov, S., Kraeva, N., Eliáš, M., Field, M. C., et al. (2016). Genome of *Leptomonas pyrrocoris*: a high-quality reference for monoxenous trypanosomatids and new insights into evolution of *Leishmania*. *Sci. Rep.* 6:23704. doi: 10.1038/srep23704
- Graff, A., and Stubner, S. (2003). Isolation and molecular characterization of thiosulfate-oxidizing bacteria from an Italian rice field soil. *Syst. Appl. Microbiol.* 26, 445–452. doi: 10.1078/072320203322497482
- Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUASt: quality assessment tool for genome assemblies. *Bioinformatics* 29, 1072–1075. doi: 10.1093/bioinformatics/btt086

- Janssen, H. J., and Steinbuchel, A. (2014). Fatty acid synthesis in *Escherichia coli* and its applications towards the production of fatty acid based biofuels. *Biotechnol. Biofuels* 7:7. doi: 10.1186/1754-6834-7-7
- Jeong, S. E., Lee, H. J., Jia, B., and Jeon, C. O. (2016). *Pandoraea terrae* sp. nov., isolated from forest soil, and emended description of the genus *Pandoraea* Coenye et al., 2000. *Int. J. Syst. Evol. Microbiol.* 66, 3524–3530. doi: 10.1099/ijsem.0.001229
- Kalyanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermini, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589. doi: 10.1038/nmeth.4285
- Kanehisa, M. (2017). Enzyme annotation and metabolic reconstruction using KEGG. *Methods Mol. Biol.* 1611, 135–145. doi: 10.1007/978-1-4939-7015-5_11
- Kanehisa, M., Sato, Y., and Morishima, K. (2016). BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J. Mol. Biol.* 428, 726–731. doi: 10.1016/j.jmb.2015.11.006
- Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010
- Kawai, Y., and Maeda, Y. (2009). GC-content of tRNA genes classifies archaea into two groups. *J. Gen. Appl. Microbiol.* 55, 403–408. doi: 10.2323/jgam.55.403
- Klein, C. C., Alves, J. M., Serrano, M. G., Buck, G. A., Vasconcelos, A. T., Sagot, M. F., et al. (2013). Biosynthesis of vitamins and cofactors in bacterium-harbouring trypanosomatids depends on the symbiotic association as revealed by genomic analyses. *PLOS ONE* 8:e79786. doi: 10.1371/journal.pone.0079786
- Kořený, L., Sobotka, R., Kovařová, J., Gnypová, A., Flegontov, P., Horváth, A., et al. (2012). Aerobic kinetoplastid flagellate *Phytomonas* does not require heme for viability. *Proc. Natl. Acad. Sci. U.S.A.* 109, 3808–3813. doi: 10.1073/pnas.1201089109
- Kostygov, A. Y., Dobaková, E., Grybchuk-Ieremenko, A., Váhala, D., Maslov, D. A., Votýpka, J., et al. (2016). Novel trypanosomatid-bacterium association: evolution of endosymbiosis in action. *mBio* 7:e01985-15. doi: 10.1128/mBio.01985-15
- Kumar, M., Singh, J., Singh, M. K., Singhal, A., and Thakur, I. S. (2015). Investigating the degradation process of kraft lignin by beta-proteobacterium, *Pandoraea* sp. ISTKB. *Environ. Sci. Pollut. Res. Int.* 22, 15690–15702. doi: 10.1007/s11356-015-4771-5
- Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054
- Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923
- Liz, J. A. Z. E., Jan-Roblero, J., de la Serna, J. Z. D., de Leon, A. V. P., and Hernandez-Rodriguez, C. (2009). Degradation of polychlorinated biphenyl (PCB) by a consortium obtained from a contaminated soil composed of *Brevibacterium*, *Pandoraea* and *Ochrobactrum*. *World J. Microbiol. Biotechnol.* 25, 165–170. doi: 10.1007/s11274-008-9875-3
- Massengo-Tiassé, R. P., and Cronan, J. E. (2009). Diversity in enoyl-acyl carrier protein reductases. *Cell. Mol. Life Sci.* 66, 1507–1517. doi: 10.1007/s00018-009-8704-7
- Maurin, M., and Raoult, D. (2001). Use of aminoglycosides in treatment of infections due to intracellular bacteria. *Antimicrob. Agents Chemother.* 45, 2977–2986. doi: 10.1128/AAC.45.11.2977-2986.2001
- McCutcheon, J. P., and Moran, N. A. (2011). Extreme genome reduction in symbiotic bacteria. *Nat. Rev. Microbiol.* 10, 13–26. doi: 10.1038/nrmicro2670
- Meincke, L., Copeland, A., Lapidus, A., Lucas, S., Berry, K. W., Del Rio, T. G., et al. (2012). Complete genome sequence of *Polynucleobacter necessarius* subsp. *asymbioticus* type strain (QLW-PIDMWA-1^T). *Stand. Genomic Sci.* 6, 74–83. doi: 10.4056/sigs.2395367
- Minh, B. Q., Nguyen, M. A., and von Haeseler, A. (2013). Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.* 30, 1188–1195. doi: 10.1093/molbev/mst024
- Moran, N. A., McCutcheon, J. P., and Nakabachi, A. (2008). Genomics and evolution of heritable bacterial symbionts. *Annu. Rev. Genet.* 42, 165–190. doi: 10.1146/annurev.genet.41.110306.130119
- Mormann, S., Lomker, A., Ruckert, C., Gaigalat, L., Tauch, A., Puhler, A., et al. (2006). Random mutagenesis in *Corynebacterium glutamicum* ATCC 13032 using an IS6100-based transposon vector identified the last unknown gene in the histidine biosynthesis pathway. *BMC Genomics* 7:205. doi: 10.1186/1471-2164-7-205
- Motta, M. C., Catta-Preta, C. M., Schenkman, S., de Azevedo Martins, A. C., Miranda, K., de Souza, W., et al. (2010). The bacterium endosymbiont of *Crithidia deanei* undergoes coordinated division with the host cell nucleus. *PLOS ONE* 5:e12415. doi: 10.1371/journal.pone.0012415
- Motta, M. C., Soares, M. J., Attias, M., Morgado, J., Lemos, A. P., Saad-Nehme, J., et al. (1997). Ultrastructural and biochemical analysis of the relationship of *Crithidia deanei* with its endosymbiont. *Eur. J. Cell Biol.* 72, 370–377.
- Moya, A., Pereto, J., Gil, R., and Latorre, A. (2008). Learning how to live together: genomic insights into prokaryote-animal symbioses. *Nat. Rev. Genet.* 9, 218–229. doi: 10.1038/nrg2319
- Nguyen, L. T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300
- Nourbakhsh, A., Collakova, E., and Gillaspay, G. E. (2014). Characterization of the inositol monophosphatase gene family in *Arabidopsis*. *Front. Plant Sci.* 5:725. doi: 10.3389/fpls.2014.00725
- Nowack, E. C., and Melkonian, M. (2010). Endosymbiotic associations within protists. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 365, 699–712. doi: 10.1098/rstb.2009.0188
- Ochman, H., and Davalos, L. M. (2006). The nature and dynamics of bacterial genomes. *Science* 311, 1730–1733. doi: 10.1126/science.1119966
- Okeke, B. C., Siddique, T., Arbestain, M. C., and Frankenberger, W. T. (2002). Biodegradation of gamma-hexachlorocyclohexane (lindane) and alpha-hexachlorocyclohexane in water and a soil slurry by a *Pandoraea* species. *J. Agric. Food Chem.* 50, 2548–2555. doi: 10.1021/jf011422a
- Oppendoes, F. R., Butenko, A., Flegontov, P., Yurchenko, V., and Lukes, J. (2016). Comparative metabolism of free-living *Bodo saltans* and parasitic trypanosomatids. *J. Eukaryot. Microbiol.* 63, 657–678. doi: 10.1111/jeu.12315
- Ozaki, S., Kishimoto, N., and Fujita, T. (2007). Change in the predominant bacteria in a microbial consortium cultured on media containing aromatic and saturated hydrocarbons as the sole carbon source. *Microbes Environ.* 22, 128–135. doi: 10.1264/jmsme.2.22.128
- Petersen, L. N., Marineo, S., Mandala, S., Davids, F., Sewell, B. T., and Ingle, R. A. (2010). The missing link in plant histidine biosynthesis: *Arabidopsis myoinositol monophosphatase-like2* encodes a functional histidinol-phosphate phosphatase. *Plant Physiol.* 152, 1186–1196. doi: 10.1104/pp.109.150805
- Petersen, T. N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* 8, 785–786. doi: 10.1038/nmeth.1701
- Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033
- Rice, P., Longden, I., and Bleasby, A. (2000). EMBOS: the European molecular biology open software suite. *Trends Genet.* 16, 276–277. doi: 10.1016/S0168-9525(00)02024-2
- Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., et al. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61, 539–542. doi: 10.1093/sysbio/sy029
- Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A., et al. (2000). Artemis: sequence visualization and annotation. *Bioinformatics* 16, 944–945. doi: 10.1093/bioinformatics/16.10.944
- Sahin, N., Tani, A., Kotan, R., Sedlacek, I., Kimbara, K., and Tamer, A. U. (2011). *Pandoraea oxalativorans* sp. nov. *Pandoraea faecigallinarum* sp. nov. and *Pandoraea vervacti* sp. nov., isolated from oxalate-enriched culture. *Int. J. Syst. Evol. Microbiol.* 61(Pt 9), 2247–2253. doi: 10.1099/ijso.026138-0
- Saier, M. H. Jr. (2015). The bacterial phosphotransferase system: new frontiers 50 years after its discovery. *J. Mol. Microbiol. Biotechnol.* 25, 73–78. doi: 10.1159/000381215
- Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi: 10.1093/bioinformatics/btu153
- Sharp, P. M., Bailes, E., Grocock, R. J., Peden, J. F., and Sockett, R. E. (2005). Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res.* 33, 1141–1153. doi: 10.1093/nar/gki242

- Sharp, P. M., Emery, L. R., and Zeng, K. (2010). Forces that influence the evolution of codon bias. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 365, 1203–1212. doi: 10.1098/rstb.2009.0305
- Sharp, P. M., Tuohy, T. M., and Mosurski, K. R. (1986). Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res.* 14, 5125–5143. doi: 10.1093/nar/14.13.5125
- Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi: 10.1093/bioinformatics/btv351
- Soderlund, C., Bomhoff, M., and Nelson, W. M. (2011). SyMAP v3.4: a turnkey synteny system with application to plant genomes. *Nucleic Acids Res.* 39, e68. doi: 10.1093/nar/gkr123
- Soding, J., Biegert, A., and Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* 33, W244–W248. doi: 10.1093/nar/gki408
- Stryjewski, M. E., LiPuma, J. J., Messier, R. H., Reller, L. B., and Alexander, B. D. (2003). Sepsis, multiple organ failure, and death due to *Pandoraea pnomenusa* infection after lung transplantation. *J. Clin. Microbiol.* 41, 2255–2257. doi: 10.1128/Jcm.41.5.2255-2257.2003
- Teixeira, M. M., Borghesan, T. C., Ferreira, R. C., Santos, M. A., Takata, C. S., Campaner, M., et al. (2011). Phylogenetic validation of the genera *Angomonas* and *Strigomonas* of trypanosomatids harboring bacterial endosymbionts with the description of new species of trypanosomatids and of proteobacterial symbionts. *Protist* 162, 503–524. doi: 10.1016/j.protis.2011.01.001
- Toh, H., Weiss, B. L., Perkin, S. A., Yamashita, A., Oshima, K., Hattori, M., et al. (2006). Massive genome erosion and functional adaptations provide insights into the symbiotic lifestyle of *Sodalis glossinidius* in the tsetse host. *Genome Res.* 16, 149–156. doi: 10.1101/gr.4106106
- Vannini, C., Ferrantini, F., Ristori, A., Verni, F., and Petroni, G. (2012). Betaproteobacterial symbionts of the ciliate *Euplotes*: origin and tangled evolutionary path of an obligate microbial association. *Environ. Microbiol.* 14, 2553–2563. doi: 10.1111/j.1462-2920.2012.02760.x
- Vannini, C., Ferrantini, F., Verni, F., and Petroni, G. (2013). A new obligate bacterial symbiont colonizing the ciliate *Euplotes* in brackish and freshwater: '*Candidatus* Protistobacter heckmanni'. *Aquat. Microb. Ecol.* 70, 233–243. doi: 10.3354/ame01657
- Vannini, C., Pockl, M., Petroni, G., Wu, Q. L., Lang, E., Stackebrandt, E., et al. (2007). Endosymbiosis in statu nascendi: close phylogenetic relationship between obligately endosymbiotic and obligately free-living *Polynucleobacter* strains (*Betaproteobacteria*). *Environ. Microbiol.* 9, 347–359. doi: 10.1111/j.1462-2920.2006.01144.x
- Votýpka, J., Kostygov, A. Y., Kraeva, N., Grybchuk-Ieremenko, A., Tesařová, M., Grybchuk, D., et al. (2014). *Kentomonas* gen. n., a new genus of endosymbiont-containing trypanosomatids of Strigomonadinae subfam. n. *Protist* 165, 825–838. doi: 10.1016/j.protis.2014.09.002
- Zhu, L., Bi, H., Ma, J., Hu, Z., Zhang, W., Cronan, J. E., et al. (2013). The two functional enoyl-acyl carrier protein reductases of *Enterococcus faecalis* do not mediate triclosan resistance. *mBio* 4:e00613-13. doi: 10.1128/mBio.00613-13
- Zientz, E., Dandekar, T., and Gross, R. (2004). Metabolic interdependence of obligate intracellular bacteria and their insect hosts. *Microbiol. Mol. Biol. Rev.* 68, 745–770. doi: 10.1128/MMBR.68.4.745-770.2004

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Kostygov, Butenko, Nenarokova, Tashyreva, Flegontov, Lukeš and Yurchenko. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

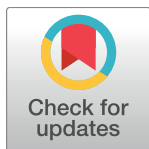
PEARLS

A paradigm shift: The mitoproteomes of procyclic and bloodstream *Trypanosoma brucei* are comparably complex

Alena Zíková^{1,2*}, Zdeněk Verner^{1,3}, Anna Nenarokova^{1,2}, Paul A. M. Michels⁴, Julius Lukeš^{1,2}

1 Institute of Parasitology, Biology Centre, Czech Academy of Sciences, České Budějovice, Czech Republic, **2** Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic, **3** Faculty of Sciences, Charles University, Prague, Czech Republic, **4** Centre for Immunity, Infection and Evolution, The University of Edinburgh, Edinburgh, United Kingdom

* azikova@paru.cas.cz



Metabolic adaptation during *Trypanosoma brucei*'s life cycle

Trypanosoma brucei is a parasitic protist that causes significant health burden in sub-Saharan countries endemic for the tsetse fly (*Glossina* spp.). During the bloodmeal of this insect vector, the flagellate is transmitted to a variety of mammals, including humans, in which *T. brucei* subs. *gambiense* and *T. brucei* subs. *rhodesiense* cause human African trypanosomiasis. During its life cycle, *T. brucei* encounters and adapts to very diverse environments that differ in available nutrients. In the mammalian host, it exists in two major forms: the replicating long-slender bloodstream form (LS-BSF) and the nondividing short-stumpy bloodstream form (SS-BSF), the latter being pre-adapted to infect the insect vector [1]. While the BSF flagellates primarily colonize the mammalian bloodstream and utilize the plentiful glucose for their energy needs, they can also be found in the cerebrospinal fluid and in extracellular spaces of several tissues, including the brain, adipose tissue, and skin [2,3]. In the insect vector, trypanosomes occur in three major forms occupying different locations within the fly: the procyclic form (PCF) resides in the midgut and proventriculus, while epimastigotes and metacyclic trypanosomes are found in the salivary glands. During the fly's bloodmeal, the latter form infects the mammalian host. All three forms experience the glucose-poor and amino acid-rich environment within the insect host. These drastic environmental changes encountered by *T. brucei* during its development require significant morphological and metabolic changes and adaptations [4,5].

The seminal work of Keith Vickerman led to the widely accepted model of a highly reduced mitochondrial metabolism in the BSF [6,7]. Its single mitochondrion is incapable of oxidative phosphorylation, and the active electron transport chain (ETC) is minimized to an alternative pathway composed of glycerol-3-phosphate dehydrogenase (Gly-3-PDH) and the so-called trypanosome alternative oxidase (AOX), which are linked to each other via a ubiquinol/ubiquinone pool [8]. The cytochrome-containing ETC is absent, and the mitochondrial transmembrane proton gradient is generated by the reverse activity of the F₀F₁-ATP synthase complex at the expense of ATP [9–11]. The proton gradient across the mitochondrial inner membrane is essential for protein import and transport of metabolites and ions so that vital mitochondrial processes such as Fe-S cluster assembly [12], RNA editing and processing [13,14], and cellular Ca²⁺ homeostasis are maintained [15,16]. The seemingly simplified biochemical composition of the BSF organelle is underlined by its tube-shaped cristae-poor

OPEN ACCESS

Citation: Zíková A, Verner Z, Nenarokova A, Michels PAM, Lukeš J (2017) A paradigm shift: The mitoproteomes of procyclic and bloodstream *Trypanosoma brucei* are comparably complex. PLoS Pathog 13(12): e1006679. <https://doi.org/10.1371/journal.ppat.1006679>

Editor: Laura J. Knoll, University of Wisconsin Medical School, UNITED STATES

Published: December 21, 2017

Copyright: © 2017 Zíková 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.

Funding: This work was supported by the ERC CZ awards LL1205 and LL1601 and by the Czech Grant Agency awards 17-22248S and 15-21974S to AZ and JL, respectively. ZV was supported by LQ1604 NPU II provided by MEYS. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

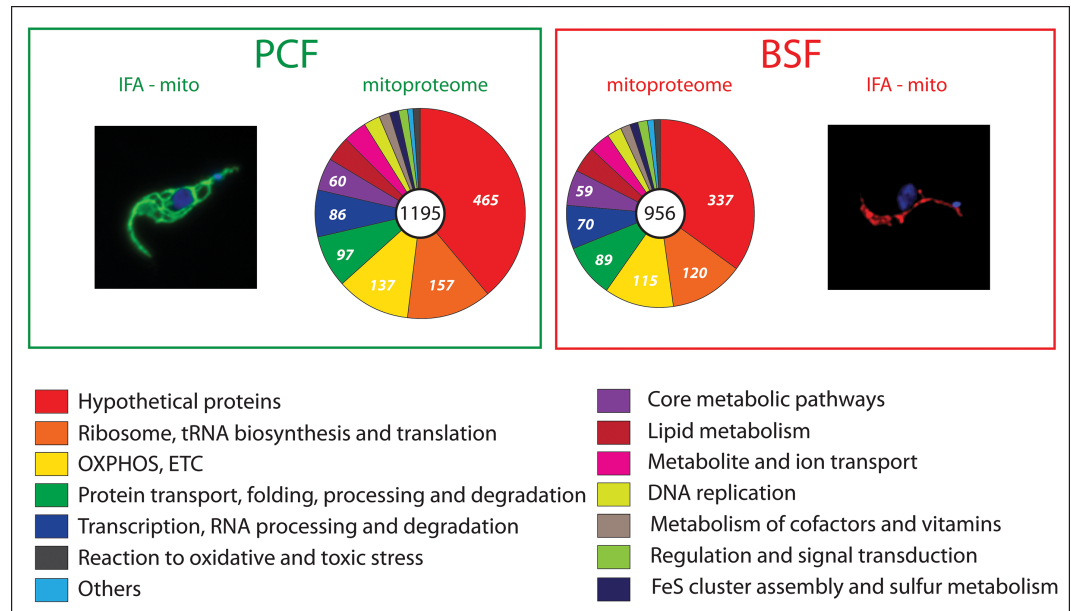


Fig 1. Pie charts showing distribution of mass spectrometry-identified mitochondrial proteins in PCF (left) and BSF (right) trypanosomes in terms of molecular functions. A total number of 1,195 and 956 proteins were assigned to PCF and BSF mitoproteome, respectively. Different colors show different metabolic pathways and categories. See also [S1 Table](#). IFA-mito in PCF (left) and in BSF cell (right). BSF, bloodstream form; Hsp70, heat shock protein 70; IFA-mito, immunofluorescence analysis of a mitochondrial Hsp70; mitoproteome, mitochondrial proteome; PCF, procyclic form.

<https://doi.org/10.1371/journal.ppat.1006679.g001>

morphology, which is in striking contrast to the extensively reticulated cristae-rich mitochondrion of the PCF flagellates (Fig 1).

Because no indications have been obtained yet for the presence of a mitochondrial ATP-producing system in the BSF, the entire cellular ATP pool is considered to be generated solely by highly active glycolysis [17]. The glycolytic pathway in trypanosomes is unique in the sense of sequestration of most of its enzymes within peroxisome-like organelles called glycosomes [18]. Because the glycosomal membrane is impermeable to large solutes like NAD(H), the essential reoxidation of glycolytically produced intraglycosomal NADH occurs by a shuttle mechanism involving the oxidation of glycerol 3-phosphate to dihydroxyacetone phosphate by the mitochondrial Gly-3-PDH [8].

Classical metabolic studies performed with trypanosomes purified from the blood of infected rodents or with in vitro-cultured BSF supported the original hypothesis of a drastically simplified mitochondrial metabolism because under aerobic conditions, glucose is almost completely catabolized to pyruvate that is excreted from the cells, indicating no need for the mitochondrial enzymes of the tricarboxylic acid cycle. In the absence of oxygen or when AOX is chemically inhibited, glycerol 3-phosphate is converted into glycerol that is produced in a 1:1 ratio with pyruvate [19,20]. Occasionally, the production of small amounts of other compounds such as acetate, succinate, and alanine has been reported; however, these products were instead attributed to the presence of a minor fraction of SS-BSF, a life cycle stage possessing a more elaborated metabolism, with some traits characteristic of the metabolically complex PCF [21].

In preparation for differentiation into PCF, the SS-BSF up-regulates a subset of mitochondrial and other proteins [21]. Moreover, these cells are metabolically active, motile, regulate their internal pH [22], and excrete end products of glucose metabolism in ratios different than

the LS-BSF and PCF cells [21]. Differentiation of LS-BSF into SS-BSF is triggered by the stumpy-inducing factor, and only pleiomorphic strains (e.g., AnTat 1.1) are able to sense this/ these yet-to-be-identified molecule(s) [23]. Extended passaging of pleiomorphic parasites in *in vitro* cultures or by syringe between laboratory animals leads to the loss of responsiveness to the stumpy-inducing factor and thus a failure to differentiate into SS-BSF. Consequently, such strains (e.g., Lister 427) are called monomorphic, i.e., they exist only as a single form [24].

Interestingly, recent analyses employing the monomorphic LS-BSF strain Lister 427 showed that, in addition to pyruvate, appreciable amounts of other carbon products (i.e., alanine, acetate, and succinate) are excreted into the cultivation medium [25], implying a need not only for cytosolic and glycosomal but also for mitochondrial enzymes thus far considered to be absent (Fig 2). An additional metabolomics study involving heavy-atom isotope-labeled glucose determined that a substantial fraction of succinate, as well as metabolic intermediates such as malate and fumarate, are glucose-derived and originate from phosphoenolpyruvate via oxaloacetate. Importantly, phosphoenolpyruvate carboxykinase, a glycosomal enzyme responsible for this conversion, is essential for the BSF parasites [26]. Moreover, the majority of excreted alanine and acetate is also derived from glucose. Alanine is most likely produced from pyruvate by the transamination reaction of alanine aminotransferase, a potentially essential enzyme [27], while glucose-derived acetate is produced from pyruvate by the mitochondrial pyruvate dehydrogenase (PDH) complex and additional subsequent enzymatic steps. A fraction of the acetate produced this way is exported to the cytosol for the *de novo* synthesis of fatty acids, which is an essential process (Fig 2) [25]. In addition to glucose, the BSF seems to uptake and metabolize amino acids such as cysteine, glutamine, phenylalanine, tryptophan, and threonine [28], implying the existence of an unexpectedly complex metabolic network in their mitochondrion.

The BSF mitoproteome

To map the BSF mitochondrial proteome (mitoproteome), we first used the available mass spectrometry data of purified PCF mitochondria [29–37] in order to assemble a comprehensive list of mitochondrial proteins. Next, we asked how many of these proteins were identified in any mass spectrometry data obtained from BSF cells [38–43]. To our surprise, out of 1,195 constituents of the PCF mitoproteome, 956 were also identified in at least one study of the BSF, suggesting that, when qualitatively measured, the corresponding mitoproteome is reduced by only approximately 20% (Fig 1; S1 Table). The surprisingly high, approximately 80% overlap with the PCF mitoproteome might also be a consequence of the heterogeneity of the examined BSF populations. The heterogeneity may be related to the experimental protocols, the environmental variations (cells grown *in vivo* versus *in vitro*), or variations within the cell cycle (e.g., ATP requirements vary between different cell cycle stages) as well as to the form type (monomorphic versus pleiomorphic). Indeed, some authors analyzed monomorphic strains grown *in vitro* [40,41], and others examined the pleiomorphic AnTat 1.1 strain grown either in immunosuppressed rats [43] or *in vitro* (S1 Table) [38]. Therefore, some LS-BSF cells may have a mitochondrion that is close to the “classical” version, while a subset of these flagellates may express an extended mitoproteome. However, no apparent differences were detected between the mitoproteomes from the pleiomorphic and monomorphic BSF cells, suggesting that, regardless of their status, a surprisingly large repertoire of mitochondrial proteins is expressed in the BSF stage.

All proteins were then organized into groups based on their Kyoto Encyclopedia of Genes and Genomes (KEGG) annotations. No striking qualitative differences were observed in the categories “oxidative phosphorylation” and “core metabolic pathways” comprising many

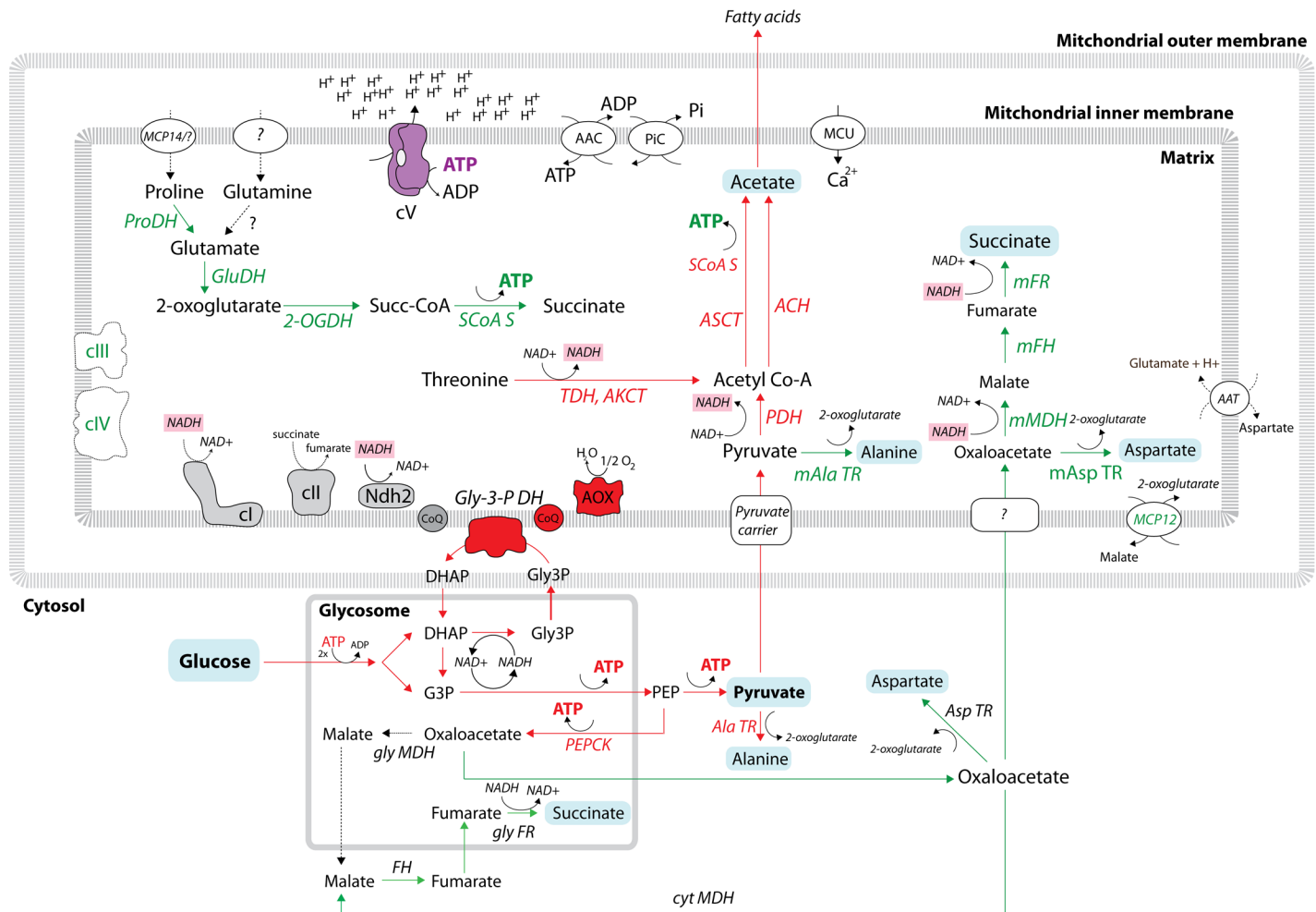


Fig 2. Schematic representation of carbon source metabolism in the bloodstream form of *T. brucei*. Red arrows represent enzymatic steps that were experimentally shown to be active in BSF. Green arrows represent enzymatic steps that might be active in BSF because the enzymes (in green) were identified in BSF proteomic data. Glucose-derived metabolites (acetate, pyruvate, succinate, alanine, aspartate) are on a blue background. NADH molecules are on a pink background. Dashed arrows indicate enzymatic steps for which no experimental proof exists. The glycosomal and mitochondrial compartments are indicated. 2-OGDH, 2-oxoglutarate dehydrogenase; AAC, ADP/ATP carrier; AAT, amino acid transporter; ACH, acetyl-CoA thioesterase; AKCT, 2-amino-3-ketobutyrate coenzyme A ligase; Ala TR, alanine transaminase; AOX, alternative oxidase; ASCT, acetate:succinate CoA-transferase; Asp TR, aspartate transaminase; BSF, bloodstream form; cI, complex I (NADH:ubiquinone oxidoreductase); cII, complex II (succinate dehydrogenase); cIII, complex III (cytochrome bc1 complex); cIV, complex IV (cytochrome c oxidase); cV, complex V (F_0F_1 ATPase); cyt, cytosolic; DHAP, dihydroxyacetone phosphate; FH, fumarate hydratase (i.e., fumarase); FR, fumarate reductase; G3P, glyceraldehyde 3-phosphate; GluDH, glutamate dehydrogenase; gly, glycosomal; Gly3P, glycerol 3-phosphate; Gly-3-PDH, glycerol-3-phosphate dehydrogenase; m, mitochondrial; MDH, malate dehydrogenase; PDH, pyruvate dehydrogenase; PEP, phosphoenolpyruvate; PEPCK, phosphoenolpyruvate carboxykinase; PIC, phosphate carrier; ProDH, proline dehydrogenase.

<https://doi.org/10.1371/journal.ppat.1006679.g002>

enzymes involved in the carbon, amino acid, and energy metabolism (Fig 1, S1 Table). Typical examples are components of the tricarboxylic acid cycle and subunits of the ETC complexes, most prominently of respiratory complexes III and IV (Fig 2, S1 Table). Nonetheless, when quantitative information was available, these proteins were often present in much lower amounts than in the PCF. While some of these proteins may not perform their expected function(s) under BSF steady state growth conditions, this finding strongly suggests that the parasite is capable of swift alterations or adjustments of its metabolism in response to various environments and differentiation cues. This ability can be exploited during environmental changes, for example when the LS-BSF migrates from the peripheral blood circulation to other

extravascular spaces (e.g., in adipose tissue, spinal and cerebral fluids) and during the differentiation to SS-BSF. Therefore, the BSF trypanosomes may uptake different substrates from the available nutrients according to their immediate needs and metabolize them via a variety of pathways.

Complex metabolic pathways in the BSF mitochondrion: Does presence equal activity?

The current metabolic model for BSF excludes a role of the mitochondrion in the ATP production by either oxidative or substrate-level phosphorylation [44]. In contrast to this premise, succinyl-CoA synthase (SCoAS), an enzyme responsible for substrate-level phosphorylation of ADP to ATP, has been detected in BSF cells, and more importantly, its RNA interference (RNAi)-mediated silencing produced a severe growth phenotype [45]. This enzyme can be involved in two ATP-producing pathways. The first one includes activity of 2-oxoglutarate dehydrogenase (2-OGDH) producing succinyl-CoA from 2-oxoglutarate that originates from amino acids such as proline and glutamine or can result from transamination reactions by mitochondrial alanine and aspartate transaminases (Fig 2). While all the enzymes involved in these reactions were detected in the LS-BSF mitoproteome (Fig 2 and S1 Table), the activity of 2-OGDH remains contradictory because some authors failed to detect it in the pleiomorphic cells [46], while others recorded its low activity in culture-adapted monomorphic LS-BSF cells [47]. Puzzlingly, the 2-OGDH subunits E1 and E2 were shown to be essential in BSF not because of their role in carbon metabolism but rather due to their moonlighting roles in glycosomes and mitochondrial DNA maintenance [46,48]. However, in an untargeted metabolomics study using isotope-labeled glucose, up to 30% of excreted succinate remained unlabeled, supporting its nonglucose origin [26] and making the occurrence of this substrate-level phosphorylation reaction even more plausible (Fig 2).

The second phosphorylation pathway includes the acetate:succinate CoA transferase/SCoAS cycle that contributes to acetate production in the BSF mitochondrion. A substrate for this reaction—acetyl-CoA—is produced by PDH, an enzymatic complex that is present and active in the BSF mitochondrion [15,25]. Moreover, PDH was shown to be indispensable for BSF cells but only in the absence of threonine because under these artificial conditions, PDH was the only system supplying acetyl-CoA for the essential acetate production [25]. Nonetheless, the mitochondrial pyruvate transporter was demonstrated to be essential for BSF *in vivo*, supporting PDH's vital role for the parasite [49]. These results imply that the BSF mitochondrion may become an ATP producer under certain conditions, perhaps just for intramitochondrial needs (Fig 2).

The presence and potential activity of the aforementioned dehydrogenases that produce NADH within the mitochondrion imply that the organelle would require reoxidation of this cofactor. Several possible scenarios for such a capacity can be deduced from the available data (S1 Table). Although complex I (NADH:ubiquinone oxidoreductase) was shown to be neither essential for BSF nor contributing to the observed NADH:ubiquinone oxidoreductase activity [50], it is still assembled in the BSF mitochondrion and may participate in NADH reoxidation under certain conditions. Reoxidation of reduced NADH molecules can also be achieved by the activity of the alternative dehydrogenase 2 (Ndh2), an enzyme shown to be important but not essential for maintaining the mitochondrial redox balance [51]. Last but not least, another scenario includes the activities of the mitochondrial malate dehydrogenase, fumarate hydratase (i.e., fumarase), and NADH-dependent fumarate reductase, with all three being present in the BSF mitoproteome (Fig 2, S1 Table). These enzymes reduce glucose-derived oxaloacetate via malate and fumarate to succinate. Indeed, 3-carbon-labeled succinate was identified in an

untargeted metabolomics study, implying that this pathway might be active [26]. Still, it should be noted that it is so far impossible to discriminate between the mitochondrial, glycosomal, and cytosolic derivations of this metabolite and that only a systematic deletion of the corresponding enzymatic isoforms followed by metabolomics would illuminate the cellular compartment in which this glucose-derived succinate is produced. To sum up, the collective activity of the aforementioned reoxidation enzymes is most likely responsible for the mitochondrial NADH regeneration. Possibly, RNAi silencing of the mitochondrial malate dehydrogenase and fumarate reductase in the background of complex I and *Ndh2* null mutants would shed light on the quantitative role of each of these enzymes in mitochondrial NADH reoxidation.

The possible occurrence of the mitochondrial substrate-level phosphorylation reactions raises an interesting question regarding the mitochondrial bioenergetics of BSF and questions the origin of ATP that is needed by mitochondrial F_0F_1 ATPase in order to maintain the mitochondrial membrane potential. The classical model presumes that ATP is imported into the organelle via the activity of the ATP/ADP carrier [52,53]. However, the available data—such as low sensitivity of BSF to treatment with bongkreic acid, an inhibitor of this carrier—raise some doubts about this assertion. Interestingly, the BSF does not respire when the mitochondrial transmembrane proton gradient is dissipated upon treatment with the F_0F_1 ATPase inhibitor oligomycin or by addition of carbonyl cyanide-4-(trifluoromethoxy)phenylhydrazone (FCCP). However, when treated with bongkreic acid, which should halt the activity of F_0F_1 ATPase by restraining its substrate, the parasite consumes oxygen at the same rate as untreated cells [54,55]. On one hand, it is possible that the mitochondrial inner membrane harbors another ATP/ADP carrier; on the other hand, it is a plausible speculation that, when specific conditions emerge, the BSF mitochondrion has the capacity to employ its complex enzymatic network to produce ATP by substrate-level phosphorylation to power the F_0F_1 ATPase.

Concluding remarks

Combined, the available data reveal that the metabolic flexibility and adaptability of the BSF mitochondrion are much larger than appreciated so far. Mitochondrial metabolism appears to be controlled at various levels; a developmental program seems to be a major contributor, but recent advances in the field suggest that other cues may also play a role through fine-tuning mechanisms. However, the triggers and signaling pathways of these mechanisms remain to be identified. Furthermore, it should be realized that almost all metabolic studies have been performed with strains well adapted to laboratory conditions. While the proteomic data do not show any significant differences between the monomorphic and pleiomorphic strains, future work combining proteomics and metabolomics with functional genomics should be extended to the mitochondrion of trypanosomes isolated not only from blood but also from other tissues to determine whether their metabolism is tissue specific and, if so, what is/are the mechanism(s) that control(s) the changes. Therefore, the virtually unexplored array of pathways and enzymes begs for attention because it may have important implications for drug target identification and future novel chemotherapeutics. Moreover, a decreased morphological complexity, which is apparently not reflected in metabolic complexity, is an interesting and novel phenomenon that can now be efficiently addressed with emerging, increasingly sensitive methods.

Supporting information

S1 Table. List of mitochondrial proteins that were identified in proteomic analysis of PCF cells (columns F, G, and H) and of BSF cells (columns I, J, K, L, M, and N). The column

color coding is green for PCF, dark grey for monomorphic BSF, and light gray for pleiomorphic BSF cells.

1, identified; 0, not identified; BSF, bloodstream form; PCF, procyclic form.

(XLSX)

References

- Rico E, Rojas F, Mony BM, Szoor B, Macgregor P, et al. (2013) Bloodstream form pre-adaptation to the tsetse fly in *Trypanosoma brucei*. *Front Cell Infect Microbiol* 3: 78. <https://doi.org/10.3389/fcimb.2013.00078> PMID: 24294594
- Trindade S, Rijo-Ferreira F, Carvalho T, Pinto-Neves D, Guegan F, et al. (2016) *Trypanosoma brucei* parasites occupy and functionally adapt to the adipose tissue in mice. *Cell Host Microbe* 19: 837–848. <https://doi.org/10.1016/j.chom.2016.05.002> PMID: 27237364
- Capewell P, Cren-Travaille C, Marchesi F, Johnston P, Clucas C, et al. (2016) The skin is a significant but overlooked anatomical reservoir for vector-borne African trypanosomes. *Elife*; 5:e17717.
- Smith TK, Bringaud F, Nolan DP, Figueiredo LM (2017) Metabolic reprogramming during the *Trypanosoma brucei* life cycle. *F1000Res* 6.
- Lukeš J, Hashimi H, Verner Z, Čičová Z (2010) The remarkable mitochondrion of trypanosomes and related flagellates. In: Structures and Organelles in Pathogenic Protists. de Souza W., editor. 1 ed: Springer-Verlag Berlin Heidelberg. pp. 227–252.
- Vickerman K (1985) Developmental cycles and biology of pathogenic trypanosomes. *Br Med Bull* 41: 105–114. PMID: 3928017
- Vickerman K (1965) Polymorphism and mitochondrial activity in sleeping sickness trypanosomes. *Nature* 208: 762–766. PMID: 5868887
- Opperdoes FR, Borst P, Bakker S, Leene W (1977) Localization of glycerol-3-phosphate oxidase in the mitochondrion and particulate NAD⁺-linked glycerol-3-phosphate dehydrogenase in the microbodies of the bloodstream form to *Trypanosoma brucei*. *Eur J Biochem* 76: 29–39. PMID: 142010
- Nolan DP, Voorheis HP (1992) The mitochondrion in bloodstream forms of *Trypanosoma brucei* is energized by the electrogenic pumping of protons catalysed by the F1F0-ATPase. *Eur J Biochem* 209: 207–216. PMID: 1327770
- Schnauffer A, Clark-Walker GD, Steinberg AG, Stuart K (2005) The F1-ATP synthase complex in bloodstream stage trypanosomes has an unusual and essential function. *EMBO J* 24: 4029–4040. <https://doi.org/10.1038/sj.emboj.7600862> PMID: 16270030
- Vercesi AE, Docampo R, Moreno SN (1992) Energization-dependent Ca²⁺ accumulation in *Trypanosoma brucei* bloodstream and procyclic trypomastigotes mitochondria. *Mol Biochem Parasitol* 56: 251–257. PMID: 1484549
- Lukeš J, Basu S (2015) Fe/S protein biogenesis in trypanosomes—A review. *Biochim Biophys Acta* 1853: 1481–1492. <https://doi.org/10.1016/j.bbamcr.2014.08.015> PMID: 25196712
- Schnauffer A, Panigrahi AK, Panicucci B, Igo RP Jr., Wirtz E, et al. (2001) An RNA ligase essential for RNA editing and survival of the bloodstream form of *Trypanosoma brucei*. *Science* 291: 2159–2162. <https://doi.org/10.1126/science.1058955> PMID: 11251122
- Read LK, Lukeš J, Hashimi H (2016) Trypanosome RNA editing: the complexity of getting U in and taking U out. *Wiley Interdiscip Rev RNA* 7: 33–51. <https://doi.org/10.1002/wrna.1313> PMID: 26522170
- Huang G, Vercesi AE, Docampo R (2013) Essential regulation of cell bioenergetics in *Trypanosoma brucei* by the mitochondrial calcium uniporter. *Nat Commun* 4: 2865. <https://doi.org/10.1038/ncomms3865> PMID: 24305511
- Docampo R, Lukeš J (2012) Trypanosomes and the solution to a 50-year mitochondrial calcium mystery. *Trends Parasitol* 28: 31–37. <https://doi.org/10.1016/j.pt.2011.10.007> PMID: 22088944
- Gualdrón-Lopez M, Brennand A, Hannaert V, Quinones W, Caceres AJ, et al. (2012) When, how and why glycolysis became compartmentalised in the Kinetoplastea. A new look at an ancient organelle. *Int J Parasitol* 42: 1–20. <https://doi.org/10.1016/j.ijpara.2011.10.007> PMID: 22142562
- Gabaldon T, Ginger ML, Michels PA (2016) Peroxisomes in parasitic protists. *Mol Biochem Parasitol* 209: 35–45. <https://doi.org/10.1016/j.molbiopara.2016.02.005> PMID: 26896770
- Clarkson AB Jr., Grady RW, Grossman SA, McCallum RJ, Brohn FH (1981) *Trypanosoma brucei brucei*: a systematic screening for alternatives to the salicylhydroxamic acid-glycerol combination. *Mol Biochem Parasitol* 3: 271–291. PMID: 6795501

20. Grant PT, Fulton JD (1957) The catabolism of glucose by strains of *Trypanosoma rhodesiense*. *Biochem J* 66: 242–250. PMID: [13445679](#)
21. van Grinsven KW, Van Den Abbeele J, Van den Bossche P, van Hellemond JJ, Tielens AG (2009) Adaptations in the glucose metabolism of procyclic *Trypanosoma brucei* isolates from tsetse flies and during differentiation of bloodstream forms. *Eukaryot Cell* 8: 1307–1311. <https://doi.org/10.1128/EC.00091-09> PMID: [19542311](#)
22. Nolan DP, Rolin S, Rodriguez JR, Van Den Abbeele J, Pays E (2000) Slender and stumpy bloodstream forms of *Trypanosoma brucei* display a differential response to extracellular acidic and proteolytic stress. *Eur J Biochem* 267: 18–27. PMID: [10601846](#)
23. Mony BM, Matthews KR (2015) Assembling the components of the quorum sensing pathway in African trypanosomes. *Mol Microbiol* 96: 220–232. <https://doi.org/10.1111/mmi.12949> PMID: [25630552](#)
24. Turner CM (1990) The use of experimental artefacts in African trypanosome research. *Parasitol Today* 6: 14–17. PMID: [15463248](#)
25. Mazet M, Morand P, Biran M, Bouyssou G, Courtois P, et al. (2013) Revisiting the central metabolism of the bloodstream forms of *Trypanosoma brucei*: production of acetate in the mitochondrion is essential for parasite viability. *PLoS Negl Trop Dis* 7: e2587. <https://doi.org/10.1371/journal.pntd.0002587> PMID: [24367711](#)
26. Creek DJ, Mazet M, Achcar F, Anderson J, Kim DH, et al. (2015) Probing the metabolic network in bloodstream-form *Trypanosoma brucei* using untargeted metabolomics with stable isotope labelled glucose. *PLoS Pathog* 11: e1004689. <https://doi.org/10.1371/journal.ppat.1004689> PMID: [25775470](#)
27. Spitznagel D, Ebikeme C, Biran M, Nic a' Bhaird N, Bringaud F, et al. (2009) Alanine aminotransferase of *Trypanosoma brucei*—a key role in proline metabolism in procyclic life forms. *FEBS J* 276: 7187–7199. <https://doi.org/10.1111/j.1742-4658.2009.07432.x> PMID: [19895576](#)
28. Creek DJ, Nijagal B, Kim DH, Rojas F, Matthews KR, et al. (2013) Metabolomics guides rational development of a simplified cell culture medium for drug screening against *Trypanosoma brucei*. *Antimicrob Agents Chemother* 57: 2768–2779. <https://doi.org/10.1128/AAC.00044-13> PMID: [23571546](#)
29. Panigrahi AK, Ogata Y, Zíková A, Anupama A, Dalley RA, et al. (2009) A comprehensive analysis of *Trypanosoma brucei* mitochondrial proteome. *Proteomics* 9: 434–450. <https://doi.org/10.1002/pmic.200800477> PMID: [19105172](#)
30. Acestor N, Panigrahi AK, Ogata Y, Anupama A, Stuart KD (2009) Protein composition of *Trypanosoma brucei* mitochondrial membranes. *Proteomics* 9: 5497–5508. <https://doi.org/10.1002/pmic.200900354> PMID: [19834910](#)
31. Zíková A, Panigrahi AK, Dalley RA, Acestor N, Anupama A, et al. (2008) *Trypanosoma brucei* mitochondrial ribosomes: affinity purification and component identification by mass spectrometry. *Mol Cell Proteomics* 7: 1286–1296. <https://doi.org/10.1074/mcp.M700490-MCP200> PMID: [18364347](#)
32. Panigrahi AK, Zíková A, Dalley RA, Acestor N, Ogata Y, et al. (2008) Mitochondrial complexes in *Trypanosoma brucei*: a novel complex and a unique oxidoreductase complex. *Mol Cell Proteomics* 7: 534–545. <https://doi.org/10.1074/mcp.M700430-MCP200> PMID: [18073385](#)
33. Acestor N, Zíková A, Dalley RA, Anupama A, Panigrahi AK, et al. (2011) *Trypanosoma brucei* mitochondrial respiratome: composition and organization in procyclic form. *Mol Cell Proteomics* 10: M110006908.
34. Niemann M, Wiese S, Mani J, Chanfon A, Jackson C, et al. (2013) Mitochondrial outer membrane proteome of *Trypanosoma brucei* reveals novel factors required to maintain mitochondrial morphology. *Mol Cell Proteomics* 12: 515–528. <https://doi.org/10.1074/mcp.M112.023093> PMID: [23221899](#)
35. Peikert CD, Mani J, Morgenstern M, Kaser S, Knapp B, et al. (2017) Charting organellar importomes by quantitative mass spectrometry. *Nat Commun* 8: 15272. <https://doi.org/10.1038/ncomms15272> PMID: [28485388](#)
36. Zíková A, Panigrahi AK, Uboldi AD, Dalley RA, Handman E, et al. (2008) Structural and functional association of *Trypanosoma brucei* MIX protein with cytochrome c oxidase complex. *Eukaryot Cell* 7: 1994–2003. <https://doi.org/10.1128/EC.00204-08> PMID: [18776036](#)
37. Zíková A, Schnauffer A, Dalley RA, Panigrahi AK, Stuart KD (2009) The F(0)F(1)-ATP synthase complex contains novel subunits and is essential for procyclic *Trypanosoma brucei*. *PLoS Pathog* 5: e1000436. <https://doi.org/10.1371/journal.ppat.1000436> PMID: [19436713](#)
38. Dejung M, Subota I, Bucerius F, Dindar G, Freiwald A, et al. (2016) Quantitative proteomics uncovers novel factors involved in developmental differentiation of *Trypanosoma brucei*. *PLoS Pathog* 12: e1005439. <https://doi.org/10.1371/journal.ppat.1005439> PMID: [26910529](#)
39. Butter F, Bucerius F, Michel M, Cicova Z, Mann M, et al. (2013) Comparative proteomics of two life cycle stages of stable isotope-labeled *Trypanosoma brucei* reveals novel components of the parasite's

- host adaptation machinery. *Mol Cell Proteomics* 12: 172–179. <https://doi.org/10.1074/mcp.M112.019224> PMID: 23090971
40. Urbaniak MD, Martin DM, Ferguson MA (2013) Global quantitative SILAC phosphoproteomics reveals differential phosphorylation is widespread between the procyclic and bloodstream form lifecycle stages of *Trypanosoma brucei*. *J Proteome Res* 12: 2233–2244. <https://doi.org/10.1021/pr400086y> PMID: 23485197
 41. Urbaniak MD, Guther ML, Ferguson MA (2012) Comparative SILAC proteomic analysis of *Trypanosoma brucei* bloodstream and procyclic lifecycle stages. *PLoS ONE* 7: e36619. <https://doi.org/10.1371/journal.pone.0036619> PMID: 22574199
 42. Urbaniak MD, Mathieson T, Bantscheff M, Eberhard D, Grimaldi R, et al. (2012) Chemical proteomic analysis reveals the drugability of the kinome of *Trypanosoma brucei*. *ACS Chem Biol* 7: 1858–1865. <https://doi.org/10.1021/cb300326z> PMID: 22908928
 43. Gunasekera K, Wuthrich D, Braga-Lagache S, Heller M, Ochsenreiter T (2012) Proteome remodelling during development from blood to insect-form *Trypanosoma brucei* quantified by SILAC and mass spectrometry. *BMC Genomics* 13: 556. <https://doi.org/10.1186/1471-2164-13-556> PMID: 23067041
 44. Verner Z, Basu S, Benz C, Dixit S, Dobáková E, et al. (2015) Malleable mitochondrion of *Trypanosoma brucei*. *Int Rev Cell Mol Biol* 315: 73–151. <https://doi.org/10.1016/bs.ircmb.2014.11.001> PMID: 25708462
 45. Zhang X, Cui J, Nilsson D, Gunasekera K, Chanfon A, et al. (2010) The *Trypanosoma brucei* MitoCarta and its regulation and splicing pattern during development. *Nucleic Acids Res* 38: 7378–7387. <https://doi.org/10.1093/nar/gkq618> PMID: 20660476
 46. Sykes SE, Hajduk SL (2013) Dual functions of alpha-ketoglutarate dehydrogenase E2 in the Krebs cycle and mitochondrial DNA inheritance in *Trypanosoma brucei*. *Eukaryot Cell* 12: 78–90. <https://doi.org/10.1128/EC.00269-12> PMID: 23125353
 47. Overath P, Czichos J, Haas C (1986) The effect of citrate/cis-aconitate on oxidative metabolism during transformation of *Trypanosoma brucei*. *Eur J Biochem* 160: 175–182. PMID: 3769918
 48. Sykes S, Szempruch A, Hajduk S (2015) The krebs cycle enzyme alpha-ketoglutarate decarboxylase is an essential glycosomal protein in bloodstream African trypanosomes. *Eukaryot Cell* 14: 206–215. <https://doi.org/10.1128/EC.00214-14> PMID: 25416237
 49. Štafková J, Mach J, Biran M, Verner Z, Bringaud F, et al. (2016) Mitochondrial pyruvate carrier in *Trypanosoma brucei*. *Mol Microbiol* 100: 442–456. <https://doi.org/10.1111/mmi.13325> PMID: 26748989
 50. Surve S, Heestand M, Panicucci B, Schnauffer A, Parsons M (2012) Enigmatic presence of mitochondrial complex I in *Trypanosoma brucei* bloodstream forms. *Eukaryot Cell* 11: 183–193. <https://doi.org/10.1128/EC.05282-11> PMID: 22158713
 51. Surve SV, Jensen BC, Heestand M, Mazet M, Smith TK, et al. (2016) NADH dehydrogenase of *Trypanosoma brucei* is important for efficient acetate production in bloodstream forms. *Mol Biochem Parasitol* 211: 57–61. <https://doi.org/10.1016/j.molbiopara.2016.10.001> PMID: 27717801
 52. Pena-Díaz P, Pelosi L, Ebikeme C, Colasante C, Gao F, et al. (2012) Functional characterization of TbMCP5, a conserved and essential ADP/ATP carrier present in the mitochondrion of the human pathogen *Trypanosoma brucei*. *J Biol Chem* 287: 41861–41874. <https://doi.org/10.1074/jbc.M112.404699> PMID: 23074217
 53. Gnipová A, Šubrtová K, Panicucci B, Horváth A, Lukeš J, et al. (2015) The ADP/ATP carrier and its relationship to OXPHOS in an ancestral protist, *Trypanosoma brucei*. *Eukaryot Cell* 14: 297–310. <https://doi.org/10.1128/EC.00238-14> PMID: 25616281
 54. Miller PG, Klein RA (1980) Effects of oligomycin on glucose utilization and calcium transport in African trypanosomes. *J Gen Microbiol* 116: 391–396. <https://doi.org/10.1099/00221287-116-2-391> PMID: 6246194
 55. Bienen EJ, Maturi RK, Pollakis G, Clarkson AB Jr. (1993) Non-cytochrome mediated mitochondrial ATP production in bloodstream form *Trypanosoma brucei brucei*. *Eur J Biochem* 216: 75–80. PMID: 8365419

Review

Not in your usual Top 10: protists that infect plants and algae

ARNE SCHWELM^{1,2,*}, JULIA BADSTÖBER², SIMON BULMAN³, NICOLAS DESOIGNIES⁴, MOHAMMAD ETEMADI², RICHARD E. FALLOON³, CLAIRE M. M. GACHON⁵, ANNE LEGREVE⁶, JULIUS LUKEŠ^{7,8,9}, UELI MERZ¹⁰, ANNA NENAROKOVA^{7,8}, MARTINA STRITTMATTER^{5†}, BROOKE K. SULLIVAN^{11,12} AND SIGRID NEUHAUSER²

¹Department of Plant Biology, Uppsala BioCentre, Linnean Centre for Plant Biology, Swedish University of Agricultural Sciences, Uppsala SE-75007, Sweden

²Institute of Microbiology, University of Innsbruck, Innsbruck 6020, Austria

³New Zealand Institute for Plant and Food Research Ltd, Lincoln 7608, New Zealand

⁴Applied Plant Ecophysiology, Haute Ecole Provinciale de Hainaut-Condorcet, Ath 7800, Belgium

⁵The Scottish Association for Marine Science, Scottish Marine Institute, Oban PA37 1QA, UK

⁶Université catholique de Louvain, Earth and Life Institute, Louvain-la-Neuve 1348, Belgium

⁷Institute of Parasitology, Biology Centre, 37005 České Budějovice (Budweis), Czech Republic

⁸Faculty of Sciences, University of South Bohemia, 37005 České Budějovice (Budweis), Czech Republic

⁹Integrated Microbial Biodiversity, Canadian Institute for Advanced Research, Toronto, Ontario M5G 1Z8, Canada

¹⁰Plant Pathology, Institute of Integrative Biology, ETH Zurich, Zurich 8092, Switzerland

¹¹School of Biosciences, University of Melbourne, Parkville, Vic. 3010, Australia

¹²School of Biosciences, Victorian Marine Science Consortium, Queenscliff, Vic. 3225, Australia

SUMMARY

Fungi, nematodes and oomycetes belong to the most prominent eukaryotic plant pathogenic organisms. Unicellular organisms from other eukaryotic lineages, commonly addressed as protists, also infect plants. This review provides an introduction to plant pathogenic protists, including algae infecting oomycetes, and their current state of research.

Keywords: algae, protist, plant pathogens, plasmodiophorids, stramenopiles, phytomonas, phytomyxae.

INTRODUCTION

Molecular Plant Pathology has published a series of the Top 10 most important plant-pathogenic viruses (Scholthof *et al.*, 2011), fungi (Dean *et al.*, 2012), bacteria (Mansfield *et al.*, 2012), nematodes (Jones *et al.*, 2013) and oomycetes (Kamoun *et al.*, 2015). The reviews of these major groups of plant pathogens do not cover a selection of protists that infect plants and algae leading to economically important diseases. These ‘non-standard’ plant pathogens are dispersed across the eukaryotic phylogenetic tree (Fig. 1), often in taxa unfamiliar to many plant pathologists as they are usually not associated with plant infections. In this

review, we would like to introduce and raise awareness of such phylogenetically diverse eukaryotic plant pathogens.

We describe diseases caused by these organisms, and the current state of research, especially with respect to their molecular biology and host interactions. We start with *Phytomonas*, plant pathogens in the trypanosomatids in the Excavata supergroup, a group better known as human and animal pathogens. They are followed by Phytomyxea, which are part of the Rhizaria supergroup and include agriculturally important plant pathogens, vectors of phytoviruses and species that infect marine plants and algae (Bulman and Braselton, 2014). Next, *Labyrinthula* are described, plant pathogens in the Stramenopiles, which are phylogenetic basal to oomycetes. Our review also includes marine oomycete parasites of red and brown algae, which impact on the fast growing aquaculture sector (Gachon *et al.*, 2010). Advancing research in this field will benefit aquacultural sustainability and our understanding of higher oomycetes because of their basal phylogenetic position inside the oomycetes (Beakes *et al.*, 2012).

Whole-genome or in-depth transcriptomic data for the species presented here are rare, with the exception of the Phytomyxea and *Phytomonas*. The organisms outlined reflect existing molecular knowledge; nevertheless, we emphasize that there are further important ‘unusual’ pathogens, especially on cultivated algae.

EXCAVATA – KINETOPLASTEA TRYPANOSOMATIDAE – *PHYTOMONAS*

Trypanosomatids are a species-rich monophyletic group of obligate parasitic flagellates that are usually transmitted by insects. They are best known as agents of human and livestock diseases,

*Correspondence: Email: arne.schwelm@uibk.ac.at

†Present address: Station Biologique de Roscoff, CNRS – UPMC, UMR7144 Adaptation and Diversity in the Marine Environment, Place Georges Teissier, CS 90074, 29688 Roscoff Cedex, France.

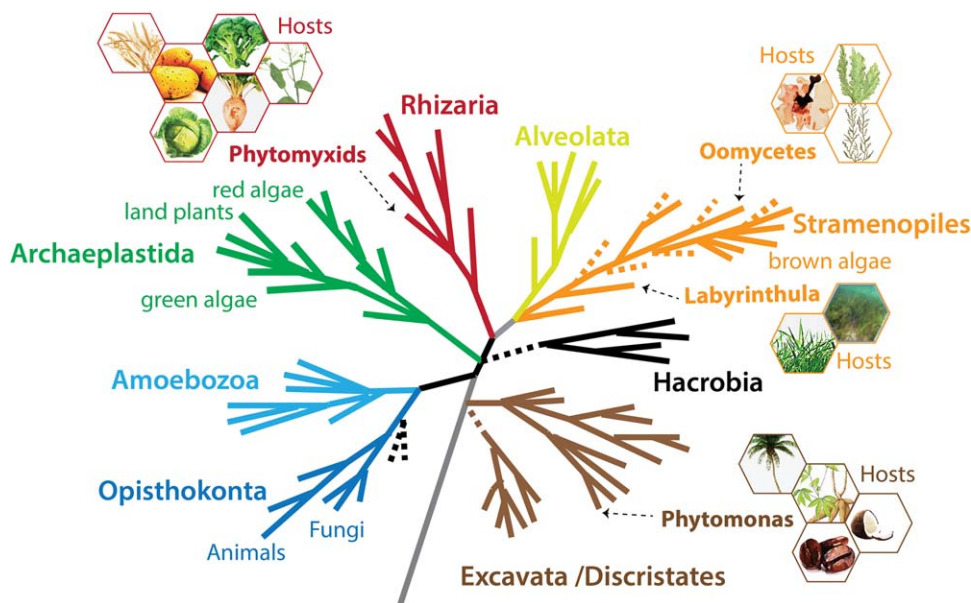


Fig. 1 A schematic current eukaryotic tree of life indicating the phylogenetic positions of the eukaryotic plant pathogens outlined in this review. The hexagons show examples of the host species for each pathogen group. The phylogenetic tree was created by S. Baldauf (Uppsala University, Uppsala, Sweden) and reproduced with permission.

such as sleeping sickness, Chagas disease and leishmaniosis, caused by *Trypanosoma brucei*, *T. cruzi* and *Leishmania* spp., respectively (Lukeš *et al.*, 2014). Trypanosomatids also include the monophyletic genus *Phytomonas* (Fig. 2), which contains all known plant-dwelling trypanosomatids, some of which are pathogenic (Seward *et al.*, 2016). The ancestral monoxenous lifestyle (development restricted to one host species) of trypanosomatids evolved at least three times independently into a dixenous strategy (Maslov *et al.*, 2013) in *Trypanosoma*, *Leishmania* and *Phytomonas* (Lukeš *et al.*, 2014). *Phytomonas* spp. are adapted to sap-sucking insects as primary hosts and plants as secondary hosts (Jaskowska *et al.*, 2015). *Phytomonas* spp. were first described from the latex of Mediterranean spurge (*Euphorbia pilulifera*) (Lafont, 1909). Currently, the genus *Phytomonas* includes more than 200 species that colonize over 20 plant families (Camargo, 1999, Jaskowska *et al.*, 2015).

Phytomonas spp. can be separated into four ecological subgroups based on whether they inhabit the latex ducts, fruits, phloem or flowers of their host plants (Camargo, 1999). Most commonly, *Phytomonas* spp. reside in latex ducts, yet the most pathogenic species are phloem dwelling, such as *P. leptovisorum* and *P. staheli*, which cause coffee phloem necrosis (CPN) and palm wilts, respectively (Jaskowska *et al.*, 2015). *Phytomonas leptovisorum* infection triggers multiple divisions of the sieve tubes in coffee roots, leading to CPN. The disease is a potential threat to Brazil as the world's largest coffee exporter, from which CPN has been reported, but never spread (Camargo, 1999). This disease occurs either acutely (plants wither and die within 2 months) or chronically (plants gradually die within a year) (Stahel, 1931).

Phytomonas staheli causes wilts of coconut (*Cocos nucifera*) and oil palms (*Elaeis guineensis*) (McGhee and McGhee, 1979).

Both deadly wilts, 'hartrot' of coconut palms and 'marchitez sorpresiva' of oil palms, are characterized by progressive leaf browning, followed by rapid rotting of fruits, spears and roots (Kastelein, 1987; Lopez, 1975). Slow wilt of oil palms ('marchitez lenta') manifests as additional chlorosis (Di Lucca *et al.*, 2013). Symptomless plants and wild hosts can harbour *Phytomonas* flagellates (Di Lucca *et al.*, 2013). Potential disease outbreaks constantly threaten palm cultivation in South and Central America. In one Surinamese district, *Phytomonas* destroyed half of the coconut population (Kastelein, 1987). The latex-inhabiting *P. françai* is linked to empty roots disease ('chochamento de raízes') of the Unha cassava (*Manihot esculenta*) variety, although its pathogenicity remains unclear (Jaskowska *et al.*, 2015; Kitajima *et al.*, 1986).

The first *Phytomonas* draft genome came from the tomato fruit-inhabiting *P. serpens* (Kořený *et al.*, 2012), which produces no significant systemic disease, but causes yellow spots on fruit (Camargo, 1999). The genomes of the pathogenic phloem-specific *Phytomonas* strain HART1 from Guyanan coconut and the non-symptomatic latex-specific strain EM1 from *Euphorbia* were generated shortly after (Porcel *et al.*, 2014). Recently, the genome of the cassava latex-inhabiting *P. françai* has been announced (Butler *et al.*, 2017), which will enable comparative genomics of *Phytomonas* spp. with different host and ecological lifestyles in the future.

The *Phytomonas* genomes are compact, consisting of single-copy genes, and are almost free of transposable elements and repeats. Therefore they are smaller (≈ 18 Mb) than most trypanosomatid genomes (26–33 Mb). *Phytomonas* spp. contain only about 6400 protein-coding genes versus approximately 10 400 found typically in trypanosomatids.

As in other biotrophs, *Phytomonas* metabolism is highly adapted to parasitic lifestyles. These plant pathogens contain

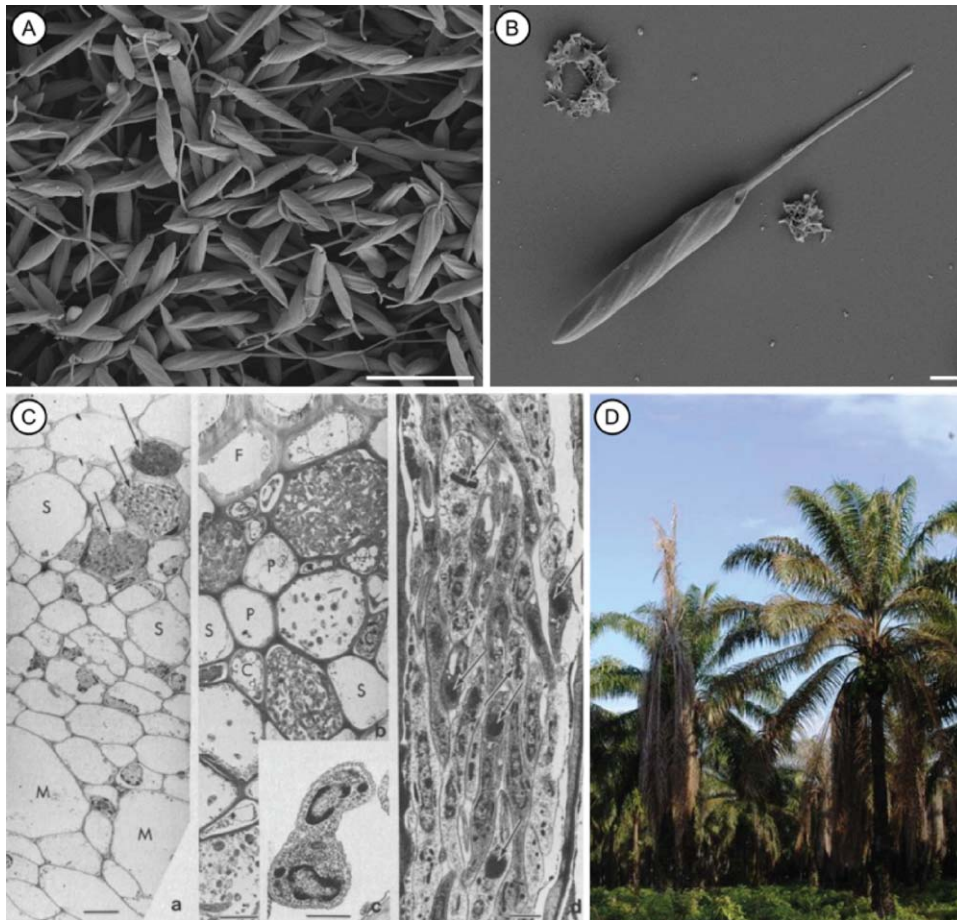


Fig. 2 *Phytomonas* sp. and palm infections. (A, B) Scanning electron micrographs of *Phytomonas serpens* cells in culture (scale bars, 10 and 1 μm). (Courtesy of Martina Tesařová.) (C) Transmission electron micrographs of *Phytomonas* sp. flagellates in the phloem of coconut palms affected by hartrot. C, companion cell; F, fibre; M, immature metaxylem; P, phloem parenchyma cell; S, sieve elements free of flagellates. (a) Transverse section of a differentiating vascular bundle, showing recently matured sieve elements filled with flagellates (scale bar, 10 μm). (b) Transverse section of the phloem in palm with advanced symptoms (scale bar, 5 μm). (c) Transverse section of a dividing flagellate (scale bar, 0.5 μm). (d) Longitudinal section of a sieve element filled with flagellates. Arrows indicate the kinetoplast DNA (scale bar, 1 μm). (Reproduced from Parthasarathy *et al.*, 1976.) (D) Coconut palms with symptoms of hartrot. (Photograph: Monica L. Elliott, Professor, Plant Pathology, University of Florida, Institute of Food and Agricultural Sciences (UF/IFAS), Gainesville, FL, USA.)

fewer genes involved in amino acid synthesis and energy metabolism and fewer protein kinases than the related *Leishmania* and *Trypanosoma* spp. Fatty acids (FAs) are synthesized via elongases instead of *de novo*, as FA synthases are missing (Porcel *et al.*, 2014). *Phytomonas* spp. have the unique capacity amongst trypanosomatids to live in the total absence of haem, although they might be able to scavenge it (Košný *et al.*, 2012). In addition, they have lost several cytochrome subunits of respiratory complexes. For energy production, *Phytomonas* may depend solely on glycolysis, whereas other trypanosomatids (at least in part of their life cycle) rely on mitochondrial amino acid metabolism as their main energy source (Jaskowska *et al.*, 2015; Porcel *et al.*, 2014). As their insect vector(s) feed on carbohydrate-rich plant juices, *Phytomonas* might not require a switch from carbohydrate to amino acid metabolism. *Phytomonas* spp. contain complete sets

of glycolytic enzymes and large numbers of glycosomes, into which glycolysis is compartmentalized (Hannaert *et al.*, 2003; Porcel *et al.*, 2014). Also unique amongst trypanosomatids, *Phytomonas* spp. possess the capacity to feed on plant polysaccharides using glucoamylase and α -glucosidase enzymes. In addition, an α, α -trehalose phosphorylase, acquired by horizontal gene transfer, enables feeding on trehalose, a common sugar in the plant and insect hosts of *Phytomonas* (Porcel *et al.*, 2014).

The *Phytomonas* HART1 and EM1 isolates share a majority of genes. However, only the phloem-restricted pathogenic HART1 encodes invertase genes for the degradation of sucrose (Porcel *et al.*, 2014), probably as an adaptation to the abundance of sucrose in the phloem. For both the HART1 and EM1 isolates, 282 secreted proteins were predicted. Their secretomes contain no plant cell wall-degrading enzymes, which reflects the feeding of

the pathogens on extracellular plant fluids. It is unknown whether *Phytomonas* spp. secrete protein effectors, which modulate host plant immune responses. However, several aspartyl proteases that are absent from the genomes of *Leishmania* and *Trypanosoma* are secreted in both *Phytomonas* strains (Porcel *et al.*, 2014). These proteases may be involved in *Phytomonas*–host interactions, as seen for oomycete and fungal plant pathogens (Jashni *et al.*, 2015). The pathogenic HART1 strain carries five copies of a cathepsin-like aspartyl protease, derived from duplication events, whereas EM1 has only a single copy. This implies that these enzymes are potential virulence factors (Porcel *et al.*, 2014). The gene family of major surface proteases, which are involved in the pathogenicity of *Leishmania*, underwent an expansion in the genus *Phytomonas* (Jackson, 2015). The surface glycoprotein 63 subfamily is present in 20 copies in HART1 and only twice in EM1, a putative adaptation of HART1 to the phloem environment (Jaskowska *et al.*, 2015; Porcel *et al.*, 2014).

Although the procyclic stage of *Phytomonas* spp. can be easily cultivated, an experimental system including their plant host is not available. Hence, our understanding of how these plant-dwelling or plant-parasitizing flagellates interact with their plant hosts is only at an early stage.

Currently, there is no treatment or prevention of the diseases caused by *Phytomonas*, except for the simple extermination of infected plants (Jaskowska *et al.*, 2015). However, it has been observed that the tomato (*Solanum lycopersicum*) is relatively resistant to *P. serpens*, as the parasite only causes yellow spots on its fruits, resulting in their lower commercial value. Interestingly, the tomato defensive alkaloids tomatine and tomatidine, surface-active saponin-like compounds, induce permeabilization and vacuolization of the parasite (Medina *et al.*, 2015). Both alkaloids inhibit the growth of *P. serpens* and therefore represent potential therapeutic agents against these phytopathogens (Medina *et al.*, 2015).

RHIZARIA

Phytomyxea – plasmodiophorids

The obligate biotrophic Plasmodiophorida (plasmodiophorids) belong to the Phytomyxea (phytomycids) in the eukaryotic supergroup Rhizaria (Fig. 1) (Adl *et al.*, 2012; Burki and Keeling, 2014; Burki *et al.*, 2010). These organisms infect a wide variety of hosts, including oomycetes and brown algae (Neuhauser *et al.*, 2014). Plasmodiophorids cause substantial damage to crops, including brassicas (*Plasmodiophora brassicae*), potatoes (*Spongospora subterranea*) and as vectors of viruses to beets, peanut and monocots (e.g. maize, rice, sugarcane, wheat, sorghum) (*Polymyxa* spp.) and potatoes (*S. subterranea*).

The plasmodiophorid life cycle consists of two phases: a sporangial stage leading to short-lived zoospores, and a sporogonic stage leading to the formation of persistent resting spores (Figs

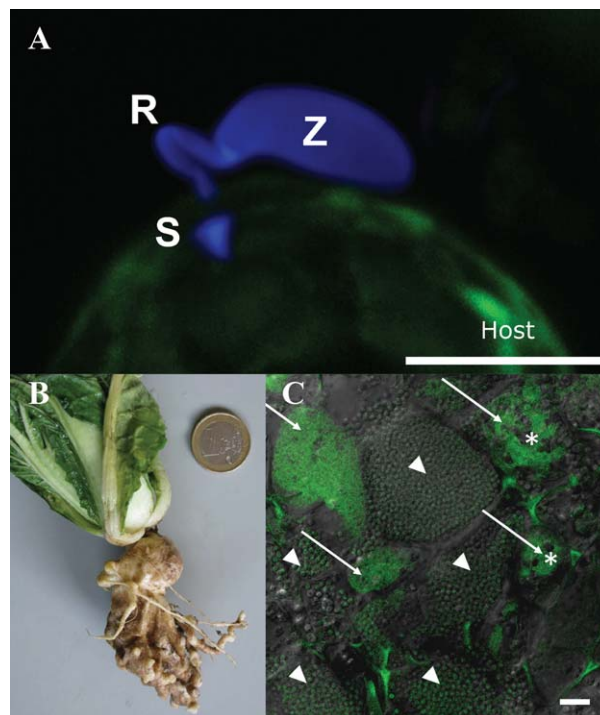
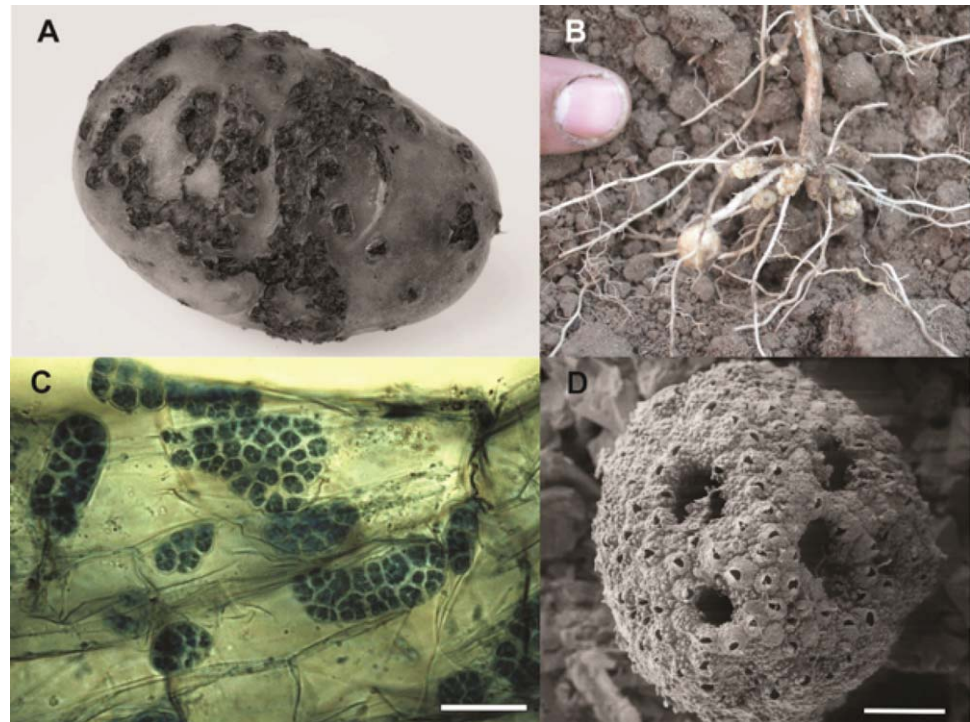


Fig. 3 Phytomyxid infection and clubroot. (A) Phytomyxean parasites infect their host via a specialized extrusosome, called a 'Rohr (R) and Stachel (S)'. The image shows a zoospore (Z) of the phagomyxid *Maullinia ectocarpii* infecting a female gametophyte of *Macrocytis pyrifera* (host). The *M. ectocarpii* spore was stained with calcofluor white and the host is visible via autofluorescence. Bar, 5 μ m. (B) Clubroot symptoms on Chinese cabbage. (C) Laser scanning micrograph of *Plasmodiophora brassicae* resting spores (arrowheads) and plasmodia (arrows) in clubroot tissue. Plasmodia of different ages can be distinguished by the presence of typical vacuoles (asterisks), which disappear when the plasmodia start to differentiate into resting spores. Overlay of a light microscopic image and the signal of a *Plasmodiophora*-specific fluorescence *in situ* hybridization (FISH) probe (green: excitation, 488 nm; emission, 510–550 nm). Bar, 20 μ m.

3–5). Resting spores give rise to biflagellate primary zoospores which inject their cellular contents into host cells via a 'Rohr und Stachel' (Aist and Williams, 1971) (Fig. 3), initiating the sporangial life cycle stage. Multinucleate plasmodia develop and produce (mitotic) secondary zoospores, which can infect host cells and develop sporogonic multinucleate plasmodia that mature into resting spores. In the sporogonic stage, gall-causing plasmodiophorids induce division and massive enlargement of host cells (for greater detail, see Bulman and Braselton, 2014).

The durability of resting spores and inconsistent chemical control make the management of plasmodiophorid diseases difficult, and biological control efforts are only beginning (Ludwig-Müller, 2016; O'Brien and Milroy, 2017). Current management mostly relies on the use of resistant host varieties and crop rotation (Bittara *et al.*, 2016; Ludwig-Müller, 2016). Pathogen detection and quantification in soil and *in planta* are important. Sequences of

Fig. 4 Potato infection by *Spongospora subterranea*. The potato pathogen *Spongospora subterranea* infects host tubers, roots and stolons, resulting in the development of powdery scab lesions (A) and galls (B). These usually appear in potato crops 2–3 months after planting, and mature to release sporosori (conglomerations of resting spores). A sporosorus contains 500–1000 resting spores, each containing a primary zoospore (D; bar, 10 μm). Secondary zoospores formed in zoosporangia (C; bar, 20 μm) emerge through root cell walls, disrupting host nutrient and water uptake.



the ribosomal operon [i.e. 18S, 28S and internal transcribed spacer (ITS) ribosomal DNA (rDNA)] are widely used for these purposes (Bulman and Marshall, 1998; Faggian and Strelkov, 2009; van de Graaf *et al.*, 2007; Vaianopoulos *et al.*, 2007; Ward *et al.*, 2004, 2005). Comparison of ITS and rDNA sequences has revealed various degrees of interspecific and intraspecific variation in plasmodiophorid species (Gau *et al.*, 2013; van de Graaf *et al.*, 2007; Qu and Christ, 2004; Schwelm *et al.*, 2016).

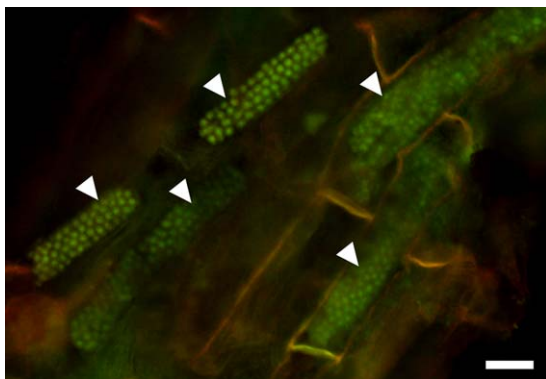


Fig. 5 Resting spores of *Polymyxa graminis* in *Poa* sp. Resting spores are arranged in typical, long and cylindrical cytosori (arrowheads). The sample was stained with acridine orange, showing the nuclei of the fully developed resting spores. Epifluorescence micrograph obtained using blue excitation with long-pass emission (Nikon B-2A filter) allowing for the detection of DNA. Bar, 20 μm .

Plasmodiophora brassicae

Plasmodiophora brassicae causes clubroot, a disease that leads to significant losses of *Brassica* oilseed and vegetable crop production worldwide (Dixon, 2009). Rapeseed cultivation for the production of biofuels, vegetable oils, industrial lubricants and rapeseed meal is of great economic importance, with a worldwide production of 27 million tonnes in 2012 (Carré and Pouzet, 2014). Clubroot has long been a major constraint for *Brassica* cultivation. A severe outbreak in 1872 in Russia led to the discovery of *Pl. brassicae* (Woronin, 1877). Clubroot causes crop losses of approximately 10% worldwide, but local losses are often greater (Dixon, 2009). Best practices for control are long crop rotation periods (although resting spores remain infective for decades), liming or cultivation of tolerant *Brassica* crops (Diederichsen *et al.*, 2009; Ludwig-Müller, 2016). Clubroot resistance genes have been identified in *Brassica* genomes (Hatakeyama *et al.*, 2013). However, resistance mechanisms are unclear and breakdown of 'resistance' has been repeatedly observed (Diederichsen *et al.*, 2009; Strelkov *et al.*, 2016; Zamani-Noor, 2017). Breeding for clubroot resistance is complicated as several pathotypes of *Pl. brassicae* exist. Genetic differences exist between *Pl. brassicae* strains, even within individual root galls, and chromosome polymorphism between strains has been suggested (Fähling *et al.*, 2003; Graf *et al.*, 2004; Klewer *et al.*, 2001). However, molecular markers for *Pl. brassicae* pathotypes have yet to be established.

The genome of a European *Pl. brassicae* single-spore isolate has been generated recently (Schwelm *et al.*, 2015), followed

shortly after by genomic data for isolates from Canada and China (Bi *et al.*, 2016; Rolfe *et al.*, 2016). The *Pl. brassicae* genome is small (24.2–25.5 Mb), as a result of a high gene density and few repetitive elements (2%–5%) (Rolfe *et al.*, 2016; Schwelm *et al.*, 2015). The first single-nucleotide polymorphism (SNP) cluster analyses of the available *Pl. brassicae* genomes indicated relationships between SNPs, host ranges and regional origins (Rolfe *et al.*, 2016). Additional genome sequencing of *Pl. brassicae* isolates should shed light on *Pl. brassicae* genomic diversity and pathotype-specific features.

The *Pl. brassicae* genomes show similar features to those of other biotrophic plant pathogens. Host dependence is evident, i.e. from a reduced number of biosynthesis genes for thiamine and certain amino acids (Rolfe *et al.*, 2016; Schwelm *et al.*, 2015). Transporter proteins may aid nutrient acquisition from the hosts (Rolfe *et al.*, 2016). The *Pl. brassicae* genome encodes few carbohydrate-active enzymes (CAZymes). Genes encoding for plant cell wall-degrading enzymes are also rare, possibly a consequence of the mechanical penetration strategy via a 'Rohr und Stachel'. However, chitin-related enzymes are enriched (Rolfe *et al.*, 2016; Schwelm *et al.*, 2015), which are probably involved in building the chitinous resting spore cell walls (Moxham and Buczacki, 1983).

In root galls, different life cycle stages of *Pl. brassicae* occur simultaneously (Fig. 3), making time course experiments difficult. The transcriptomics of isolated plasmodia show a highly active metabolism, i.e. the high expression of glyoxylate cycle-related genes suggests a high turnover from carbohydrates and lipids in the plasmodia (Schwelm *et al.* 2015). Lipids start to accumulate in the plasmodial stage and are stored in organelles in the plasmodia and resting spores (Bi *et al.*, 2016; Moxham and Buczacki, 1983). The lipids are potential energy sources for resting spores and, as *Pl. brassicae*, like *Phytophthora*, does not contain an FA synthase (Schwelm *et al.*, 2015), it might synthesize the lipids from host-derived precursors.

Depending on the strain sequenced, 553–590 secreted *Pl. brassicae* proteins were predicted. Effector candidates including the amino acid motif RxLR, known from *Phytophthora* effectors (Kamoun *et al.*, 2015), are rare in *Pl. brassicae* (Rolfe *et al.*, 2016; Schwelm *et al.*, 2015). Crinkler (CRN)-related proteins were found in *Pl. brassicae* (Zhang *et al.*, 2016a), but their functions are unknown. No effector candidates containing the chitin-binding LysM-motif, known to interfere with chitin detection in fungal-plant interactions (Kombink and Thomma, 2013), were detected in *Pl. brassicae*.

Plasmodiophora brassicae infection results in a heavily altered host metabolism (Jubault *et al.*, 2013): transcriptional and proteomic changes occur in pathways involved in lipid, flavonoid and plant hormone metabolism, defence responses, and carbohydrate and cell wall synthesis of the *Brassica* hosts (Agarwal *et al.*, 2011; Chen *et al.*, 2015; Ludwig-Müller *et al.*, 2009; Päsold *et al.*, 2010; Siemens *et al.*, 2009; Zhang *et al.*, 2016b). In *Arabidopsis*, gall

formation results from increased host vascular cambium activity combined with significant reduction of xylem development (Malinowski *et al.*, 2012). Conversely, higher activity of lignification-related genes occurs in less susceptible plants (Chen *et al.*, 2015; Song *et al.*, 2016).

On inoculation, amino acid transport and metabolism vary between tolerant and susceptible hosts, i.e. arginine and proline metabolism are less active in less susceptible *B. rapa* than in susceptible genotypes (Chen *et al.*, 2015; Jubault *et al.*, 2008; Song *et al.*, 2016). Arginine and proline biosynthesis in *Pl. brassicae* also seems to be incomplete (Rolfe *et al.*, 2016; Schwelm *et al.*, 2015). Similar to other gall-forming plant diseases, galled roots also provoke hypoxic responses (Gravot *et al.*, 2016). Infections by *Pl. brassicae* and morphogenic changes within roots leading to gall formation are accompanied by changes in phytohormone homeostasis of auxin, cytokinin and brassinosteroids (Agarwal *et al.*, 2011; Ludwig-Müller *et al.*, 2009; Schuller *et al.*, 2014), but the exact mechanisms are not yet known. The contributions of plant hormones in clubroot have been addressed using *Arabidopsis* mutants altered in phytohormone biosynthesis, metabolism and signalling (Ludwig-Müller *et al.*, 2017). In *Arabidopsis*, elevated cytokinins are associated with increased cell division early during infection. When galls are formed, the expression of host cytokinin biosynthetic genes is repressed, as is the expression of host cytokinin oxidases and dehydrogenases (Devos *et al.*, 2006; Siemens *et al.*, 2006). *Plasmodiophora brassicae*-produced cytokinins probably play a minor role in cytokinin homeostasis in infected tissues (Malinowski *et al.*, 2016). *Arabidopsis* mutants of auxin conjugate synthesis, as well as auxin receptors, were more susceptible to the pathogen (Jahn *et al.*, 2013), whereas nitrilase mutants were more tolerant (Grsic-Rausch *et al.*, 2000). A *Pl. brassicae* protein can conjugate auxin and jasmonic acid to amino acids *in vitro* (Schwelm *et al.*, 2015), but whether it manipulates host hormones in clubroots is unknown.

Effector-triggered immunity is likely to be important in host resistance to *Pl. brassicae*. During infection, resistance (*R*) genes and pathogen-related (*PR*) genes are expressed more strongly in tolerant than in susceptible plants, whereas the pathogen-associated molecular pattern (PAMP)-triggered immune response appears to be similar in both host types (Chen *et al.*, 2015; Zhang *et al.*, 2016b).

One *Pl. brassicae* effector candidate is a predicted secreted methyltransferase, PbBSMT. Biochemical expression assays have shown that this protein can mediate the methylation of salicylic acid (SA) (Ludwig-Müller *et al.*, 2015). PbBSMT might interfere with SA signalling in infected root tissue. SA-mediated pathways are involved in resistance to *Pl. brassicae* (Agarwal *et al.*, 2011; Lemarié *et al.*, 2015; Lovelock *et al.*, 2013). Accordingly, SA-responsive gene expression is increased in tolerant hosts (Chen *et al.*, 2015; Song *et al.*, 2016) and higher SA levels during early infection correlate with resistance (Chen *et al.*, 2015; Zhang *et al.*, 2016b).

Spongospora subterranea

Spongospora subterranea causes powdery scab of potato tubers (*Solanum tuberosum*) (Fig. 4A), an important blemish disease in most major potato-growing regions worldwide. This disease can result in the rejection of whole seed potato lots. The pathogen also causes root gall (Fig. 4B) and is the vector for the *Potato mop top virus* (PMTV, *Pomovirus*, *Virgaviridae*) (Merz and Falloon, 2009; Tamada and Kondo, 2013). Root membrane dysfunction, which reduces water uptake and plant growth, has also been attributed to *S. subterranea* (Falloon *et al.*, 2016). All of these diseases devalue potato crops, causing potato tuber yield losses of >20% in severely diseased crops (Johnson and Cummings, 2015; Merz and Falloon, 2009; Shah *et al.*, 2012). Mature tuber lesions and root galls are filled with clusters of resting spores (sporosori; Fig. 4D), each containing a primary zoospore. Root infection results in the development of zoosporangia (Fig. 4C) producing secondary zoospores. Both types of zoospore infect the host tuber, root epidermis cells and root hairs, and can transmit PMTV.

Disease management is mainly preventative through the use of disease-free seed tubers and non-contaminated fields. Powdery scab and root gall susceptibility differ across potato cultivars (Bittara *et al.*, 2016; Falloon *et al.*, 2003), but no genetic basis of resistance has yet been identified. Metabolites of potato root exudates induce *S. subterranea* resting spore germination, but as L-glutamine and tyramine have the strongest effects, this might not be host specific (Balendres *et al.*, 2016). This may explain reports of primary infection by *S. subterranea* in a range of non-solanaceous host plants (Merz and Falloon, 2009).

Spongospora subterranea ITS rDNA and microsatellite analyses indicate much greater genetic diversity in South American strains (the presumed origin of this pathogen) than elsewhere (Bulman and Marshall, 1998; Gau *et al.*, 2013). After the initial dispersal from South America, Europe was probably the main source of spread of *S. subterranea* (Gau *et al.*, 2013). Molecular data suggest possible substructures between root gall and tuber scab causing *S. subterranea* lineages from South America (Gau *et al.*, 2013). Evidence for recombination in *S. subterranea* is limited, and there is little understanding of sexual recombination in phyto-myxids (Bulman and Braselton, 2014).

Limited genomic sequences, including an assembled mitochondrial genome, are available from *S. subterranea* (Bulman *et al.*, 2011; Gutiérrez *et al.*, 2014, 2016). By comparison, relatively comprehensive *S. subterranea* transcriptomic datasets are available from root galls (Burki *et al.*, 2010; Schwelm *et al.*, 2015). As for *Pl. brassicae*, the current data suggest intron-rich genes, a paucity of CAZymes, but an enrichment of chitin-related enzymes in *S. subterranea*. By contrast, transposable elements are likely to be more common and expressed in *S. subterranea* than in *Pl. brassicae* (Bulman *et al.*, 2011; Gutiérrez *et al.*, 2014; Schwelm *et al.*, 2015). For *S. subterranean*, 613 secreted proteins were

predicted – enriched in ankyrin and protein domains – typical of effectors from other plant pathogens. Few are shared with *Pl. brassicae*, but a putative PbBSMT homologue was detected.

Although no genome has been published, genome sequences from *S. subterranea* are being generated. These will identify *S. subterranea*-specific features and allow research of the transcriptional interaction with its hosts.gg

Polymyxa spp

The genus *Polymyxa* includes two morphologically indistinguishable agriculturally important species: *Polymyxa graminis* (Fig. 5) and *Polymyxa betae*. Both differ in their rDNA sequences and host ranges. The host range of *Px. betae* is restricted to Chenopodiaceae and related plants, whereas *Px. graminis* infects mainly graminaceous plants (Legreve *et al.*, 2000, 2002). Infection by these obligate root endoparasites is asymptomatic (Desoignies, 2012). Unlike *Pl. brassicae* and *S. subterranea*, *Polymyxa* spp. do not cause root galls on infected hosts, but indirectly cause damage as vectors of plant viruses. *Polymyxa graminis* transmits viruses belonging to *Benyvirus*, *Bymovirus*, *Furovirus* and *Pecluvirus*. They include economically important viruses of different grain crops, such as *Barley yellow mosaic virus* (BaYMV) and *Soil-borne wheat mosaic virus* (SBWMV), and also cause virus diseases on other cereals, sugar cane and peanuts [*Peanut clump virus* (PCV)] (Dieryck *et al.*, 2011; Tamada and Kondo, 2013). *Polymyxa betae* transmits *Beet necrotic yellow vein virus* (BNYVV), causing 'rhizomania' in sugar beet (McGrann *et al.*, 2009).

Polymyxa betae is a well-defined species, whereas, in *Px. graminis*, five *formae speciales* or six ribotypes exist, with sub-type classifications based on ecological, molecular and biological characteristics, including specificity in virus transmission (Cox *et al.*, 2014; Dieryck *et al.*, 2011; Kanyuka *et al.*, 2003; Legreve *et al.*, 2002; Smith *et al.*, 2013; Vaianopoulos *et al.*, 2007; Ward *et al.*, 2005; Ziegler *et al.*, 2016).

Obtaining genomic data from *Polymyxa* spp. is more difficult than for the gall-forming plasmodiophorids as high-density infections with substantial amounts of parasite DNA cannot be identified. *Polymyxa betae* cultures on sugar beet hairy roots (Desoignies and Legreve, 2011) and in its non-natural host *A. thaliana* (Desoignies and Legreve, 2011; Smith *et al.*, 2011) were tested, but were difficult to maintain. A suppression subtractive hybridization experiment identified most currently known *Polymyxa* gene models (Desoignies *et al.*, 2014), including 76 *Px. betae* and 120 sugar beet expressed sequence tags (ESTs) putatively involved in the early stages of the host–pathogen interaction. The *Px. betae* ESTs included chitin synthase, polysaccharide deacetylases, ankyrins and galactose lectin domain-encoding transcripts, proteins which are also enriched in *Pl. brassicae* and *S. subterranea* (Bulman *et al.*, 2011; Desoignies *et al.*, 2014; Schwelm *et al.*, 2015). Genes encoding for profilin and a von

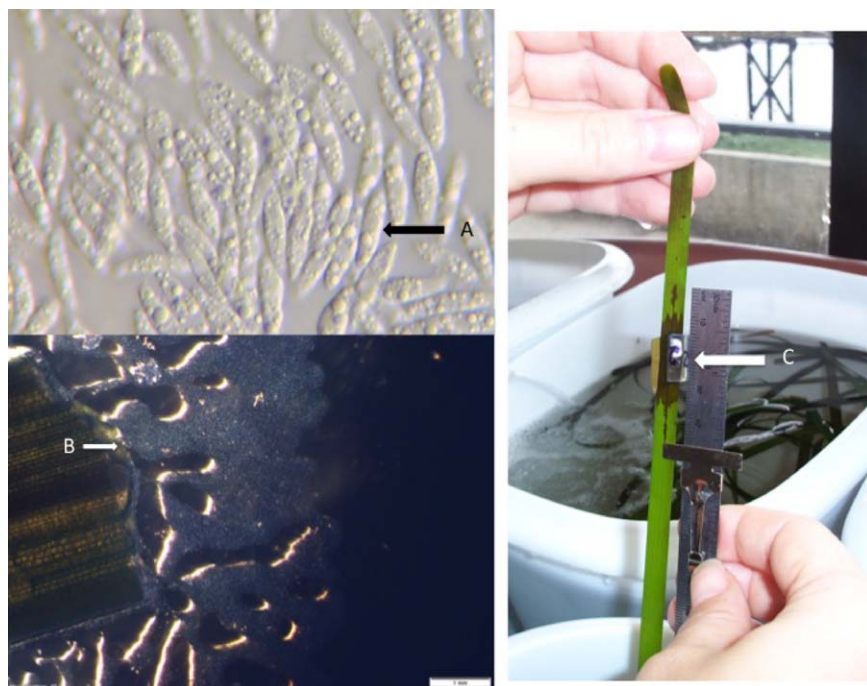


Fig. 6 *Labyrinthula* and disease symptoms. (A) Single fusiform cells of the unicellular Labyrinthulomycota *Labyrinthula* protist. (B) *Labyrinthula* cells emerging from a seagrass leaf on serum seawater agar. Cells move through colonies of self-generated ectoplasmic networks or 'slimeways', a net-like tube within which *Labyrinthula* are able to move. (C) Symptoms of the seagrass wasting disease 4 days following the artificial infection of seagrass blades.

Willebrand factor domain-containing protein were also highly expressed. The sugar beet response to *Px. betae* infection, especially during the plasmodial stage, includes the over-expression of some defence genes, including those that encode PR proteins or lectins (Desoignes *et al.*, 2014).

Other Phytomyxea

Other phytomyxids infect freshwater and marine organisms (Neuhauser *et al.*, 2011). *Maullinia ectocarpii* (Fig. 3) and *M. brasseltonii* are plasmodiophorids infecting brown algae. *Plasmodiophora diplantherea*, *Pl. bicaudata*, *Pl. halophile* and *Tetramyxa parasitica* cause galls on seagrasses, and, in the case of *T. parasitica*, also other estuarine plants (Bulman and Braselton, 2014; Neuhauser *et al.*, 2010). *Spongospora nasturtii* causes crook root on watercress and transmits the *Watercress yellow spot virus* (Walsh *et al.*, 1989), impacting watercress cultivation.

Stramenopiles – *Labyrinthula*

Labyrinthula spp. are protists in the Labyrinthulida (Stramenopila), and are phylogenetically basal to oomycetes (Pan *et al.*, 2017; Tsui *et al.*, 2009). High-throughput environmental DNA sampling, ITS and ribosomal sequences suggest that *Labyrinthula* spp. are highly diverse, and globally distributed (Bockelmann *et al.*, 2013; Collado-Mercado *et al.*, 2010; Martin *et al.*, 2016; Pan *et al.*, 2017). These organisms are saline tolerant, and can be saprobes, coral inhabitants, endosymbionts of amoebae or endophytic facultative parasites of marine and terrestrial plants (Amon, 1978; Bigelow *et al.*, 2005; Pan *et al.*, 2017; Sullivan *et al.*, 2013).

Marine *Labyrinthula*, such as *L. zosterae*, which causes seagrass wasting disease (SWD) (Sullivan *et al.*, 2013), are usually associated with mangrove, macroalgal and seagrass ecosystems (Lindholm *et al.*, 2016; Pan *et al.*, 2017). Rapid blight disease (RBD) in turfgrasses is caused by the terrestrial species *L. terrestris* in high-salinity environments, such as salt lakes and golf course turf (Douhan *et al.*, 2009; Kerrigan *et al.*, 2012). This pathogen may have become important in specialized turfgrass because of increased salinity in irrigation or the use of reclaimed water, causing increased turf salinification (Olsen, 2007; Stowell *et al.*, 2005). Both *L. zosterae* and *L. terrestris* vary greatly in virulence to their hosts (Chitrampalam *et al.*, 2015; Douhan *et al.*, 2009; Martin *et al.*, 2016). Although the exact mechanism is uncertain, SWD and RBD manifest through the penetration of host leaf epidermis cells of individual *Labyrinthula* cells.

After infection, *Labyrinthula* spp. destroy the host chloroplasts and advance to neighbouring cells. This creates lesions, sometimes killing entire leaves or plants through interruption of photosynthesis (Fig. 6). These pathogens are therefore found on the edges of progressing infections rather than within the host lesions (Muehlstein, 1992; Olsen, 2007; Sullivan *et al.*, 2016). They can be isolated from infected leaf tissues as they emerge from tissues plated onto serum seawater agar solutions (Fig. 6). The individual spindle- to oval-shaped *Labyrinthula* cells move through colonies of self-generated ectoplasmic networks or 'slimeways', which are thought to originate from specialized organelles called bothriosomes. In conjunction with pseudopodium extension, a net-like tube is created within which the cells move. The movement of cells occurs through the utilization of an actomyosin system

(Preston and King, 2005). The slimeways are also thought to aid nutrient absorption (Vishniac, 1955). *Labyrinthula* cells contain two vacuoles, thought to serve as excretory organs in the cell and may also regulate osmotic pressure, as their presence depends on the environmental salinity (Young, 1943).

The seagrass–*Labyrinthula* pathosystem is the best-studied relationship for this group. Quantitative polymerase chain reaction (PCR) has shown that *Labyrinthula* spp. occur in most marine eelgrass populations in Europe, but pathogenic species may only cause disease when infection is coupled with host stress (Bockelmann *et al.*, 2013; Brakel *et al.*, 2014). However, the potential impact of SWD was observed in the 1930s, when *Labyrinthula* killed up to 90% of *Zostera marina*, the most abundant Northern Hemisphere seagrass (reviewed in Muehlstein, 1989; Sullivan *et al.*, 2013). Seagrass meadows are ecologically rich and productive marine ecosystems, and important carbon sinks (Christianen *et al.*, 2013; Fourqurean *et al.*, 2012). They support commercial fish nurseries (Jackson *et al.*, 2001) and influence bacterial pathogen populations (Lamb *et al.*, 2017). Despite the important ecological and economic roles of their hosts, and widespread evidence of their cause of severe disease, research in *Labyrinthula* pathology is still under development.

Labyrinthula spp. tolerate high temperatures up to 28 °C, but, in tropical and subtropical seagrasses, increased temperature results in reduced virulence (Olsen and Duarte, 2015). Low salinity also inhibits *Labyrinthula* growth (Muehlstein *et al.*, 1988), and so seagrass meadows in high-salinity waters may have an advantage compared with those in truly marine locations (Vergeer *et al.*, 1995). The transcriptomic host response to a *Labyrinthula* infection of seagrasses includes the down-regulation of genes related to reactive oxygen species (ROS) and chitinases, whereas a phenolic acid synthesis gene is highly expressed (Brakel *et al.*, 2014). Phenolic metabolites may produce 'synergistic' host benefits. Resistance to *Labyrinthula* is density dependent, and diseased leaves have enhanced phenolic metabolite concentrations and these may reduce host susceptibility to *Labyrinthula* (Groner *et al.*, 2016; McKone and Tanner, 2009; Trevathan-Tackett *et al.*, 2015). The first seagrass genome (of *Z. marina*) has been published recently (Olsen *et al.*, 2016). As a host for *Labyrinthula*, this expands the ability to investigate the genetic and molecular interactions between *Labyrinthula* and seagrass, and to improve our understanding of this potentially devastating pathogen.

Stramenopiles – oomycetes as algal parasites

Oomycetes cause considerable damage in aquatic crops, including red (Rhodophyta) and brown (Phaeophyceae) algae. Worldwide algal industries have increased dramatically (Loureiro *et al.*, 2015). In 2012, global macroalgal production was more than 23 million tonnes (dry weight), with a market value greater than six billion US\$ (FAO, 2014). Most of this production (approximately 80%) is

used for human consumption, and the remainder for fertilizers, animal feed additives and in medical and biotechnological applications, including biofuel production (Loureiro *et al.*, 2015; Stengel and Connan, 2015). Seaweed farming is also often integrated into fish and shellfish aquaculture (Loureiro *et al.*, 2015). The total market value for red seaweed reached 3.8 billion US\$ (FAO, 2014). Best known in the form of Nori (sushi wrap), *Pyropia* (formerly *Porphyra*) spp. are the most common cultivated red algae. Brown algae are often the predominant primary producers in temperate and cold marine coastal ecosystems (Rodgers and Shears, 2016), and are phylogenetically distant from plants, green and red algae. They differ from red and green algae in cell wall composition (Michel *et al.*, 2010), halogen metabolism (La Barre *et al.*, 2010), oxylipin synthesis (Ritter *et al.*, 2008) and life cycles (Coelho *et al.*, 2011). Brown algae include edible seaweeds (e.g. kombu – *Undaria pinnatifida*, wakame – *Saccharina japonica* and sugar kelp – *Saccharina latissima*), and some species are commercially used to produce alginate. Collectively, red and brown algae are affected by many diseases (reviewed in Gachon *et al.*, 2010). Because of the economic importance of *Pyropia* cultivation, and the growing economic burden of diseases for this crop (up to 50% of farm costs are spent on disease management: Kim *et al.*, 2014), this review focuses on *Pythium porphyrae* and *Olpidiopsis* sp., the two main oomycetes that cause diseases on this crop.

Olpidiopsis diseases (previously 'chytrid rot') caused Korean Nori farms to lose nearly 25% of their resale value in 2012–2013 (Kim *et al.*, 2014), but local losses can be greater (Klochkova *et al.*, 2012; Loureiro *et al.*, 2015). Environmental factors, such as temperature and seasonality, affect the severity of disease outbreaks.

Pythium porphyrae causes red rot disease, which is one of the most damaging diseases affecting *Pyropia* farming (Fig. 7) with production losses being greater than 20% (Kawamura *et al.*, 2005). Distinct bleached patches on the algal blades characterize the initial infections. The diversity of *Olpidiopsis* is beginning to be described using molecular tools, with the recognition of new species, such as *O. pyropiae* from Korean farms (Klochkova *et al.*, 2016; Sekimoto *et al.*, 2008), in addition to the Japanese *O. porphyrae*.

Olpidiopsis spp

Olpidiopsis pathogens are obligate intracellular pathogens with biotrophic lifestyles. During the off-season of algal cultivation, *Olpidiopsis* may survive in alternative red algal hosts (e.g. *Heterosiphonia* sp.) or as dormant cysts (Klochkova *et al.*, 2012, 2016). Germinating zoospores form germ tubes which penetrate algal cell walls. Within the cells, multinucleate walled thalli form, which quickly develop into sporangia, which release zoospores. With advancing infection, host cells break down and lesions in the blades become prominent.

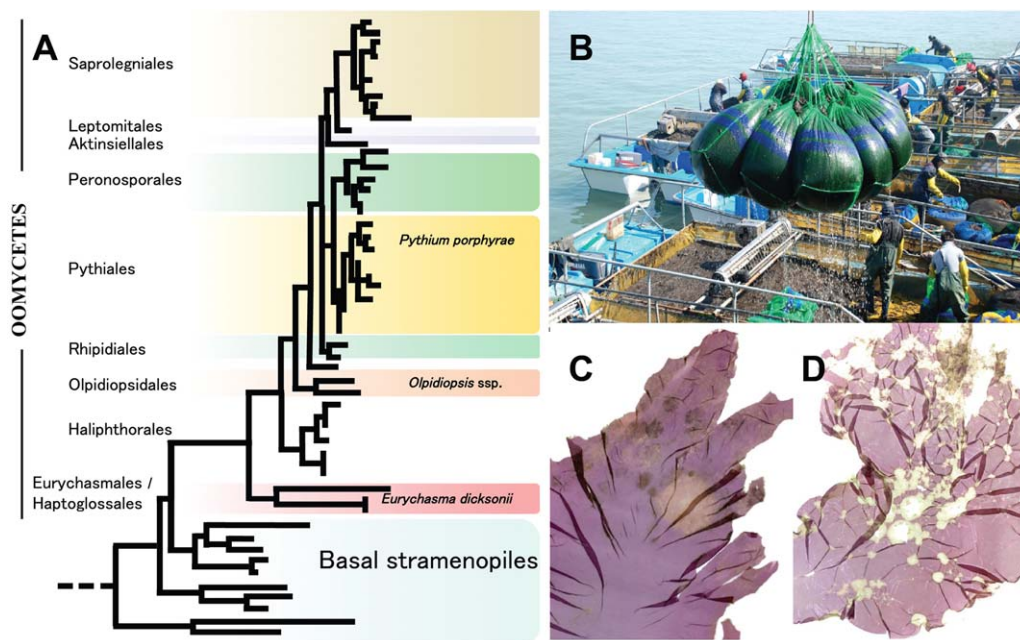


Fig. 7 Oomycete phylogeny, *Pyropia* farming, *Pythium porphyrae* and *Olpidiopsis* symptoms. (A) Schematic phylogenetic tree of Oomycetes based on Beakes *et al.* (2012) indicating the positions of the discussed pathogens of marine algae. (B) *Pyropia* seaweed harvest on a commercial farm in South Korea (photograph: H. Kim). (C, D) *Pyropia* blade with lesions caused by *Pythium porphyrae* (C) and *Olpidiopsis* (D) infection. Photographs were originally published in Kim *et al.* (2014) which includes more detailed descriptions of *Pyropia* diseases.

The establishment of *Olpidiopsis* sp. and *Pyropia* pathosystems for research is difficult as the infected host disintegrates in a matter of days. However, with alternative hosts, such as *Heterosiphonia japonica*, stable dual cultures can be achieved (Klochkova *et al.*, 2012). *Olpidiopsis* infection in this system is cell type specific, and occurs on the extended rhizoid-like apical cells. This specificity has been attributed to d(+)-mannose in host cell walls, indicating a specific lectin–carbohydrate interaction during host–parasite recognition, necessary for zoospore attachment to host cells (Klochkova *et al.*, 2012). Until recently, the only available treatment for these diseases was to wash algal blades with acid, a practice now banned because of environmental concerns (Kim *et al.*, 2014).

Pythium porphyrae

Red rot disease, caused by *Py. porphyrae*, was first described by Arasaki (1947). The disease spreads via zoospores and starts with distinct, small, red patches on the host blades in which the zoospores germinate. The pathogen develops extensive cell-to-cell spreading mycelium. Dead host cells change colour to violet–red and green before they degenerate, generating holes that finally destroy entire blades (Fig. 7). Red rot disease management is only effective during the early stages of infection, and PCR methods are important to detect the pathogen early during the algal cultivation period (Park *et al.*, 2001, 2006). Disease control involves immersion of cultivation nets into organic acid, freezing of

infected cultures and the application of fungicides (Amano *et al.*, 1995; Hwang *et al.*, 2009; Park *et al.*, 2006). These treatments have significant costs and environmental impacts (Park and Hwang, 2015). Disease-resistant host cultivars are an alternative control strategy. Partially resistant *Pyropia yezoensis* cultivars, generated from living cells in lesions of infected tissue, have altered cell wall polysaccharide contents (Park and Hwang, 2015). Sulfated galactans (e.g. porphyran) of the algal cell walls may be essential for cyst attachment and infection of *Py. porphyrae*, although the attraction and contact of zoospores are independent of host exudates (Uppalapati and Fujita, 2000). On resistant *Pyropia* sp., cysts with germ tubes frequently grow on the host thallus surfaces without penetration, and show no or delayed induction of appressoria (Uppalapati and Fujita, 2001). Although *Py. porphyrae* zoospores attach and encyst on a number of red algal species, red rot disease only develops on *Pyropia yezoensis* and *Bangia atropurpurea* (Uppalapati and Fujita, 2000).

Pythium porphyrae grows best in low-salinity water, possibly explaining why red rot occurs in farms near river banks (Klochkova *et al.*, 2017). The pathogen can also infect and grow on land plants, including Chinese cabbage and rice. *Pythium porphyrae* carried from the land into coastal waters may increase damage in seaweed farms close to river inlets (Klochkova *et al.*, 2017). This could enable molecular research on *Py. porphyrae* using the model hosts rice and *A. thaliana*. Genomic data are already available for *Pyropia* hosts (<http://dbdata.rutgers.edu/nori/index.php>)

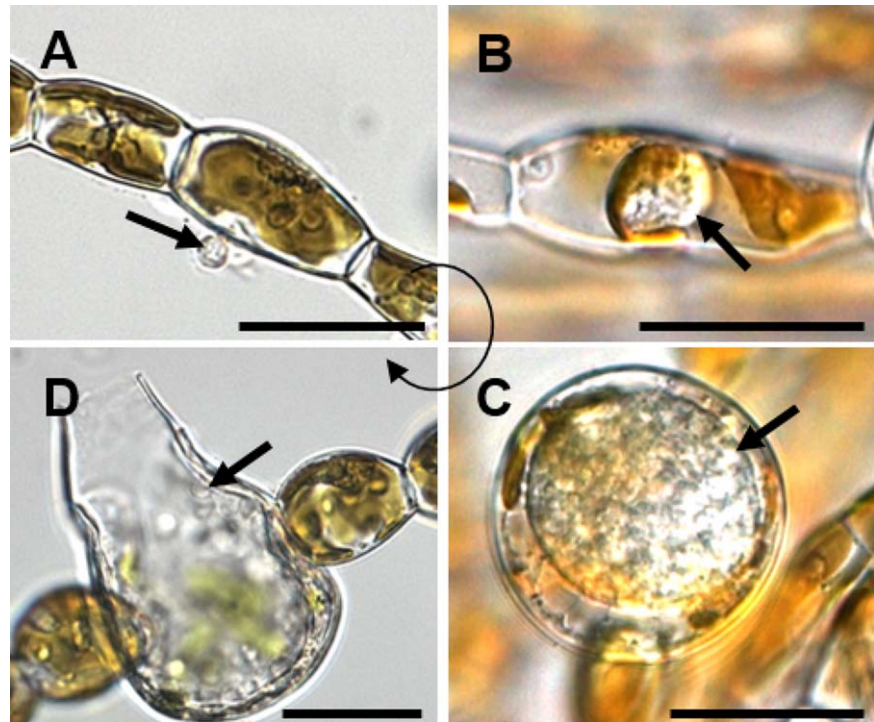


Fig. 8 Life cycle of *Eurychasma dicksonii* in its brown algal host *Ectocarpus siliculosus*. (A) A spore (arrow) attaches to the algal surface and injects its content into the host. (B) Within the algal cytoplasm, the *Eu. dicksonii* thallus (arrow) develops which, at the early stage of infection, is unwallled. (C) Later, the pathogen thallus (arrow) has a cell wall and causes hypertrophic expansion of the algal host cell. (D) At the final stage, the complete thallus differentiates into a sporangium from which motile zoospores (arrow) are produced, completing the life cycle of the pathogen. Scale bars equal to 25 μm . (Figure reproduced from Strittmatter *et al.* 2016.)

(Nakamura *et al.*, 2013; Wu *et al.*, 2014) and are currently being generated for *Py. porphyrae* and *Olpidiopsis* sp.

Eurychasma dicksonii

The most frequently recorded eukaryotic pathogen of brown algae is the biotrophic oomycete *Eurychasma dicksonii*. This phylogenetically basal oomycete (Beakes *et al.*, 2012) is geographically widespread, tolerates a broad temperature range (4–23 °C) and infects at least 45 different species of brown algae in laboratory cultures (Müller *et al.*, 1999). Similar to *Olpidiopsis* spp., *Eu. dicksonii* is a holocarpic endoparasite (Sekimoto *et al.* 2008). Zoospores attach, encyst and build adhesion-like structures at the host surfaces. The parasite cytoplasm is transferred into the host via a needle-like structure which is associated with the formation of the adhesion chamber at the host–spore contact point (Tsirigoti *et al.*, 2015), similar to the plasmodiophorid ‘Rohr und Stachel’. After penetration, multinucleate non-walled immature thalli, with double membrane envelopes of host and parasite (Sekimoto *et al.*, 2008), develop and expand in the infected host cells, until each cell is almost filled. The plasmodial thallus develops into a sporangium with peripheral primary cysts (Fig. 8), which release biflagellate zoospores through apical exit tubes. The empty cyst walls form a net-like sporangium, which is a distinctive morphological feature of this pathogen (Petersen, 1905).

Eurychasma dicksonii can be cultured in *Ectocarpus siliculosus*, the first brown alga to be genomically sequenced (Cock *et al.*, 2010), explaining why the *Eurychasma*–*Ectocarpus* pathosystem is

the most thoroughly investigated parasitic interaction in brown algae. A cDNA analysis of *Ec. siliculosus* infected with *Eu. dicksonii* identified 3086 unigenes of oomycete origin. The dataset of *Eu. dicksonii* included 351 proteins predicted to be secreted, but contained no CRN or RxLR effector candidates (Grenville-Briggs *et al.*, 2011). The *Eu. dicksonii* genes included glucanases and a potential alginate lyase, for which no homologues in land plant-infecting oomycetes have been identified. Alginates and glucans are key components of brown algal cell walls. Similar to higher oomycetes, which secrete cell wall-degrading enzymes involved in host penetration, this lyase is probably an adaptation to the marine host. In brown algae, β -1,3-glucans are usually not part of the cell walls, but are storage polysaccharides. Cell wall modification is a putative host defence mechanism against *Eu. dicksonii*. On infection, cell wall thickening and increased amounts of β -1,3-glucans at the penetration site may build physical barriers to pathogen invasion. Large amounts of β -1,3-glucan occur at cell surfaces of partially resistant *Ectocarpus* strains (Tsirigoti *et al.*, 2015).

Although the infection mechanisms remain largely unexplored, molecular data exist on the host response to infection by *Eu. dicksonii*. Host genes differentially expressed during infection include those encoding for proteins involved in the detoxification of ROS and halogen metabolism (Strittmatter *et al.*, 2016). The host genome includes candidate immune receptors of the leucine-rich and tetratricopeptide repeat families, which quickly evolve via an original exon shuffling mechanism (Zambounis *et al.*, 2012).

Different hosts display different levels of susceptibility to *Eurychasma* (Gachon *et al.*, 2009), and the resistance mechanisms are currently being investigated using cytological and molecular approaches. A targeted movement of host nuclei to pathogen penetration sites has been observed (Grenville-Briggs *et al.*, 2011), and microtubule disorganization in the host occurs only when zoospore germination of the pathogen begins (Tsirigoti *et al.*, 2015).

OUTLOOK

Our understanding of eukaryotic plant pathogens is built on studies of fungi, animals (both opisthokonts) and oomycetes (stramenopiles). For the plant pathogens introduced here, the biochemical interactions with their plant hosts are just beginning to be unravelled through the introduction of study systems (e.g. the *Eu. dicksonii*-brown algae interaction) or the generation of reference genomes (*Pl. brassicae*, *Phytomonas* spp.). This will allow the presented pathogens to take a more prominent place in the molecular plant pathology field in the coming years, create deeper insights into how these pathogens interact with their hosts and how they have evolved. This should finally lead to new strategies for the control of these pathogens.

AUTHOR CONTRIBUTIONS

A.S. and S.N. initiated and organized the manuscript, and the other authors are listed alphabetically. Section contributions are as follows: *Phytomonas* (J.L., A.N., A.S.), plasmodiophorids (A.S., S.N., S.B., R.E.F., U.M., N.D., A.L.), *Labyrinthula* (S.N., B.K.S.), oomycetes (C.M.M.G., M.S., A.S., S.N., J.B., M.E.). All the authors read the manuscript and agreed to publication.

ACKNOWLEDGEMENTS

A.S. was funded by Formas, the Swedish Research Council. S.N., J.B. and M.E. were funded by the Austrian Science Fund (grant Y0810-B16). S.B. and R.E.F. were funded by the New Zealand Ministry for Business Innovation and Employment (Programme LINX0804). J.L. was supported by the Czech Grant Agency award 15-21974 and the ERC CZ LL1601. We would like to thank Sandra Baldauf, Gwang Hoon Kim and Monica L. Elliott for providing the photographs used in the figures. The authors have no conflicts of interest to declare.

REFERENCES

- Adl, S.M., Simpson, A.G.B., Lane, C.E., Lukes, J., Bass, D., Bowser, S.S., Brown, M.W., Burki, F., Dunthorn, M., Hampl, V., Heiss, A., Hoppenrath, M., Lara, E., Le Gall, L., Lynn, D.H., McManus, H., Mitchell, E.A., Mozley-Stanridge, S.E., Parfrey, L.W., Pawlowski, J., Rueckert, S., Shadwick, L., Schoch, C.L., Smirnov, A. and Spiegel, F.W. (2012) The revised classification of eukaryotes. *J. Eukaryot. Microbiol.* **59**, 429–493.
- Agarwal, A., Kaul, V., Faggian, R., Rookes, J.E., Ludwig-Muller, J. and Cahill, D.M. (2011) Analysis of global host gene expression during the primary phase of the *Arabidopsis thaliana*-*Plasmodiophora brassicae* interaction. *Funct. Plant Biol.* **38**, 462–478.
- Aist, J.R. and Williams, P.H. (1971) The cytology and kinetics of cabbage root hair penetration by *Plasmodiophora brassicae*. *Can. J. Bot.* **49**, 2023–2034.
- Amano, H., Suginaga, R., Arashima, K. and Noda, H. (1995) Immunological detection of the fungal parasite, *Pythium* sp. – the causative organism of red rot disease in *Porphyra-yezoensis*. *J. Appl. Phycol.* **7**, 53–58.
- Amon, J.P. (1978) Thraustochytrids and labyrinthulids of terrestrial, aquatic and hypersaline environments of the Great Salt Lake, USA. *Mycologia*, **70**, 1299–1301.
- Arasaki, S. (1947) Studies on the rot of *Porphyra tenera* by *Pythium*. *Nippon Suisan Gakkaishi*, **13**, 74–90.
- Balendres, M.A., Nichols, D.S., Tegg, R.S. and Wilson, C.R. (2016) Metabolomes of potato root exudates: compounds that stimulate resting spore germination of the soil-borne pathogen *Spongospora subterranea*. *J. Agric. Food Chem.* **64**, 7466–7474.
- Beakes, G.W., Glockling, S.L. and Sekimoto, S. (2012) The evolutionary phylogeny of the oomycete “fungi”. *Protoplasma*, **249**, 3–19.
- Bi, K., He, Z., Gao, Z., Zhao, Y., Fu, Y., Cheng, J., Xie, J., Jiang, D. and Chen, T. (2016) Integrated omics study of lipid droplets from *Plasmodiophora brassicae*. *Sci. Rep.* **6**, 36 965.
- Bigelow, D., Olsen, M. and Gilbertson, R. (2005) *Labyrinthula terrestris* sp. nov., a new pathogen of turf grass. *Mycologia*, **97**, 185–190.
- Bittara, F.G., Thompson, A.L., Gudmestad, N.C. and Secor, G.A. (2016) Field evaluation of potato genotypes for resistance to powdery scab on tubers and root gall formation caused by *Spongospora subterranea*. *Am. J. Potato Res.* **93**, 497–508.
- Bockelmann, A.-C., Tams, V., Ploog, J., Schubert, P.R. and Reusch, T.B. (2013) Quantitative PCR reveals strong spatial and temporal variation of the wasting disease pathogen, *Labyrinthula zosterae* in northern European eelgrass (*Zostera marina*) beds. *PLoS One*, **8**, e62169.
- Brakel, J., Werner, F.J., Tams, V., Reusch, T.B.H. and Bockelmann, A.C. (2014) Current European *Labyrinthula zosterae* are not virulent and modulate seagrass (*Zostera marina*) defense gene expression. *PLoS One*, **9**, e92448.
- Bulman, S. and Braselton, J.P. (2014) Rhizaria: Phytomyxea. In: *The Mycota VII, Part A, Systematics and Evolution* (McLaughlin, D. J. and Spatafora, J. W., eds.), pp. 99–112. Springer Berlin Heidelberg.
- Bulman, S., Candy, J.M., Fiers, M., Lister, R., Conner, A.J. and Eady, C.C. (2011) Genomics of biotrophic, plant-infecting plasmodiophorids using *in vitro* dual cultures. *Protist*, **162**, 449–461.
- Bulman, S.R. and Marshall, J.W. (1998) Detection of *Spongospora subterranea* in potato tuber lesions using the polymerase chain reaction (PCR). *Plant Pathol.* **47**, 759–766.
- Burki, F. and Keeling, P.J. (2014) Rhizaria. *Curr. Biol.* **24**, R103–R107.
- Burki, F., Kudryavtsev, A., Matz, M.V., Aglyamova, G.V., Bulman, S., Fiers, M., Keeling, P.J. and Pawlowski, J. (2010) Evolution of Rhizaria: new insights from phylogenomic analysis of uncultivated protists. *BMC Evol. Biol.* **10**, 377.
- Butler, C.E., Jaskowska, E. and Kelly, S. (2017) Genome sequence of *Phytomonas françai*, a cassava (*Manihot esculenta*) latex parasite. *Genome Announc.* **5**, e01266–16.
- Camargo, E.P. (1999) *Phytomonas* and other trypanosomatid parasites of plants and fruit. *Adv. Parasitol.* **42**, 29–112.
- Carré, P. and Pouzet, A. (2014) Rapeseed market, worldwide and in Europe. *OCL*, **21**, D102.
- Chen, J., Pang, W., Chen, B., Zhang, C. and Piao, Z. (2015) Transcriptome analysis of *Brassica rapa* near-isogenic lines carrying clubroot-resistant and -susceptible alleles in response to *Plasmodiophora brassicae* during early infection. *Front. Plant Sci.* **6**, 1183.
- Chitrapalam, P., Goldberg, N. and Olsen, M.W. (2015) *Labyrinthula* species associated with turfgrasses in Arizona and New Mexico. *Eur. J. Plant Pathol.* **143**, 485–493.
- Christianen, M.J.A., van Belzen, J., Herman, P.M.J., van Katwijk, M.M., Lamers, L.P.M., van Leent, P.J.M. and Bouma, T.J. (2013) Low-canopy seagrass beds still provide important coastal protection services. *PLoS One*, **8**, e62413.
- Cock, J.M., Sterck, L., Rouze, P., Scornet, D., Allen, A.E., Amoutzias, G., Anthouard, V., Artiguenave, F., Aury, J.M., Badger, J.H., Beszteri, B., Billiau, K., Bonnet, E., Bothwell, J.H., Bowler, C., Boyen, C., Brownlee, C., Carrano, C.J., Charrier, B., Cho, G.Y., Coelho, S.M., Collén, J., Corre, E., Da Silva, C., Delage, L., Delarouge, N., Dittami, S.M., Doubeau, S., Elias, M., Farnham, G., Gachon, C.M., Gschloessl, B., Heesch, S., Jabbari, K., Jubin, C., Kawai, H., Kimura, K., Kloareg, B., Küpper, F.C., Lang, D., Le Bail, A., Leblanc, C., Lerouge, P., Lohr, M., Lopez, P.J., Martens, C., Maumus, F., Michel, G., Miranda-Saavedra, D., Morales, J., Moreau, H., Motomura, T., Nagasato, C., Napoli, C.A., Nelson, D.R., Nyvall-Collén, P., Peters, A.F., Pommier, C., Potin, P., Poulain, J., Quesneville, H., Read, B., Rensing, S.A., Ritter, A.,

- Rousvoal, S., Samanta, M., Samson, G., Schroeder, D.C., Ségurens, B., Strittmatter, M., Tonon, T., Tregear, J.W., Valentin, K., von Dassow, P., Yamagishi, T., Van de Peer, Y. and Wincker, P. (2010) The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. *Nature*, **465**, 617–621.
- Coelho, S.M., Godfroy, O., Arun, A., Le Corguillé, G., Peters, A.F. and Cock, J.M. (2011) Genetic regulation of life cycle transitions in the brown alga *Ectocarpus*. *Plant Signal. Behav.* **6**, 1858–1860.
- Collado-Mercado, E., Radway, J.C. and Collier, J.L. (2010) Novel uncultivated labyrinthulomycetes revealed by 18S rDNA sequences from seawater and sediment samples. *Aquat. Microb. Ecol.* **58**, 215–228.
- Cox, B.A., Luo, H. and Jones, R. (2014) *Polymyxa graminis* isolates from Australia: identification in wheat roots and soil, molecular characterization and wide genetic diversity. *Phytopathology*, **98**, 1567–1575.
- Dean, R., Van Kan, J.A.L., Pretorius, Z.A., Hammond-Kosack, K.E., Di Pietro, A., Spanu, P.D., Rudd, J.J., Dickman, M., Kahmann, R., Ellis, J. and Foster, G.D. (2012) The Top 10 fungal pathogens in molecular plant pathology. *Mol. Plant Pathol.* **13**, 414–430.
- Desoignies, N. (2012) *Polymyxa betae* - *Beta vulgaris*: understanding the molecular interactions through transcriptome and plant defense analysis. PhD thesis, Université catholique de Louvain, Belgium.
- Desoignies, N. and Legreve, A. (2011) *In vitro* dual culture of *Polymyxa betae* in *Agrobacterium rhizogenes* transformed sugar beet hairy roots in liquid media. *J. Eukaryot. Microbiol.* **58**, 424–425.
- Desoignies, N., Carbonell, J., Moreau, J.S., Conesa, A., Dopazo, J. and Legreve, A. (2014) Molecular interactions between sugar beet and *Polymyxa betae* during its life cycle. *Ann. Appl. Biol.* **164**, 244–256.
- Devos, S., Laukens, K., Deckers, P., Van der Straeten, D., Beeckman, T., Inze, D., Van Onckelen, H., Witters, E. and Prinsen, E. (2006) A hormone and proteome approach to picturing the initial metabolic events during *Plasmodiophora brassicae* infection on *Arabidopsis*. *Mol. Plant–Microbe Interact.* **19**, 1431–1443.
- Di Lucca, A.G.T., Chipana, E.F.T., Albuja, M.J.T., Peralta, W.D., Piedra, Y.C.M. and Zelada, J.L.A. (2013) Slow wilt: another form of Marchitez in oil palm associated with trypanosomatids in Peru. *Trop. Plant Pathol.* **38**, 522–533.
- Diederichsen, E., Frauen, M., Linders, E.G.A., Hatakeyama, K. and Hirai, M. (2009) Status and perspectives of clubroot resistance breeding in crucifer crops. *J. Plant Growth Regul.* **28**, 265–281.
- Dieryck, B., Weyns, J., Doucet, D., Bragard, C. and Legreve, A. (2011) Acquisition and transmission of peanut clump virus by *Polymyxa graminis* on cereal species. *Phytopathology*, **101**, 1149–1158.
- Dixon, G.R. (2009) The occurrence and economic impact of *Plasmodiophora brassicae* and clubroot disease. *J. Plant Growth Regul.* **28**, 194–202.
- Douhan, G.W., Olsen, M.W., Herrell, A., Winder, C., Wong, F. and Entwistle, K. (2009) Genetic diversity of *Labyrinthula terrestris*, a newly emergent plant pathogen, and the discovery of new Labyrinthulid organisms. *Mycol. Res.* **113**, 1192–1199.
- Faggian, R. and Strelkov, S.E. (2009) Detection and measurement of *Plasmodiophora brassicae*. *J. Plant Growth Regul.* **28**, 282–288.
- Fähling, M., Graf, H. and Siemens, J. (2003) Pathotype separation of *Plasmodiophora brassicae* by the host plant. *J. Phytopathol.* **151**, 425–430.
- Falloon, R.E., Genet, R.A., Wallace, A.R. and Butler, R.C. (2003) Susceptibility of potato (*Solanum tuberosum*) cultivars to powdery scab (caused by *Spongospora subterranea* f. sp. *subterranea*), and relationships between tuber and root infection. *Australas. Plant Pathol.* **32**, 377–385.
- Falloon, R.E., Merz, U., Butler, R.C., Curtin, D., Lister, R.A. and Thomas, S.M. (2016) Root infection of potato by *Spongospora subterranea*: knowledge review and evidence for decreased plant productivity. *Plant Pathol.* **65**, 422–434.
- FAO (2014) Food and Agriculture Organization of the United Nations. Fisheries and Aquaculture Information and Statistics Services. URL <http://www.fao.org/figis/> [accessed on Jul 26, 2014].
- Fourqurean, J.W., Duarte, C.M., Kennedy, H., Marba, N., Holmer, M., Mateo, M.A., Apostolaki, E.T., Kendrick, G.A., Krause-Jensen, D., McGlathery, K.J. and Serrano, O. (2012) Seagrass ecosystems as a globally significant carbon stock. *Nat. Geosci.* **5**, 505–509.
- Gachon, C.M., Strittmatter, M., Muller, D.G., Kleinteich, J. and Kupper, F.C. (2009) Detection of differential host susceptibility to the marine oomycete pathogen *Eurychasma dicksonii* by real-time PCR: not all algae are equal. *Appl. Environ. Microbiol.* **75**, 322–328.
- Gachon, C.M., Sime-Ngando, T., Strittmatter, M., Chambouvet, A. and Kim, G.H. (2010) Algal diseases: spotlight on a black box. *Trends Plant Sci.* **15**, 633–640.
- Gau, R.D., Merz, U., Falloon, R.E. and Brunner, P.C. (2013) Global genetics and invasion history of the potato powdery scab pathogen, *Spongospora subterranea* f. sp. *subterranea*. *PLoS One*, **8**, e67944.
- van de Graaf, P., Wale, S.J. and Lees, A.K. (2007) Factors affecting the incidence and severity of *Spongospora subterranea* infection and galling in potato roots. *Plant Pathol.* **56**, 1005–1013.
- Graf, H., Fähling, M. and Siemens, J. (2004) Chromosome polymorphism of the obligate biotrophic parasite *Plasmodiophora brassicae*. *J. Phytopathol.* **152**, 86–91.
- Gravot, A., Richard, G., Lime, T., Lemarié, S., Jubault, M., Lariagon, C., Lemoine, J., Vicente, J., Robert-Seilantant, A., Holdsworth, M.J. and Manzanares-Dauleux, M.J. (2016) Hypoxia response in *Arabidopsis* roots infected by *Plasmodiophora brassicae* supports the development of clubroot. *BMC Plant Biol.* **16**, 251.
- Grenville-Briggs, L., Gachon, C.M.M., Strittmatter, M., Sterck, L., Kupper, F.C. and van West, P. (2011) A molecular insight into algal–oomycete warfare: cDNA analysis of *Ectocarpus siliculosus* infected with the basal oomycete *Eurychasma dicksonii*. *PLoS One*, **6**, e24500.
- Groner, M.L., Burge, C.A., Kim, C.J.S., Rees, E., Van Alstyne, K.L., Yang, S., Wyllie-Echeverria, S. and Harvell, C.D. (2016) Plant characteristics associated with widespread variation in eelgrass wasting disease. *Dis. Aquat. Organ.* **118**, 159–168.
- Grsic-Rausch, S., Kobelt, P., Siemens, J.M., Bischoff, M. and Ludwig-Müller, J. (2000) Expression and localization of nitrilase during symptom development of the clubroot disease in *Arabidopsis*. *Plant Physiol.* **122**, 369–378.
- Gutiérrez, P., Bulman, S., Alzate, J.F., Ortiz, M.C. and Marín, M. (2016) Mitochondrial genome sequence of the potato powdery scab pathogen *Spongospora subterranea*. *Mitochondrial DNA A DNA Mapp. Seq. Anal.* **27**, 58–59.
- Gutiérrez, P.A., Alzate, J.F. and Montoya, M.M. (2014) Analysis of carbohydrate metabolism genes of *Spongospora subterranea* using 454 pyrosequencing. *Rev. Fac. Nat. Agr. Medellín*. **67**, 7247–7260.
- Hannaert, V., Saavedra, E., Duffieux, F., Szikora, J.P., Rigden, D.J., Michels, P.A.M. and Opperdoes, F.R. (2003) Plant-like traits associated with metabolism of Trypanosoma parasites. *Proc. Natl. Acad. Sci. USA*, **100**, 1067–1071.
- Hatakeyama, K., Suwabe, K., Tomita, R.N., Kato, T., Nunome, T., Fukuoka, H. and Matsumoto, S. (2013) Identification and characterization of Crr1a, a gene for resistance to clubroot disease (*Plasmodiophora brassicae* Woronin) in *Brassica rapa* L. *PLoS One*, **8**, e54745.
- Hwang, E.K., Park, C.S. and Kakinuma, M. (2009) Physicochemical responses of *Pythium porphyrae* (Oomycota), the causative organism of red rot disease in *Porphyra* to acidification. *Aquacult. Res.* **40**, 1777–1784.
- Jackson, A.P. (2015) Genome evolution in trypanosomatid parasites. *Parasitology*, **142**, S40–S56.
- Jackson, E.L., Rowden, A.A., Attrill, M.J., Bossey, S.J. and Jones, M.B. (2001) The importance of seagrass beds as a habitat for fishery species. *Oceanogr. Mar. Biol.* **39**, 269–303.
- Jahn, L., Mucha, S., Bergmann, S., Horn, C., Staswick, P., Steffens, B., Siemens, J. and Ludwig-Müller, J. (2013) The clubroot pathogen (*Plasmodiophora brassicae*) influences auxin signaling to regulate auxin homeostasis in *Arabidopsis*. *Plants*, **2**, 726–749.
- Jashni, M.K., Mehrabi, R., Collemare, J., Mesarich, C.H. and de Wit, P.J. (2015) The battle in the apoplast: further insights into the roles of proteases and their inhibitors in plant–pathogen interactions. *Front. Plant Sci.* **6**, 584.
- Jaskowska, E., Butler, C., Preston, G. and Kelly, S. (2015) Phytomonas: trypanosomatids adapted to plant environments. *PLoS Pathog.* **11**, e1004484.
- Johnson, D.A. and Cummings, T.F. (2015) Effect of powdery scab root galls on yield of potato. *Plant Dis.* **99**, 1396–1403.
- Jones, J.T., Haegeman, A., Danchin, E.G.J., Gaur, H.S., Helder, J., Jones, M.G., Kikuchi, T., Manzanilla-López, R., Palomares-Rius, J.E., Wesemael, W.M. and Perry, R.N. (2013) Top 10 plant-parasitic nematodes in molecular plant pathology. *Mol. Plant Pathol.* **14**, 946–961.
- Jubault, M., Hamon, C., Gravot, A., Lariagon, C., Delourme, R., Bouchereau, A. and Manzanares-Dauleux, M.J. (2008) Differential regulation of root arginine catabolism and polyamine metabolism in clubroot-susceptible and partially resistant *Arabidopsis* genotypes. *Plant Physiol.* **146**, 2008–2019.
- Jubault, M., Lariagon, C., Taconnat, L., Renou, J.-P., Gravot, A., Delourme, R. and Manzanares-Dauleux, M.J. (2013) Partial resistance to clubroot in *Arabidopsis* is based on changes in the host primary metabolism and targeted cell division and expansion capacity. *Funct. Integr. Genomics*, **13**, 191–205.
- Kamoun, S., Furzer, O., Jones, J.D.G., Judelson, H.S., Ali, G.S., Dalio, R.J., Roy, S.G., Schena, L., Zambounis, A., Panabières, F. and Cahill, D. (2015) The

- Top 10 oomycete pathogens in molecular plant pathology. *Mol. Plant Pathol.* **16**, 413–434.
- Kanyuka, K., Ward, E. and Adams, M.J. (2003) *Polymyxa graminis* and the cereal viruses it transmits: a research challenge. *Mol. Plant Pathol.* **4**, 393–406.
- Kastelein, P. (1987) Investigations on 'Hartrot' of coconut and oilpalms in Suriname. PhD dissertation, Rijksuniversiteit te Utrecht, Netherlands.
- Kawamura, Y., Yokoo, K., Tojo, M. and Hishiike, M. (2005) Distribution of *Pythium porphyrae*, the causal agent of red rot disease of *Porphyra* spp., in the Ariake Sea, Japan. *Plant Dis.* **89**, 1041–1047.
- Kerrigan, J.L., Olsen, M.W. and Martin, S.B. (2012) Rapid blight of turfgrass. *Plant Health Instructor*, <https://www.apsnet.org/edcenter/intropp/lessons/fungi/other/Pages/RapidBlight.aspx> [accessed on Aug 1, 2017].
- Kim, G.H., Moon, K.H., Kim, J.Y., Shim, J. and Klochkova, T.A. (2014) A reevaluation of algal diseases in Korean *Pyropia* (*Porphyra*) sea farms and their economic impact. *Algae*, **29**, 249–265.
- Kitajima, E.W., Vainstein, M.H. and Silveira, J.S.M. (1986) Flagellate protozoan associated with poor development of the root-system of cassava in the Espírito-Santo State, Brazil. *Phytopathology*, **76**, 638–642.
- Klewer, A., Luerben, H., Graf, H. and Siemens, J. (2001) Restriction fragment length polymorphism markers to characterize *Plasmodiophora brassicae* single-spore isolates with different virulence patterns. *J. Phytopathol.* **149**, 121–127.
- Klochkova, T.A., Shim, J.B., Hwang, M.S. and Kim, G.H. (2012) Host–parasite interactions and host species susceptibility of the marine oomycete parasite, *Olpidiopsis* sp., from Korea that infects red algae. *J. Appl. Phycol.* **24**, 135–144.
- Klochkova, T.A., Shin, Y.J., Moon, K.H., Motomura, T. and Kim, G.H. (2016) New species of unicellular obligate parasite, *Olpidiopsis pyropiae* sp. nov., that plagues *Pyropia* sea farms in Korea. *J. Appl. Phycol.* **28**, 73–83.
- Klochkova, T.A., Jung, S. and Kim, G.H. (2017) Host range and salinity tolerance of *Pythium porphyrae* may indicate its terrestrial origin. *J. Appl. Phycol.* **29**, 371–379.
- Kombrink, A. and Thomma, B.P.H.J. (2013) LysM effectors: secreted proteins supporting fungal life. *PLoS Pathog.* **9**, e1003769.
- Kořený, L., Sobotka, R.J.K., Gnipová, A., Flegontov, P., Horváth, A., Oborník, M., Ayala, F.J. and Lukes, J. (2012) Aerobic kinetoplastid flagellate *Phytomonas* does not require heme for viability. *Proc. Natl. Acad. Sci. USA*, **109**, 3808–3813.
- La Barre, S., Potin, P., Leblanc, C. and Delage, L. (2010) The halogenated metabolism of brown algae (*Phaeophyta*), its biological importance and its environmental significance. *Mar. Drugs*, **8**, 988–1010.
- Lafont, A. (1909) Sur la présence d'un *Leptomonas*, parasite de la classe des Flagelles dans le latex de l'*Euphorbia pilulifera*. *C.r. séances Soc. biol. ses. fil.* **66**, 1011–1013.
- Lamb, J.B., van de Water, J.A.J.M., Bourne, D.G., Altier, C., Hein, M.Y., Fiorenza, E.A., Abu, N., Jompa, J. and Harvell, C.D. (2017) Seagrass ecosystems reduce exposure to bacterial pathogens of humans, fishes, and invertebrates. *Science*, **355**, 731–733.
- Legreve, A., Vanpee, B., Delfosse, P. and Maraite, H. (2000) Host range of tropical and sub-tropical isolates of *Polymyxa graminis*. *Eur. J. Plant Pathol.* **106**, 379–389.
- Legreve, A., Delfosse, P. and Maraite, H. (2002) Phylogenetic analysis of *Polymyxa* species based on nuclear 5.8S and internal transcribed spacers ribosomal DNA sequences. *Mycol. Res.* **106**, 138–147.
- Lemarié, S., Robert-Seilaniantz, A., Lariagon, C., Lemoine, J., Marnet, N., Jubault, M., Manzaneres-Dauleux, M.J. and Gravot, A. (2015) Both the jasmonic acid and the salicylic acid pathways contribute to resistance to the biotrophic clubroot agent *Plasmodiophora brassicae* in *Arabidopsis*. *Plant Cell Physiol.* **56**, 2158–2168.
- Lindholm, T., Lindqvist, C. and Sjöqvist, C. (2016) Occurrence and activity of slime nets, *Labyrinthula* sp. among aquatic plants in cold and oligohaline Baltic Sea waters. *Ann. Bot. Fennici*, **53**, 139–143.
- Lopez, G., Genty, P. and Ollagnier, M. (1975) Control preventivo de la "Marchitez sorpresiva" del *Elaeis guineensis* en America Latina. *Oleagineux* **30**, 243–250.
- Loureiro, R., Gachon, C.M. and Rebours, C. (2015) Seaweed cultivation: potential and challenges of crop domestication at an unprecedented pace. *New Phytol.* **206**, 489–492.
- Lovelock, D.A., Donald, C.E., Conlan, X.A. and Cahill, D.M. (2013) Salicylic acid suppression of clubroot in broccoli (*Brassica oleracea* var. *italica*) caused by the obligate biotroph *Plasmodiophora brassicae*. *Australas. Plant Pathol.* **42**, 141–153.
- Ludwig-Müller, J. (2016) Belowground defence strategies against clubroot (*Plasmodiophora brassicae*). In: *Belowground Defence Strategies in Plants* (Vos, C. M. F. and Kazan, K., eds.), pp. 195–219. Cham: Springer International Publishing.
- Ludwig-Müller, J., Prinsen, E., Rolfe, S.A. and Scholes, J.D. (2009) Metabolism and plant hormone action during clubroot disease. *J. Plant Growth Regul.* **28**, 229–244.
- Ludwig-Müller, J., Jülke, S., Geiß, K., Richter, F., Mithöfer, A., Sola, I., Rusak, G., Keenan, S. and Bulman, S. (2015) A novel methyltransferase from the intracellular pathogen *Plasmodiophora brassicae* methylates salicylic acid. *Mol. Plant Pathol.* **16**, 349–364.
- Ludwig-Müller, J., Auer, S., Jülke, S. and Marschollek, S. (2017) Manipulation of auxin and cytokinin balance during the *Plasmodiophora brassicae*–*Arabidopsis thaliana* interaction. In: *Auxins and Cytokinins in Plant Biology: Methods and Protocols* (Dandekar, T. and Naseem, M., eds.), pp. 41–60. New York, NY: Springer.
- Lukes, J., Skalicky, T., Tyc, J., Votypka, J. and Yurchenko, V. (2014) Evolution of parasitism in kinetoplastid flagellates. *Mol. Biochem. Parasitol.* **195**, 115–122.
- Malinowski, R., Smith, J.A., Fleming, A.J., Scholes, J.D. and Rolfe, S.A. (2012) Gall formation in clubroot-infected *Arabidopsis* results from an increase in existing meristematic activities of the host but is not essential for the completion of the pathogen life cycle. *Plant J.* **71**, 226–238.
- Malinowski, R., Novák, O., Borhan, M.H., Spíchal, L., Strnad, M. and Rolfe, S.A. (2016) The role of cytokinins in clubroot disease. *Eur. J. Plant Pathol.* **145**, 543–557.
- Mansfield, J., Genin, S., Magori, S., Citovsky, V., Sriariyanum, M., Ronald, P., Dow, M.A.X., Verdier, V., Beer, S.V., Machado, M.A. and Toth, I.A.N. (2012) Top 10 plant pathogenic bacteria in molecular plant pathology. *Mol. Plant Pathol.* **13**, 614–629.
- Martin, D.L., Chiari, Y., Boone, E., Sherman, T.D., Ross, C., Wyllie-Echeverria, S., Gaydos, J.K. and Boettcher, A.A. (2016) Functional, phylogenetic and host-geographic signatures of *Labyrinthula* spp. provide for putative species delimitation and a global-scale view of seagrass wasting disease. *Estuar. Coasts*, **39**, 1–19.
- Maslov, D.A., Votypka, J., Yurchenko, V. and Lukes, J. (2013) Diversity and phylogeny of insect trypanosomatids: all that is hidden shall be revealed. *Trends Parasitol.* **29**, 43–52.
- McGhee, R.B. and McGhee, A.H. (1979) Biology and structure of *Phytomonas staheli* sp.n. a trypanosomatid located in sieve tubes of coconut and oil palms. *J. Protozool.* **26**, 348–351.
- McGrann, G.R.D., Grimmer, M.K., Mutasa-Goettgens, E.S. and Stevens, M. (2009) Progress towards the understanding and control of sugar beet rhizomania disease. *Mol. Plant Pathol.* **10**, 129–141.
- McKone, K.L. and Tanner, C.E. (2009) Role of salinity in the susceptibility of eelgrass *Zostera marina* to the wasting disease pathogen *Labyrinthula zosterae*. *Mar. Ecol. Prog. Ser.* **377**, 123–130.
- Medina, J.M., Rodrigues, J.C.F., Moreira, O.C., Atella, G., de Souza, W. and Barrabin, H. (2015) Mechanisms of growth inhibition of *Phytomonas serpens* by the alkaloids tomatine and tomatidine. *Mem. Inst. Oswaldo Cruz*, **110**, 48–55.
- Merz, U. and Falloon, R.E. (2009) Review: powdery scab of potato—increased knowledge of pathogen biology and disease epidemiology for effective disease management. *Potato Res.* **52**, 17–37.
- Michel, G., Tonon, T., Scornet, D., Cock, J.M. and Kloareg, B. (2010) The cell wall polysaccharide metabolism of the brown alga *Ectocarpus siliculosus*. Insights into the evolution of extracellular matrix polysaccharides in Eukaryotes. *New Phytol.* **188**, 82–97.
- Moxham, S.E. and Buczacki, S.T. (1983) Chemical-composition of the resting spore wall of *Plasmodiophora-brassicae*. *Trans. Br. Mycol. Soc.* **80**, 297–304.
- Muehlstein, L.K. (1989) Perspectives on the wasting disease of eelgrass *Zostera marina*. *Dis. Aquat. Organ.* **7**, 211–221.
- Muehlstein, L.K. (1992) The host–pathogen interaction in the wasting disease of eelgrass, *Zostera marina*. *Can. J. Bot.* **70**, 2081–2088.
- Muehlstein, L.K., Porter, D. and Short, F.T. (1988) *Labyrinthula* sp., a marine slime-mold producing the symptoms of wasting disease in eelgrass, *Zostera marina*. *Mar. Biol.* **99**, 465–472.
- Müller, D.G., Küpper, F.C. and Küpper, H. (1999) Infection experiments reveal broad host ranges of *Eurychasma dicksonii* (Oomycota) and *Chytridium polysiphoniae* (Chytridiomycota), two eukaryotic parasites in marine brown algae (*Phaeophyceae*). *Phycol. Res.* **47**, 217–223.
- Nakamura, Y., Sasaki, N., Kobayashi, M., Ojima, N., Yasuie, M., Shigenobu, Y., Satomi, M., Fukuma, Y., Shiwaku, K., Tsujimoto, A. and Kobayashi, T. (2013) The first symbiont-free genome sequence of marine red alga, *Susabi-nori* (*Pyropia yezoensis*). *PLoS One*, **8**, e57122.
- Neuhauser, S., Bulman, S. and Kirchmair, M. (2010) Plasmodiophorids: The Challenge to Understand Soil-Borne, Obligate Biotrophs with a Multiphasic Life Cycle.

- In: *Molecular Identification of Fungi* (Gherbawy, Y. and Voigt, K., eds.), pp. 51–78. Berlin, Heidelberg: Springer Berlin Heidelberg.
- Neuhauser, S., Kirchmair, M. and Gleason, F.H. (2011) The ecological potentials of *Phytomyxa* ("plasmodiophorids") in aquatic food webs. *Hydrobiologia*, **659**, 23–35.
- Neuhauser, S., Kirchmair, M., Bulman, S. and Bass, D. (2014) Cross-kingdom host shifts of phytomyxid parasites. *BMC Evol. Biol.* **14**, 33.
- O'Brien, P.A. and Milroy, S.P. (2017) Towards biological control of *Spongospora subterranea* f. sp. *subterranea*, the causal agent of powdery scab in potato. *Australas. Plant Pathol.* **46**, 1–10.
- Olsen, J.L., Rouze, P., Verhelst, B., Lin, Y.C., Bayer, T., Collen, J., Dattolo, E., De Paoli, E., Dittami, S., Maumus, F. and Michel, G. (2016) The genome of the seagrass *Zostera marina* reveals angiosperm adaptation to the sea. *Nature*, **530**, 331–335.
- Olsen, M.W. (2007) *Labyrinthula terrestris*: a new pathogen of cool-season turf-grasses. *Mol. Plant Pathol.* **8**, 817–820.
- Olsen, Y.S. and Duarte, C.M. (2015) Combined effect of warming and infection by *Labyrinthula* sp. on the Mediterranean seagrass *Cymodocea nodosa*. *Mar. Ecol. Prog. Ser.* **532**, 101–109.
- Pan, J.W., del Campo, J. and Keeling, P.J. (2017) Reference tree and environmental sequence diversity of Labyrinthulomycetes. *J. Eukaryot. Microbiol.* **64**, 88–96.
- Park, C.S. and Hwang, E.K. (2015) Biochemical characterization of *Pyropia yezoensis*-AP1 strain accompanies the resistance reaction to the red rot disease pathogen, *Pythium porphyrae*. *J. Appl. Phycol.* **27**, 2149–2156.
- Park, C.S., Kakinuma, M. and Amano, H. (2001) Detection and quantitative analysis of zoospores of *Pythium porphyrae*, causative organism of red rot disease in *Porphyra*, by competitive PCR. *J. Appl. Phycol.* **13**, 433–441.
- Park, C.S., Kakinuma, M. and Amano, H. (2006) Forecasting infections of the red rot disease on *Porphyra yezoensis* Ueda (*Rhodophyta*) cultivation farms. *J. Appl. Phycol.* **18**, 295–299.
- Parthasarathy, M.V., Van Slobbe, W.G. and Soudant, C. (1976) Trypanosomatid flagellate in the phloem of diseased coconut palms. *Science*, **192**, 1346–1348.
- Päsold, S., Siegel, I., Seidel, C. and Ludwig-Müller, J. (2010) Flavonoid accumulation in *Arabidopsis thaliana* root galls caused by the obligate biotrophic pathogen *Plasmodiophora brassicae*. *Mol. Plant Pathol.* **11**, 545–562.
- Petersen, H.E. (1905) Contributions à la connaissance des Phycomycetes marines. *Overs. K. Danske Vidensk. Selsk. Forh.* **5**, 439–488.
- Porcel, B.M., Deneoud, F., Opperdoes, F., Noel, B., Madoui, M.A., Hammarton, T.C., Field, M.C., Da Silva, C., Couloux, A., Poulain, J. and Katinka, M. (2014) The streamlined genome of *Phytomonas* spp. relative to human pathogenic kinetoplastids reveals a parasite tailored for plants. *PLoS Genet.* **10**, e1004007.
- Preston, T.M. and King, C.A. (2005) Actin-based motility in the net slime mould *Labyrinthula*: evidence for the role of myosin in gliding movement. *J. Eukaryot. Microbiol.* **52**, 461–475.
- Qu, X.S. and Christ, B.J. (2004) Genetic variation and phylogeny of *Spongospora subterranea* f.sp. *subterranea* based on ribosomal DNA sequence analysis. *Am. J. Potato Res.* **81**, 385–394.
- Ritter, A., Goultiquet, S., Salaun, J.P., Tonon, T., Correa, J.A. and Potin, P. (2008) Copper stress induces biosynthesis of octadecanoid and eicosanoid oxygenated derivatives in the brown algal kelp *Laminaria digitata*. *New Phytol.* **180**, 809–821.
- Rodgers, K.L. and Shears, N.T. (2016) Modelling kelp forest primary production using in situ photosynthesis, biomass and light measurements. *Mar. Ecol. Prog. Ser.* **553**, 67–79.
- Rolfe, S.A., Strelkov, S.E., Links, M.G., Clarke, W.E., Robinson, S.J., Djavaheri, M., Malinowski, R., Haddadi, P., Kagale, S., Parkin, I.A. and Taheri, A. (2016) The compact genome of the plant pathogen *Plasmodiophora brassicae* is adapted to intracellular interactions with host *Brassica* spp. *BMC Genomics*, **17**, 1–15.
- Scholthof, K.B.G., Adkins, S., Czosnek, H., Palukaitis, P., Jacquot, E., Hohn, T., Hohn, B., Saunders, K., Candresse, T., Ahlquist, P. and Hemenway, C. (2011) Top 10 plant viruses in molecular plant pathology. *Mol. Plant Pathol.* **12**, 938–954.
- Schuller, A., Kehr, J. and Ludwig-Müller, J. (2014) Laser microdissection coupled to transcriptional profiling of *Arabidopsis* roots inoculated by *Plasmodiophora brassicae* indicates a role for brassinosteroids in clubroot formation. *Plant Cell Physiol.* **55**, 392–411.
- Schwelm, A., Fogelqvist, J., Knaust, A., Jülke, S., Lilja, T., Bonilla-Rosso, G., Karlsson, M., Shevchenko, A., Dhandapani, V., Choi, S.R. and Kim, H.G. (2015) The *Plasmodiophora brassicae* genome reveals insights in its life cycle and ancestry of chitin synthases. *Sci. Rep.* **5**, 11 153.
- Schwelm, A., Berney, C., Bass, D., Dixelius, C. and Neuhauser, S. (2016) The large subunit rDNA sequence of *Plasmodiophora brassicae* does not contain intraspecific polymorphism. *Protist*, **167**, 544–554.
- Sekimoto, S., Beakes, G.W., Gachon, C.M.M., Muller, D.G., Kupper, F.C. and Honda, D. (2008) The development, ultrastructural cytology, and molecular phylogeny of the basal oomycete *Eurychasma dicksonii*, infecting the filamentous phaeophyte algae *Ectocarpus siliculosus* and *Pylaiella littoralis*. *Protist*, **159**, 299–318.
- Seward, E.A., Votycka, J., Kment, P., Lukeš, J. and Kelly, S. (2016) Description of *Phytomonas oxycareni* n. sp. from the salivary glands of *Oxycareus lavaterae*. *Protist*, **168**, 71–79.
- Shah, F.A., Falloon, R.E., Butler, R.C. and Lister, R.A. (2012) Low amounts of *Spongospora subterranea* sporosorus inoculum cause severe powdery scab, root galling and reduced water use in potato (*Solanum tuberosum*). *Australas. Plant Pathol.* **41**, 219–228.
- Siemens, J., Keller, I., Sarx, J., Kunz, S., Schuller, A., Nagel, W., Schülling, T., Parniske, M. and Ludwig-Müller, J. (2006) Transcriptome analysis of *Arabidopsis* clubroots indicate a key role for cytokinins in disease development. *Mol. Plant-Microbe Interact.* **19**, 480–494.
- Siemens, J., Bulman, S., Rehn, F. and Sundelin, T. (2009) Molecular biology of *Plasmodiophora brassicae*. *J. Plant Growth Regul.* **28**, 245–251.
- Smith, M.J., Adams, M.J. and Ward, E. (2011) Evidence that *Polymyxa* species may infect *Arabidopsis thaliana*. *FEMS Microbiol. Lett.* **318**, 35–40.
- Smith, M.J., Adams, M.J. and Ward, E. (2013) Ribosomal DNA analyses reveal greater sequence variation in *Polymyxa* species than previously thought and indicate the possibility of new ribotype-host-virus associations. *Environ. Microbiol. Rep.* **5**, 143–150.
- Song, T., Chu, M., Lahlali, R., Yu, F. and Peng, G. (2016) Shotgun label-free proteomic analysis of clubroot (*Plasmodiophora brassicae*) resistance conferred by the gene Rcr1 in *Brassica rapa*. *Front. Plant Sci.* **7**, 1013.
- Stahel, G. (1931) Zur Kenntnis der Siebröhrenkrankheit (Phloemnekrose) des Kaffeebaumes in Surinam. II. *Phytopathol. Z.* **4**, 539–548.
- Stengel, D.B. and Connan, S. (2015) Marine algae: a source of biomass for biotechnological applications. In: *Natural Products from Marine Algae: Methods and Protocols* (Stengel, D. B. and Connan, S., eds.), pp. 1–37. New York, NY: Springer.
- Stowell, L.J., Martin, S.B., Olsen, M.W., Bigelow, D., Kohout, M., Peterson, P.D., Camberato, J. and Gelernter, W.D. (2005) Rapid blight: a new plant disease. *APSnet Features* July 2005. Available at: <http://www.apsnet.org/publications/apsnetfeatures/Pages/RapidBlight.aspx>. [accessed on Aug 1, 2017].
- Strelkov, S.E., Hwang, S.F., Manolii, V.P., Cao, T. and Feindel, D. (2016) Emergence of new virulence phenotypes of *Plasmodiophora brassicae* on canola (*Brassica napus*) in Alberta, Canada. *Eur. J. Plant Pathol.* **145**, 517–529.
- Strittmatter, M., Grenville-Briggs, L.J., Breithut, L., van West, P., Gachon, C.M.M. and Kupper, F.C. (2016) Infection of the brown alga *Ectocarpus siliculosus* by the oomycete *Eurychasma dicksonii* induces oxidative stress and halogen metabolism. *Plant Cell Environ.* **39**, 259–271.
- Sullivan, B.K., Sherman, T.D., Damare, V.S., Lilje, O. and Gleason, F.H. (2013) Potential roles of *Labyrinthula* spp. in global seagrass population declines. *Fungal Ecol.* **6**, 328–338.
- Sullivan, B.K., Robinson, K.L., Trevathan-Tackett, S.M., Lilje, E.S., Gleason, F.H. and Lilje, O. (2016) The first isolation and characterisation of the protist *Labyrinthula* sp. in Southeastern Australia. *J. Eukaryot. Microbiol.* **64**, 504–513.
- Tamada, T. and Kondo, H. (2013) Biological and genetic diversity of plasmodiophorid-transmitted viruses and their vectors. *J. Gen. Plant Pathol.* **79**, 307–320.
- Trevathan-Tackett, S.M., Lane, A.L., Bishop, N. and Ross, C. (2015) Metabolites derived from the tropical seagrass *Thalassia testudinum* are bioactive against pathogenic *Labyrinthula* sp. *Aquat. Bot.* **122**, 1–8.
- Tsirigoti, A., Beakes, G.W., Herve, C., Gachon, C.M. and Katsaros, C. (2015) Attachment, penetration and early host defense mechanisms during the infection of filamentous brown algae by *Eurychasma dicksonii*. *Protoplasma*, **252**, 845–856.
- Tsui, C.K., Marshall, W., Yokoyama, R., Honda, D., Lippmeier, J.C., Craven, K.D., Peterson, P.D. and Berbee, M.L. (2009) Labyrinthulomycetes phylogeny and its implications for the evolutionary loss of chloroplasts and gain of ectoplasmic gliding. *Mol. Phylogenet. Evol.* **50**, 129–140.
- Uppalapati, S.R. and Fujita, Y. (2000) Carbohydrate regulation of attachment, encystment, and appressorium formation by *Pythium porphyrae* (*Oomycota*) zoospores on *Porphyra yezoensis* (*Rhodophyta*). *J. Phycol.* **36**, 359–366.
- Uppalapati, S.R. and Fujita, Y. (2001) The relative resistances of *Porphyra* species (*Bangiales*, *Rhodophyta*) to infection by *Pythium porphyrae* (*Peronosporales*, *Oomycota*). *Bot. Mar.* **44**, 1–7.

- Vaianopoulos, C., Bragard, C., Moreau, V., Maraite, H. and Legreve, A. (2007) Identification and quantification of *Polymyxa graminis* f. sp. *temperata* and *P. graminis* f. sp. *tepida* on barley and wheat. *Plant Dis.* **91**, 857–864.
- Vergeer, L.H.T., Aarts, T.L. and Degroot, J.D. (1995) The wasting disease and the effect of abiotic factors (light-intensity, temperature, salinity) and infection with *Labyrinthula zosterae* on the phenolic content of *Zostera marina* shoots. *Aquat. Bot.* **52**, 35–44.
- Vishniac, H.S. (1955) The nutritional requirements of isolates of *Labyrinthula* spp. *J. Gen. Microbiol.* **12**, 455–463.
- Walsh, J.A., Clay, C.M. and Miller, A. (1989) A new virus disease of watercress in England. *EPPO Bull.* **19**, 463–470.
- Ward, E., Kanyuka, K., Motteram, J., Korniyukhin, D. and Adams, M.J. (2005) The use of conventional and quantitative real-time PCR assays for *Polymyxa graminis* to examine host plant resistance, inoculum levels and intraspecific variation. *New Phytol.* **165**, 875–885.
- Ward, L.I., Fenn, M.G.E. and Henry, C.M. (2004) A rapid method for direct detection of *Polymyxa* DNA in soil. *Plant Pathol.* **53**, 485–490.
- Woronin, M. (1877) *Plasmodiophora brassicae*, der Organismus, der die unter dem Namen Hernie bekannte Krankheit der Kohlpflanzen verursacht. *Arb. St. Petersburg naturf. Ges.* **8**, 169–201.
- Wu, S., Sun, J., Chi, S., Wang, L., Wang, X., Liu, C., Li, X., Yin, J., Liu, T. and Yu, J. (2014) Transcriptome sequencing of essential marine brown and red algal species in China and its significance in algal biology and phylogeny. *Acta Oceanol. Sin.* **33**, 1–12.
- Young, E.L. (1943) Studies on *Labyrinthula*. The etiologic agent of the wasting disease of eel-grass. *Am. J. Bot.* **30**, 586–593.
- Zamani-Noor, N. (2017) Variation in pathotypes and virulence of *Plasmodiophora brassicae* populations in Germany. *Plant Pathol.* **66**, 316–324.
- Zambounis, A., Elias, M., Sterck, L., Maumus, F., Gachon, C.M. (2012) Highly dynamic exon shuffling in candidate pathogen receptors ... what if brown algae were capable of adaptive immunity? *Mol. Biol. Evol.* **29**, 1263–1276.
- Zhang, D.P., Burroughs, A.M., Vidal, N.D., Iyer, L.M. and Aravind, L. (2016a) Transposons to toxins: the provenance, architecture and diversification of a widespread class of eukaryotic effectors. *Nucleic Acids Res.* **44**, 3513–3533.
- Zhang, X., Liu, Y., Fang, Z., Li, Z., Yang, L., Zhuang, M., Zhang, Y. and Lv, H. (2016b) Comparative transcriptome analysis between Broccoli (*Brassica oleracea* var. *italica*) and wild cabbage (*Brassica macrocarpa* Guss.) in response to *Plasmodiophora brassicae* during different infection stages. *Front. Plant Sci.* **7**, 1929.
- Ziegler, A., Fomitcheva, V., Zakri, A.M. and Kastir, U. (2016) Occurrence of *Polymyxa graminis* ribotypes in Germany and their association with different host plants and viruses. *Cereal Res. Commun.* **44**, 251–262.

Farming, slaving and enslavement: histories of endosymbioses during kinetoplastid evolution

Jane Harmer¹, Vyacheslav Yurchenko^{2,3,4}, Anna Nenarokova^{2,5}, Julius Lukeš^{2,5} and Michael L. Ginger¹

Special Issue Review

Cite this article: Harmer J, Yurchenko V, Nenarokova A, Lukeš J, Ginger ML (2018). Farming, slaving and enslavement: histories of endosymbioses during kinetoplastid evolution. *Parasitology* **145**, 1311–1323. <https://doi.org/10.1017/S0031182018000781>

Received: 26 January 2018

Revised: 29 March 2018

Accepted: 8 April 2018

First published online: 13 June 2018

Key words:

Angomonas deanei; *Candidatus* Kinetoplastibacterium; cytostome; *Kentomonas*; *Novymonas esmeraldas*; *Pandoraea*

Authors for correspondence: Jane Harmer and Michael L. Ginger, E-mail: J.Harmer@hud.ac.uk and M.Ginger@hud.ac.uk

¹Department of Biological Sciences, School of Applied Sciences, University of Huddersfield, Huddersfield, HD1 3DH, UK; ²Biology Centre, Institute of Parasitology, Czech Academy of Sciences, 370 05 České Budějovice (Budweis), Czechia; ³Faculty of Science, Life Science Research Centre, University of Ostrava, 710 00 Ostrava, Czechia; ⁴Martsinovsky Institute of Medical Parasitology, Tropical and Vector Borne Diseases, Sechenov University, Moscow, Russia and ⁵Faculty of Sciences, University of South Bohemia, České Budějovice (Budweis), Czechia

Abstract

Parasitic trypanosomatids diverged from free-living kinetoplastid ancestors several hundred million years ago. These parasites are relatively well known, due in part to several unusual cell biological and molecular traits and in part to the significance of a few – pathogenic *Leishmania* and *Trypanosoma* species – as aetiological agents of serious neglected tropical diseases. However, the majority of trypanosomatid biodiversity is represented by osmotrophic monoxenous parasites of insects. In two lineages, novymonads and strigomonads, osmotrophic lifestyles are supported by cytoplasmic endosymbionts, providing hosts with macromolecular precursors and vitamins. Here we discuss the two independent origins of endosymbiosis within trypanosomatids and subsequently different evolutionary trajectories that see entrainment vs tolerance of symbiont cell divisions cycles within those of the host. With the potential to inform on the transition to obligate parasitism in the trypanosomatids, interest in the biology and ecology of free-living, phagotrophic kinetoplastids is beginning to enjoy a renaissance. Thus, we take the opportunity to additionally consider the wider relevance of endosymbiosis during kinetoplastid evolution, including the indulged lifestyle and reductive evolution of basal kinetoplastid *Perkinsella*.

Introduction

Kinetoplastids are one of three major groups of organisms that belong to the evolutionarily divergent protist phylum Euglenozoa (Cavalier-Smith, 2016). Although divergent, euglenozoans are ubiquitous; representatives from all three groups are easily isolated from many freshwater, marine and soil environments and, in part due to their overall abundance, contribute significantly to ecosystem ecology (von der Heyden *et al.*, 2004; Edgcomb *et al.*, 2011; Lukeš *et al.*, 2015; Mukherjee *et al.*, 2015; Flegontova *et al.*, 2016; Flegontova *et al.*, 2018).

Systematically, the kinetoplastids separate into the monophyletic, obligatory parasitic trypanosomatids and a wide diversity of free-living, bi-flagellate phagotrophs, with occasional examples of parasites and symbionts populating three major clades (von der Heyden *et al.*, 2004; Simpson *et al.*, 2006; Kaufer *et al.*, 2017; Yazaki *et al.*, 2017). It is the uniflagellate trypanosomatids that are the best known due to the role of some as the aetiological agents of serious, neglected tropical diseases (Nussbaum *et al.*, 2010). The defining characteristic common to both free-living and parasitic kinetoplastids is the coalescence (in trypanosomatids the catenation) of several thousand circular DNA molecules to form distinctive mitochondrial genome architectures, known more commonly as kinetoplasts, and which give rise to the class name Kinetoplastea (Lukeš *et al.*, 2002). Uridine-insertion and -deletion editing of mRNA on a massive scale is essential for gene expression from these genomes and provides a second example of extreme or unusual biology that defines and pervades throughout the kinetoplastids (Aphasizhev and Aphasizheva, 2014; David *et al.*, 2015; Read *et al.*, 2016). For further examples of extreme kinetoplastid biology that have peripheral relevance for this review – peroxisome-compartmentalized carbohydrate metabolism, loss of transcriptional control on protein-coding gene expression, flagellar pocket dynamics – readers are directed towards articles by Haanstra *et al.* (2016), Morales *et al.* (2016a), Clayton (2014), and Field and Carrington (2009).

Although trypanosomatid species are widely known as the causative agents for diseases of medical, veterinary and agricultural importance (Jaskowska *et al.*, 2015; Giordani *et al.*, 2016; Field *et al.*, 2017; Kaufer *et al.*, 2017), most members of the family are simply monoxenous parasites of insects (Podlipaev *et al.*, 2004; Maslov *et al.*, 2013; Lukeš *et al.*, 2014; Kaufer *et al.*, 2017) with not always a clear indication that these protists are pathogenic towards their invertebrate host(s). Also less widely recognized is that at least twice, the symbiosis between a bacterial endosymbiont and a host trypanosomatid has occurred (Du *et al.*, 1994; de Souza and Motta, 1999; Votýpka *et al.*, 2014; Kostygov *et al.*, 2016) (Fig. 1). Trypanosomatid taxa involved in these events are not particularly closely related, and different evolutionary trajectories are possibly evident for each symbiosis: in the Strigomonadinae, growth and division of a single bacterial endosymbiont is entrained within the cell cycle of

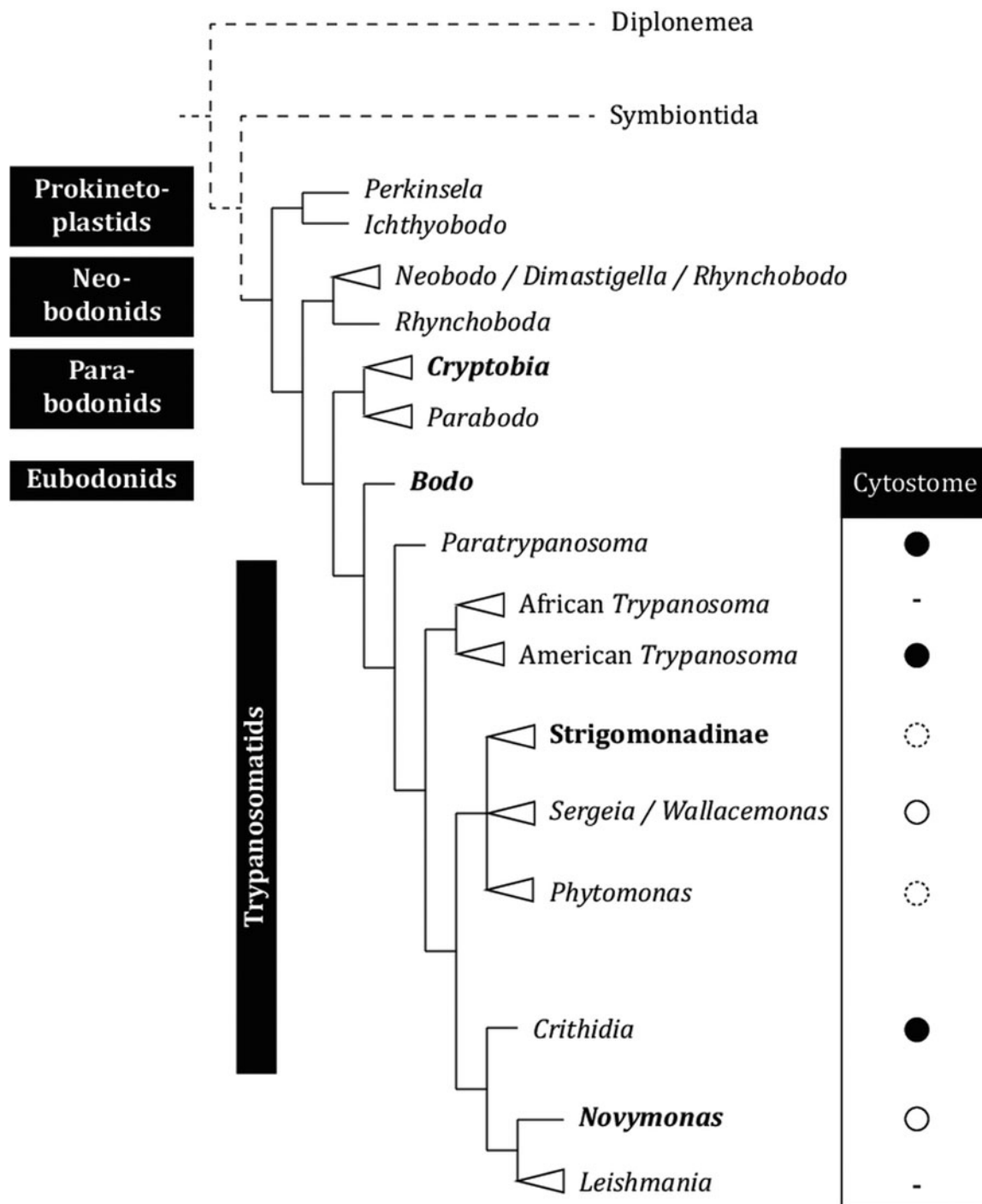


Fig. 1. Kinetoplastid phylogeny and a history of endosymbiosis. Taxa in possession of bacterial endosymbionts are highlighted in bold. Filled circles denote presence of a cytotome–cytopharynx complex in some trypanosomatid taxa; open and dashed circles denote uncertainty (as defined by an absence of data) or an unlikelihood (based on extensive, published electron microscopy studies), respectively, with regard to the presence of these structures in others; – denotes absence of a cytotome–cytopharynx from *Leishmania* and African trypanosome species.

the host cell (Motta *et al.*, 2010), whereas in recently discovered *Novymonas* less stringent regulation on the number of β -proteobacterial *Pandoraea* endosymbionts could reflect either symbiont farming or a snap-shot of an early transitional phase in the establishment of a novel endosymbiont–host relationship (Kostygov *et al.*, 2016, 2017). In this mini-review, we consider metabolic advantages conferred by bacterial endosymbionts to their partner trypanosomatids, how the biology of the host cell potentially influences the establishment, reductive evolution and subsequent entrainment of the endosymbiont(s), and we survey the literature with regard to endosymbioses within free-living phagotrophic kinetoplastids. Finally, we also consider the

fascinating example of *Perkinsela*, a basal kinetoplastid and itself an endosymbiont of *Paramoeba* sp. (Dyková *et al.*, 2003; Tanifuji *et al.*, 2011). Here, the evolutionary path from protist to obligate endosymbiont has been accompanied by streamlining and loss of much cell biology that defines and characterizes the Kinetoplastea (Tanifuji *et al.*, 2017).

Independent origins of endosymbiosis among trypanosomatids

Species belonging to four trypanosomatid genera, spanning two monophyletic groups (Fig. 1), are characterized by the presence

of a bacterial endosymbiont. Phylogenetic analyses indicate *Novymonas esmeraldas*, the most recently characterized endosymbiont-bearing trypanosomatid isolated in Ecuador from a scentless plant bug (*Niesthrea vincentii*) (Kostygov *et al.*, 2016), is closely related to the genus *Leishmania* that encompasses more than 30 species of dioxenous parasites found variously in tropical and sub-tropical countries across the New and Old World and include the aetiological agents of cutaneous, mucocutaneous and visceral human disease (Akhoundi *et al.*, 2017). *Novymonas esmeraldas* is considered to be monoxenous and non-pathogenic, despite its close relatedness with *Leishmania*. It contains a β -proteobacterial endosymbiont, *Candidatus* *Pandoraea novymonadis* (order Burkholderiales; family Burkholderiaceae) (Kostygov *et al.*, 2017). Environmental DNA reads corresponding to 18S rRNA and trypanosomatid spliced leader RNA gene sequences point to the presence of trypanosomatid taxa very closely related to *N. esmeraldas* in Central Africa (Kostygov *et al.*, 2016), raising the question of whether such taxa also contain similar *Pandoraea*-related endosymbionts.

In contrast, strigomonads form a discrete monophyletic clade most closely related to the genera *Wallacemonas* and *Sergeia* and some distance removed from the leishmanias (Teixeira *et al.*, 2011; Votýpka *et al.*, 2014). In the common ancestor of the three genera-forming Strigomonadinae – *Angomonas*, *Strigomonas* and *Kentomonas* – an endosymbiotic association with a different member of the Burkholderiales (family Alcaligenaceae) occurred. Considered to be a more ancient relationship than the endosymbiosis occurring in *N. esmeraldas*, the cell cycle of the endosymbiont in strigomonads is firmly entrained within that of its host cell. The peptidoglycan and outermost layers of the endosymbiont cell envelope are absent or heavily reduced (Motta *et al.*, 1997; de Souza and Motta, 1999), potentially facilitating easy metabolite transfer between host and endosymbiont (discussed further in the ‘Interface with host cell biology’). Bacterial endosymbionts in Strigomonadinae are known as *Candidatus* Kinetoplastibacterium spp. (Alves *et al.*, 2013). The known distribution of strigomonads is also more cosmopolitan than that of *Novymonas*, as they are variously found in heteropteran and dipteran insects. Moreover, different *Angomonas* species have been isolated from Europe, the Americas, Africa and Australia, while *Kentomonas* has been isolated from Ecuador and the Philippines, and *Strigomonas* was encountered in different regions of the Americas (Maslov *et al.*, 2013; Votýpka *et al.*, 2014).

Classic hallmarks of the transition to an obligate endosymbiotic life cycle are evident in all trypanosomatid endosymbionts: a reduced GC-content (in comparison with free-living relations), reduced genome size, a paucity of mobile elements, and a reduced gene content (Table 1). Among different *Candidatus* Kinetoplastibacterium spp. from the strigomonads there are only slight variations in overall gene content and near-complete preservation of synteny (Alves *et al.*, 2013; Silva *et al.*, 2018) indicating reductive evolution of these endosymbionts had progressed nearly to completion prior

to the divergence of the last common *Strigomonas/Angomonas/Kentomonas* ancestor(s) or that reductive evolution of endosymbionts followed parallel trajectories in each strigomonad lineage (Alves *et al.*, 2013). We pick up briefly in the discussion of the ‘Interface with host cell biology’ how the reductive evolution of trypanosomatid endosymbiont gene content commonly incorporates loss of processes associated with a free-living lifestyle and perception of environmental change.

Common metabolic gains in endosymbiont-containing trypanosomatids

A consequence of any endosymbiosis is conferment of new metabolic capability for the host cell. Taken to extremes, an endosymbiont’s cell cycle can become entrained within that of its host and the advent of translocon-mediated protein targeting from host to endosymbiont classically marks the transition from endosymbiont to ‘organelle’ (Cavalier-Smith and Lee, 1985; Theissen and Martin, 2006; Keeling *et al.*, 2015; McCutcheon, 2016). Among eukaryotes, the most easily recognizable products of endosymbiotic relationships are mitochondria, which conferred cytochrome-dependent oxidative phosphorylation upon an archaeal host cell of ill-defined metabolic capability (Sousa *et al.*, 2016; Eme *et al.*, 2017; Zachar and Szathmáry, 2017), and chloroplasts responsible for photosynthesis; they have evolved independently twice as the consequence of a primary endosymbiotic event (Nowack and Grossman, 2012; Singer *et al.*, 2017). These organelles were pivotal in the radiation of eukaryotic diversity with chloroplasts, notably of red algal origin, also becoming widely established in many protist lineages as consequences of secondary and tertiary endosymbiosis (Keeling, 2013). On a global scale, chloroplast functions remain integral to carbon cycle dynamics (Pan *et al.*, 2011; Phillips and Lewis, 2014; Worden *et al.*, 2015). At a species level, a few taxa are also secondarily photosynthetic owing to transient retention of chloroplasts (and transcriptionally active nuclei) from their algal prey (Dorrell and Howe, 2012). This phenomenon is termed ‘kleptoplastidy’; such opportunistic oxygenic photosynthesis potentially confers several advantages, including aerobic respiration within anoxic environments (Esteban *et al.*, 2009). A wide variety of other endosymbioses also exist in eukaryotic evolution that confer alternative physiological advantage(s) for the host cell as consequences of different metabolic gains, e.g. N₂ fixation and N₂ recycling (from waste host urea, ammonium products) occurring in termite gut-dwelling parabasalid and oxymonad flagellates and in some diatoms or, among anaerobic (non-photosynthetic) ciliates, CO₂ fixation by methanogenic bacterial endosymbionts that utilize H₂ produced as a metabolic end-product by the host cell (Nowack and Melkonian, 2010; Allen *et al.*, 2011; Carpenter *et al.*, 2013; Tai *et al.*, 2016).

Among the Trypanosomatidae, endosymbiosis likely confers physiological advantage within nutritionally challenging

Table 1. Genome properties of trypanosomatid endosymbionts and related taxa

Characteristic	<i>Ca. Pan. Nov</i>	<i>f</i> <i>Pan. spp.</i>	<i>Ca. Kin.</i>	<i>Tay. equ</i>	<i>Ach. xyl</i>
Genome size (Mb)	1.16	4.46–6.50	0.74–0.83	1.70	7.36
GC-content (%)	44	63–65	25–33	37	66
No. of protein-coding genes	968	4181–5342	670–742	1556	6815
No. of pseudogenes	13	76–361	1–20	0	0
Reference	Kostygov <i>et al.</i> (2017)		Alves <i>et al.</i> (2013) Silva <i>et al.</i> (2018)		

Ca. Pan. nov., *Candidatus* *Pandoraea novymonadis*; *f* *Pan. spp.*, free-living *Pandoraea* species; *Ca. Kin.*, *Candidatus* Kinetoplastibacterium; *Tay. equ.*, *Taylorella equigenitalis* (pathogenic bacterium closely related to *Ca. Kinetoplastibacterium*; Alves *et al.*, 2013); *Ach. xyl.*, *Achromobacter xylosoxidans* (free-living bacterium closely related to *Ca. Kinetoplastibacterium*; Alves *et al.*, 2013).

environments offered by the digestive tracts of their invertebrate vectors (or hosts). However, it is neither N₂ nor CO₂ fixation or an ability to utilize or provide alternative carbon sources or electron acceptors for energy generation that differentiate endosymbiont-bearing trypanosomatids from other trypanosomatids. Instead, their endosymbionts render strigomonads and *Novymonas* autotrophic for vitamins (or cofactor precursors), amino acids, purines, and heme which are all essential nutrients in other trypanosomatids (Table 2). The curious exception is the endosymbiont from *Kentomonas sorsogonicus*, which is missing the haem biosynthetic pathway and the host cell is thus reliant upon an exogenous source of haem within its culture medium (Silva *et al.*, 2018). As highlighted in Table 2, in many instances complete biosynthetic pathways are encoded within endosymbiont genomes; in other instances, metabolite exchange between endosymbiont and host is required to complete amino acid, haem or vitamin provision. For *Angomonas deanei*, *Strigomonas culicis* and *S. oncopelti*, predictions for autotrophy arising from genome annotations are consistent with early descriptions of minimal culture media (Newton, 1957; Mundim *et al.*, 1974; De Menezes *et al.*, 1991).

Whether the enhanced autotrophies of endosymbiont-containing trypanosomatids serve to widen the range of vectors that can be colonized and/or offers these trypanosomatids a competitive edge over other microbiota that may compete for the gut niche is not known. At first glance, the relative rarity of endosymbiont-containing trypanosomatids in ecological surveys argues against either of these possibilities. However, it is moot whether susceptibility to antibiotics typically applied during isolation into the culture of trypanosomatids from ecological surveys limits the frequency with which endosymbiont-bearing taxa are found. Insect digestive tracts colonized by trypanosomatids are ill-understood environments, but although they clearly provide sufficient haem, purines, vitamins of the group B and other precursors to support parasite replication in different regions of the alimentary tract, they are also unequivocally nutritionally challenging environments. Several pieces of evidence support this assertion of a nutritional 'knife-edge': (i) with rare exception, trypanosomatid species present (in comparison with other parasites) complex and robust metabolic networks for central energy metabolism and anabolism (notably in the extent of sterol and other lipid biosynthetic pathways) (Ginger, 2006; Kraeva *et al.*, 2015; Opperdoes *et al.*, 2016); (ii) retention in some trypanosomatids of enzymes to (a) complete biosynthetic pathways for which gut microbiota can provide initial precursors – e.g. the

importance of homoserine kinase coupled to the expression of threonine synthase in tsetse-dwelling forms of the African trypanosome *Trypanosoma brucei* (Ong *et al.*, 2015) or (b) catabolize carbon sources likely specific to the insect vectors of some trypanosomatids – e.g. histidine in the reduviid vector of the American trypanosome *T. cruzi* (Berriman *et al.*, 2005); (iii) the extensive reductive evolution of central metabolism that does occur in trypanosomatids when they become adapted to live in particularly nutrient rich environments – e.g. *Phytomonas* in sugar-rich plant sap (Kořený *et al.*, 2012; Porcel *et al.*, 2014) or kinetoplast loss in mechanically – rather than tsetse-transmitted African trypanosomes (Lai *et al.*, 2008). Intriguingly, the loss of respiratory complexes III and IV in *Phytomonas* (Nawathean and Maslov, 2000) may have helped facilitate an ability of *P. françai* to colonize its cyanide-rich cassava host. Comparative analysis of proteome and annotated genomes of endosymbiont-containing *A. deanei* and *S. culicis* have indicated no obvious moderation of the central metabolic networks seen in better studied *Leishmania* or *Trypanosoma* parasites (Motta *et al.*, 2013).

Interface with host cell biology I: strigomonads the slavers; *Novymonas* the farmer

Despite some variations in cell shape, all endosymbiont-containing trypanosomatids adopt liberform morphologies where the flagellum is not attached for an extended region to the cell body following exit from the flagellar pocket (Fig. 2).

In strigomonads, their endosymbiont is positioned proximate to the nucleus and its replication and division in the cell cycle entrained (Motta *et al.*, 2010): endosymbiont duplication occurs early in the cell cycle preceding the host cell's discrete kinetoplast S-phase and segregation, which is coupled to flagellar basal body segregation (Ogbadoyi *et al.*, 2003); endosymbiont division is followed by movement of the endosymbionts such that each is positioned on opposite outer-faces of the nucleus; mitosis (with each nucleus associated with a single endosymbiont) and new flagellum elongation beyond the flagellar pocket exit point conclude the latter stages of the cell cycle prior to cytokinesis. Annotation of *Ca. Kinetoplastibacterium* genomes reveal they lack much of the machinery associated with bacterial cell division, indicating involvement from the host cell in that regard (Alves *et al.*, 2013; Motta *et al.*, 2013). The co-ordination of endosymbiont division within that of the host cell is illustrated further by the effect of the addition of aphidicolin, an inhibitor of eukaryotic replication DNA polymerases, or the eukaryotic translation inhibitor cycloheximide to *A. deanei* or *S. culicis* (Catta-Preta *et al.*, 2015). Application of either eukaryotic growth inhibitor resulted in cessation of host cell growth and division and also blocked endosymbiont division but not endosymbiont replication. Application of aphidicolin in *S. culicis* additionally caused filamentation of bacteria indicating re-entry of the endosymbiont into subsequent cell cycles and continued DNA replication but without any completion of cytokinesis (Catta-Preta *et al.*, 2015).

Candidatus Pandoraea novymonadis replicates more readily within the cytoplasm of its host cell (Fig. 3A and B). In multiplicative *N. esmeraldas* promastigotes, ~70% of the population contain between two and six endosymbionts, with 10 or more present in ~5% of cells (Kostygov *et al.*, 2016). Approximately 6% of *Novymonas* cells are aposymbiotic although the extreme difficulty in cloning such cells, the retention of intracellular bacteria in cultures since their isolation, and the significant deceleration of an aposymbiotic cell line growth as compared with wild-type highlight the importance of *Ca. Pandoraea* to host cell fitness (Kostygov *et al.*, 2017). This contrasts with strigomonads where aposymbiotic populations, albeit replicating more slowly than parental lines and with increased nutritional

Table 2. Metabolic gains for endosymbiont-containing trypanosomatids

Metabolic gain	RT	<i>Ne</i>	<i>A/K/S</i>
Haem biosynthesis	–	+	+ ^a
Purine provision	–	+	+
Branched chain a.a. synthesis (leu, iso, val)	–	+	+ ^b
Aromatic a.a. synthesis (phe, trp, tyr)	–	+	+
Lys biosynthesis	–	+	+
<i>De novo</i> folic acid production	–	+	+
Thiamine, nicotinic acid, biotin provision	–	+	–
Riboflavin, pantothenic acid, vitamin B ₆ provision	–	+	+ ^c

RT, regular trypanosomatids; *Ne*, *Novymonas esmeraldas*; *A/K/S*, *Angomonas/Kentomonas/Strigomonas*.

^aWith the exception of the endosymbiont from the sole characterized *Kentomonas* species (*K. sorsogonicus*) where the haem biosynthetic pathway is absent from both host and its endosymbiont (Silva *et al.*, 2018).

^bRequires use of host cell branched-chain amino acid aminotransferase.

^cPantothenic acid synthesis utilizes enzymes from host and endosymbiont.

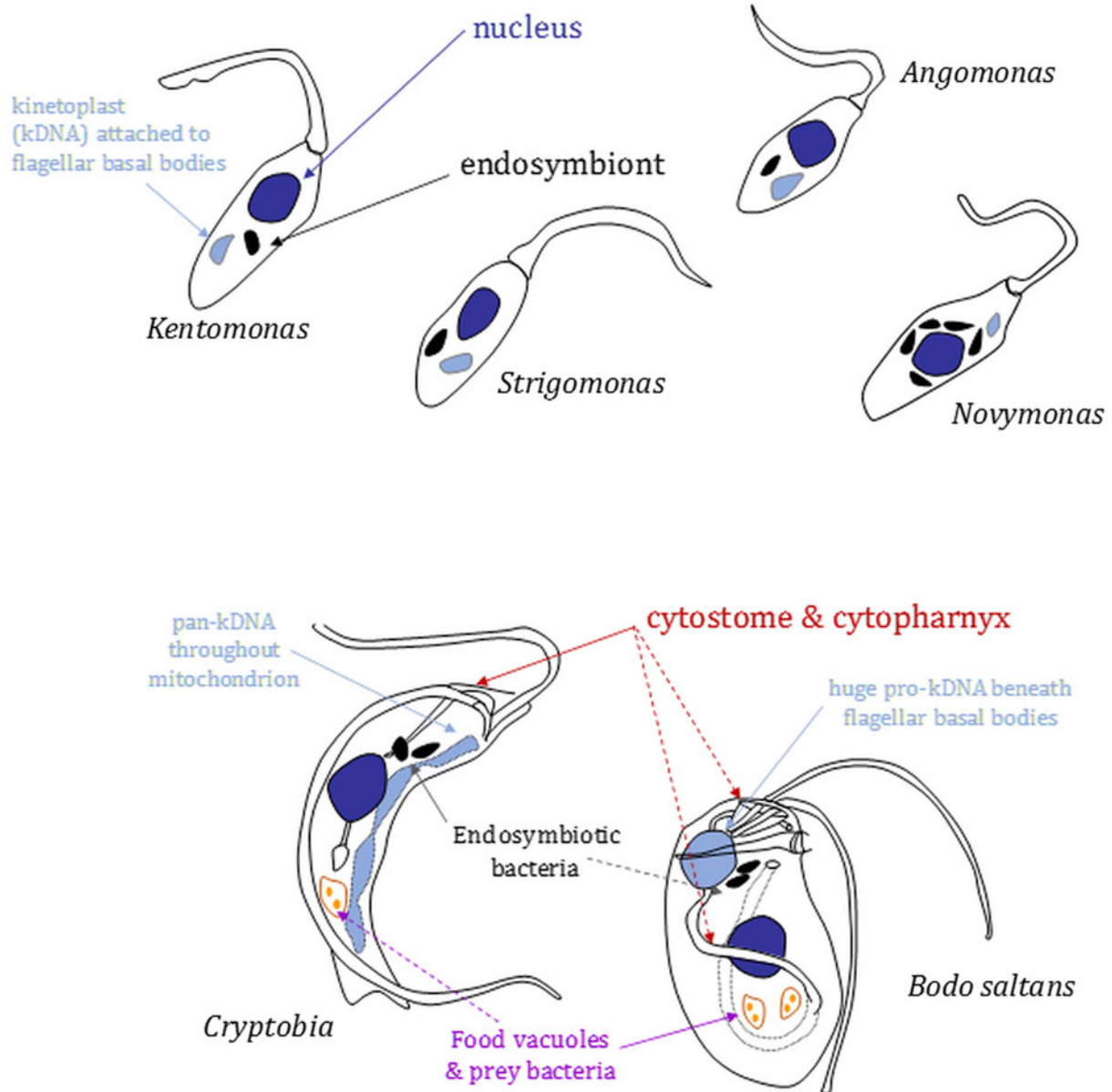


Fig. 2. Morphology and nucleus–mitochondrial genome–endosymbiont organization in endosymbiont-containing kinetoplastids. Cartoons (not to scale) are based on images shown in Kostygov *et al.* (2016), Teixeira *et al.* (2011) and Votýpka *et al.* (2014) or original drawings in Brooker (1971a) and Vickerman (1977). Relative positions of several organelles discussed in the main text are shown. Shading: black, bacterial endosymbionts; dark grey, nuclei; light grey, mitochondrial genomes [kinetoplasts (kDNA) or (in *Cryptobia*) pan-kDNA and (in *Bodo saltans*) pro-kDNA].

requirements, can be readily obtained by treatment of cultures with chloramphenicol (de Souza and Motta, 1999). Intriguingly, studies of aposymbiotic strigomonads reveal another possible dimension to the host–endosymbiont interface with differences evident in cell surface carbohydrate composition between symbiont-containing and symbiont-lacking *S. culicis* cultivated in equivalent media and the indication that altered surface composition negatively influences interaction of the trypanosomatid with permissive insect hosts (Dwyer and Chang 1976; Catta-Preta *et al.*, 2013; d’Avila-Levy *et al.*, 2015). Significantly, culture conditions have been shown to influence the composition of the cell surface of other trypanosomatids, demonstrating common links between nutritional status and cell surface properties (Vassella *et al.*, 2000; Morris *et al.*, 2002).

Fusion of *Novymonas* lysosomes with *Ca. P. novymonadis* provides an indication that the host ‘farms’ its endosymbiont, presumably taking amino acids, haem, purines and other molecules liberated in lysosomes to satisfy dietary requirements. In

agreement with the ‘lax’ control on endosymbiont multiplication evident in *Novymonas*, *Ca. P. novymonadis* retains more genes associated with bacterial cell division than *Ca. Kinetoplastibacterium* (Kostygov *et al.*, 2017). However, other findings from *Ca. P. novymonadis* genome annotation point to a well-established host–endosymbiont relationship and provide a note of caution for any assumption of how readily *Ca. P. novymonadis* might multiply free from the host cell in different, commonly used bacterial growth media. For instance, cellular characteristics associated with perception and response to environmental change are either absent (genes for pilus and flagellum assemblies, ‘wsp’ chemotaxis proteins, ‘pel’ proteins involved in biofilm formation) or minimized (two-component signalling). There is also a drastic reduction in the number of nutrient transporters/exporters present, including members of ABC-transporter and major facilitator superfamilies and in the ability of *Ca. P. novymonadis* to catabolize diverse carbon sources in comparison with free-living *Pandora* (Fig. 4).

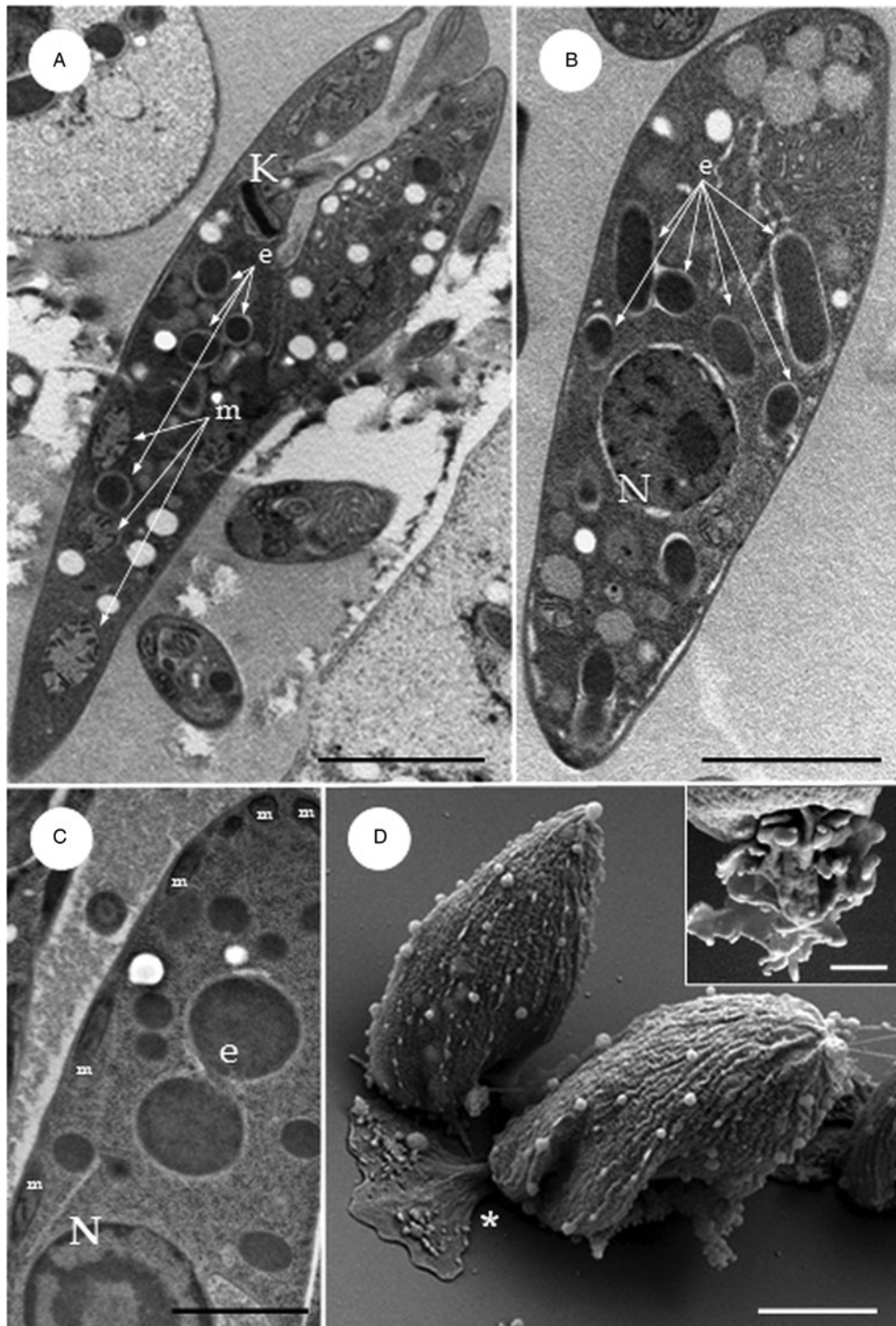


Fig. 3. Electron microscopy of the endosymbiont–host cell association and cell form in *Novymonas* and *Kentomonas*. (A and B) Longitudinal sections through *N. esmeraldas* promastigotes showing the presence of multiple endosymbiont profiles (e). Also highlighted are the kinetoplast (K), nucleus (N) and cross-sections through the mitochondrion (m). (C) Longitudinal section through a *Kentomonas sorsogonicus* choanomastigote illustrating (i) a dividing bacterial endosymbiont and (ii) mitochondrial hypertrophy and loss of typical microtubule spacing within the sub-pellicular array. (D) Sessile *N. esmeraldas* choanomastigote attached to the substrate surface via a modified flagellum (asterisk). Inset, the modified flagellum of a sessile choanomastigote revealing a possible open collar structure to the flagellar pocket exit point. Scale bars (A) and (B) 2 μm ; (C) 1 μm ; (D) 2 μm (inset, 400 nm). Images in (D) are reproduced from Kostygov *et al.* (2016) under the terms of a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported licence.

However, metabolic dependencies in endosymbiotic relationships go both ways. In the trypanosomatid examples, owing in large part to the close proximity of strigomonad endosymbionts to host cell mitochondria and glycosomes, strigomonads have for many years been considered to provide ATP to their intracellular partners (see Loyola-Machado *et al.*, 2017 for recent consideration of this topic). This assertion is supported by the paucity of options for efficient oxidative phosphorylation by *Ca. Kinetoplastibacterium* spp.

Genomes of both *Ca. Kinetoplastibacterium* and *Ca. P. novymonadis* contain genes for *nuo*-type NADH:ubiquinone oxidoreductases (Kostygov *et al.*, 2017), but in the former its electron transport chain is truncated to a cytochrome *bd* terminal oxidase for transfer of electrons from ubiquinone to O_2 – the type of terminal oxidase favoured by numerous bacteria, including *Escherichia coli* under low O_2 availability. In contrast to *Ca. Kinetoplastibacterium* spp., however, whilst the carbon source(s)

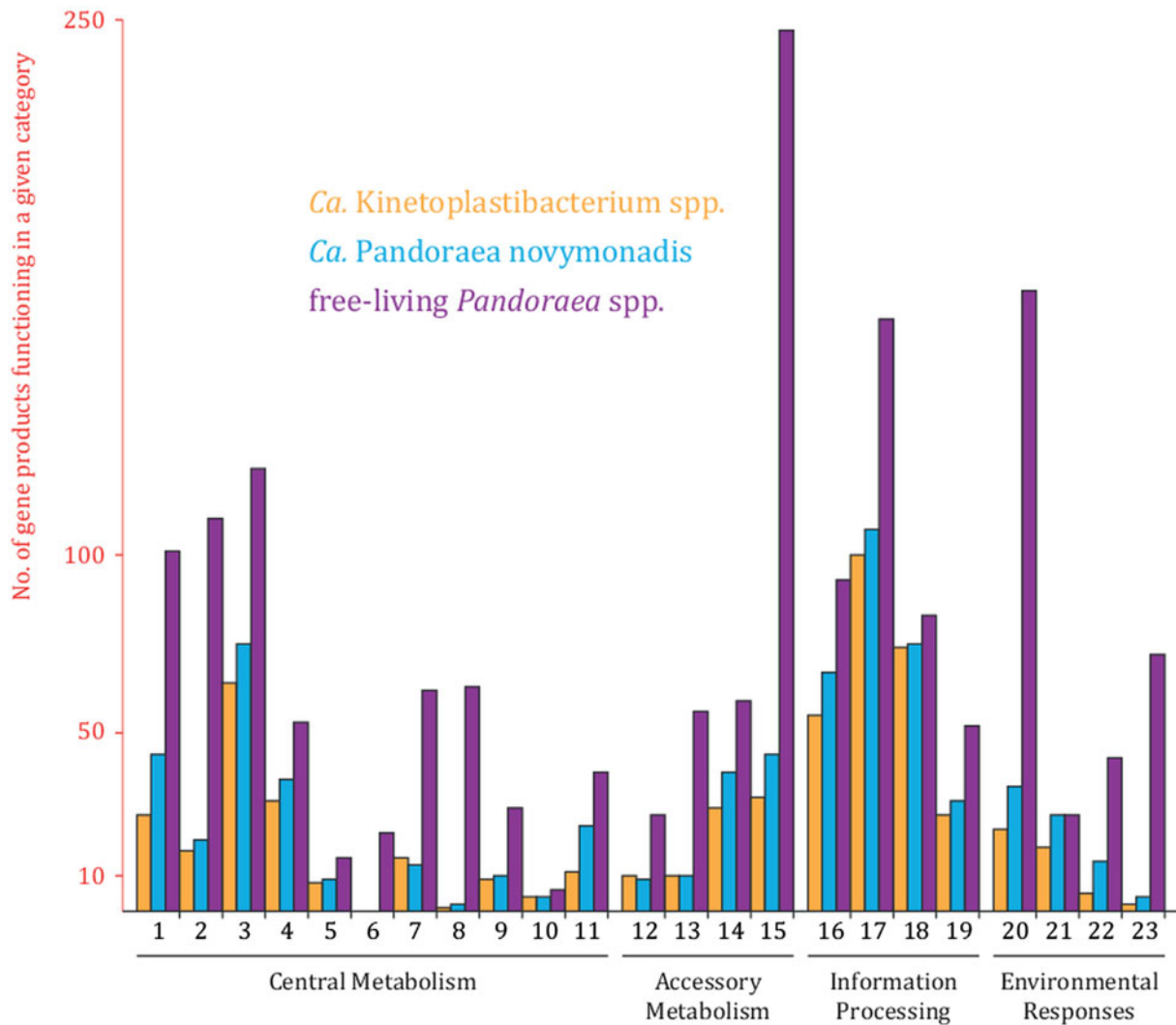


Fig. 4. *In silico* annotated proteomes illustrate the reductive evolution of *Ca. Pandoraea novymonadis* and *Ca. Kinetoplastibacterium*. Predicted protein repertoires for *Ca. P. novymonadis*, 5 *Ca. Kinetoplastibacterium* spp. and 11 free-living *Pandoraea* species (Kostygov *et al.*, 2017) were analysed according to within the KEGG Orthology (KO). 2728 KO functions were analysed. For *Ca. Kinetoplastibacterium* spp. and *Pandoraea* spp. annotation of gene products in 3 or 5 genomes, respectively, were required for inclusion in the chart shown. Known nearest free-living relatives of *Ca. Kinetoplastibacterium* are evolutionarily more distant than for *Ca. P. novymonadis*, and were not therefore included in the analysis although we note the closest *Ca. Kinetoplastibacterium* free-living relative, *A. xylosoxidans*, is more gene-rich than free-living *Pandoraea* spp. (Table 2). Individual gene products were scored once and appear in only one of the following categories. *Central metabolism*: category 1, carbohydrate usage (including lipopolysaccharide and peptidoglycan assembly); 2, amino acid catabolism; 3, amino acid biosynthesis (including glycolysis); 4, fatty acid and terpenoid metabolism; 5, inositol phosphate and glycerophospholipid metabolism; 6, butanoate and propanoate metabolism; 7, pyruvate, glyoxylate, and dicarboxylate metabolism; 8, degradation of aromatics; 9, pentose phosphate and antioxidant metabolism; 10, Krebs cycle; 11, respiration and oxidative phosphorylation. *Accessory metabolism*: 12, porphyrin metabolism; 13, miscellaneous (including carbon fixation, sulphur and methane metabolism, urease); 14, vitamin and cofactor biosynthesis; 15, transporters and ATPases. *Information processing*: 16, replication and DNA repair; 17, purine and pyrimidine metabolism (including tRNA processing and core transcription); 18, ribosome and translation; 19, chaperones. *Environmental responses*: 20, two-component signaling, transcriptional regulation, quorum sensing and phosphate metabolism; 21, cell division; 22, secondary metabolism and antibiotic defence/attack; 23, flagellum, pilus, biofilm formation.

utilized by *Novymonas* endosymbionts remains enigmatic – fructose, common in the diet of plant-feeding insects, is the most likely carbon source (Kostygov *et al.*, 2017) – the novymonad endosymbiont appears more self-sufficient for energy generation. Perhaps as a consequence of their greater autonomy with regard to their rate of cell division, and thus a greater need for intra-symbiont ATP generation, *Ca. P. novymonadis* retains a more expansive electron transport chain. Here the metabolism includes a capacity for oxidative phosphorylation from *c*-type cytochrome-dependent respiration.

Currently, the least explored facet of the interface from host to endosymbiont is the degree to which the host cell targets nuclear-encoded proteins to the symbiont. One example is known for *A. deanei* (Morales *et al.*, 2016b), but this is a long way short of the number of host-targeted proteins that might be required to question whether trypanosomatid endosymbionts begin to blur boundaries between endosymbiont and organelles.

Interface with host cell biology II: symbiont acquisition by closed-mouth, osmotrophic trypanosomatids – how?

In contrast to phagotrophic bodonids and other free-living kinetoplastids, trypanosomatids are obligate osmotrophs. A robust sub-pellicular mono-layer of microtubules cross-linked to one another and the over-laying plasma membrane provides a corset that defines characteristic trypanosomatid cell morphologies and prevents general endocytosis or membrane invagination across the cell surface. Membrane invagination occurs only at points where the sub-pellicular corset is absent which, in well-studied African trypanosomes and *Leishmania*, is where the flagellar pocket forms around the single flagellum emerging from the cell body. In these trypanosomatids, the flagellar pocket is the site of endo- and exocytic traffic (Field and Carrington, 2009). At the flagellum exit point, an essential collar marks the flagellum

pocket boundary (Bonhivers *et al.*, 2008) limiting the size and rate of macromolecular traffic into the pocket lumen (Gadelha *et al.*, 2009). Given these constraints, how, following radiation of various trypanosomatid lineages, have trypanosomatid–endosymbiont associations occurred on at least two occasions?

Several possibilities can explain the conundrum of how *Novymonas* and a strigomonad ancestor acquired their respective bacterial endosymbionts. Conserved in free-living kinetoplastids and present in some trypanosomatids (Brooker, 1971a, 1971b; Brugerolle *et al.*, 1979; Attias *et al.*, 1996; Alcantara *et al.*, 2017; Skalický *et al.*, 2017) is a cytotome–cytopharynx complex, sitting in close proximity to the flagellar pocket (Figs 1 and 5). In *T. cruzi* (Porto-Carreiro *et al.*, 2000) and apparently in *Crithidia fasciculata* (Brooker, 1971b) the cytotome is a site of endo- and

pinocytosis. In free-living kinetoplastids, the cytotome leading to the cytopharynx, in conjunction with the anterior flagellum, is used for phagotrophic feeding on bacterial prey. Early microscopy analyses indicate extensive distension of the feeding apparatus in order to ingest large prey (Brooker, 1971a; Burzell, 1973, 1975). Enzymatic machinery necessary for digestion of complex macromolecular structures from live prey is considered to have been lost at an early point following divergence of the last common trypanosomatid ancestor (Skalický *et al.*, 2017), coincident with the advent of obligate osmotrophy but also indicating that fortuitous uptake of a bacterium by a cytotome-bearing trypanosomatid would not necessarily be followed by its digestion.

Although clearly absent from African trypanosomes and *Leishmania* (Skalický *et al.*, 2017), a paucity of data cannot

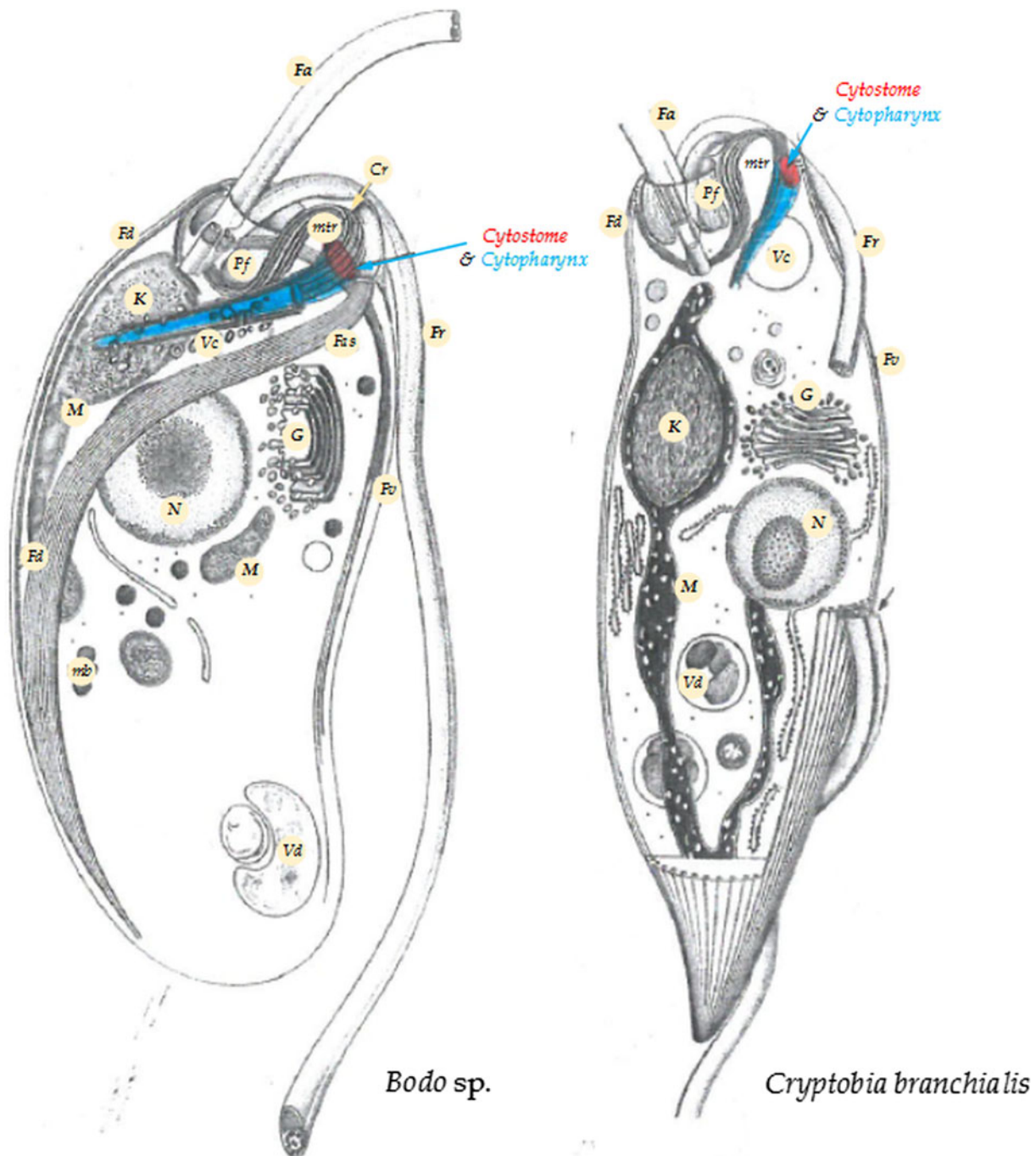


Fig. 5. Relative positions of flagella, cytotome, cytopharynx and other cellular features in free-living *Bodo* and *Cryptobia* kinetoplastids. Images were adapted from original drawings in Figs 4–6 from Brugerolle *et al.* (1979). Abbreviations (translated from the original French): Cr, oral ridge; Fas, 'microtubule fibre' associated with the 'striatal plaque'; Fd, 'dorsal fibre'; Fr, recurrent flagellum; Fv, 'ventral fibre'; Fa, anterior flagellum; G, Golgi; K, kintetoplast; M, mitochondrion; mb, microbodies; mtr, 'reinforced microtubules'; N, nucleus; Pf, flagellar pocket; Vc, contractile vacuole; Vd, food vacuole.

yet allow insight into how often and when the cytostome–cytopharynx was lost during trypanosomatid evolution. Whilst this organelle complex has never been seen from detailed ultrastructural analyses of extant strigomonads (Bombaça *et al.*, 2017; Loyola-Machado *et al.*, 2017) or analysis of *Phytomonas* sp. (e.g. Postell and McGhee, 1981; Milder *et al.*, 1990) and functionality of the *Crithidia* ‘cytostome’ has not, to our knowledge been revisited since the early 1970s, the critical questions are whether an ancestral cytostome was present and could have played a role in endosymbiont uptake by strigomonad and/or novymonad ancestors. A cytostome–cytopharynx is retained in the basal trypanosomatid *Paratrypanosoma confusum* (Skalický *et al.*, 2017); coupled to the monophyly of the trypanosomes, plus the relatively close relationship between the leishmanias and *C. fasciculata*, the pattern of organelle degeneration and thence loss was likely complex. The observation that cytostome–cytopharynx assembly in *T. cruzi* is stage-regulated (Vidal *et al.*, 2016) also leaves open the possibility of a cryptic or hidden cytostome in other extant trypanosomatids. Thus, cell entry *via* a cytostome is a plausible route for the acquisition of *Novymonas* or strigomonad endosymbionts.

To consider alternative acquisition routes, a hypertrophied mitochondrion is a diagnostic trait for the Strigomonadinae and its invasion of the spacing between sub-pellicular microtubules (Fig. 3C) is often considered to be a consequence of endosymbiosis with the ATP requirements of the endosymbiont driving mitochondrial expansion and an increased rate of energy generation by the host cell. Looser organization of the kinetoplast, relative to other trypanosomatids, is another strigomonad-specific characteristic (Teixeira *et al.*, 2011; Votýpka *et al.*, 2014), conceivably facilitates high rates of mitochondrial gene expression, and, thus, potentially an enhanced capacity for oxidative phosphorylation relative to some other trypanosomatids (careful, cross-species quantitative assessment of metabolic rate as a function of growth rate(s) under equivalent conditions will be necessary to determine if this is the case). Considered less often, however, is the possibility that mitochondrial hypertrophy and/or disruption of sub-pellicular microtubule spacing preceded endosymbiont acquisition. In this instance, a release of constraints on plasma membrane invagination would facilitate another route for endosymbiont uptake in the ancestor of the Strigomonadinae.

Looking further at the influence(s) of mitochondrial hypertrophy, rather than the endosymbiont itself might exert on host cell biology, then another strigomonad synapomorphy is the extensive reduction of paraflagellar rod (PFR) architecture. This results in a vestigial structure extended along only the proximal third of the axoneme (Gadelha *et al.*, 2006). Reductive PFR evolution was driven, at least in part, by the loss of genes encoding the major PFR2 protein. The extreme alteration of PFR form is intriguing not least because of the essentiality of this flagellar structure in other trypanosomatids (Maga *et al.*, 1999; Ginger *et al.*, 2013; Lander *et al.*, 2015). If the view that the PFR provides an important function in maintaining intraflagellar nucleotide homeostasis is correct (Pullen *et al.*, 2004; Ginger *et al.*, 2008), then a significant increase to the efficiency of mitochondrial ATP production in strigomonads could have provided a selective driver for the enigmatic reduction of PFR form seen in this trypanosomatid group.

Mitochondrial hypertrophy is not so evident in *Novymonas* with sub-pellicular microtubules having a spacing reminiscent of that found in most trypanosomatids. Acquisition of its symbiont is thus unlikely to have occurred *via* invagination of the plasma membrane. Sessile *N. esmeraldas* choanomastigotes, however, attach to surfaces *via* their flagellum and attached in this way exhibit a drastically altered flagellum structure (Fig. 3D) reminiscent of the flagellum surface attachment remodelling seen also in *P. confusum* (Skalický *et al.*, 2017). In scanning electron

micrographs of detached *N. esmeraldas* choanomastigotes (Fig. 3D; inset) the altered flagellum morphology hints at a more open flagellar pocket collar through which a flagellum membrane-attached bacterium could putatively be ingested.

Endosymbioses within free-living phagotrophic kinetoplastids

There is currently sparse data with regard to endosymbionts and their role(s) in free-living phagotrophic kinetoplastids. This is not surprising given that attention to their molecular cell biology using modern approaches is only recently forthcoming (Gomaa *et al.*, 2017). However, constraints that leave the conundrum of how at least two trypanosomatids acquired their endosymbionts – arrayed sub-pellicular microtubules; a closed flagellar pocket – are not conspicuous among free-living kinetoplastids. Plus, there are likely significant insights to be made with regard to niche adaptation and exploitation; anoxic environments provide an obvious example with kinetoplastids being one of the few protist groups for which there is only limited evidence of adaptation (Priya *et al.*, 2008). The current lack of known anaerobic kinetoplastids contrasts with observations of obligately aerobic metabolism in trypanosomatid and *Bodo saltans* genomes that nonetheless showcases several anaerobic hallmarks (Michels *et al.*, 1997; Annoura *et al.*, 2005; Opperdoes *et al.*, 2016).

Surveying the literature indicates that the presence of endosymbiotic bacteria is not an obligate characteristic of free-living kinetoplastids, e.g. an absence from *Rhynchomonas metabolita* (Burzell, 1973). When present, however, endosymbionts are found in the anterior region of the cytoplasm or in close proximity to the nucleus of other kinetoplastids, albeit far from the posterior cell region that tends to be dominated by food vacuoles containing bacteria ingested *via* the cytostome–cytopharynx (Fig. 2) (Brooker 1971a; Burzell, 1975; Vickerman, 1977). Likely bacterial epibionts have been noted on the surface of *Cryptobia vaginalis* (Vickerman, 1977) and in others endosymbiont multiplication keeps pace with host cell division and a reduced peptidoglycan layer of the endosymbiont cell wall is in evidence, again indicative of the establishment of long-term endosymbioses.

Only distantly related to the kinetoplastids, but nonetheless of interest, another clade of euglenozoans – Symbiontida – is characterized by a dense layer of ectosymbiotic bacteria, present on their surface. This poorly studied group of flagellates, consisting of only three known species, inhabit low-oxygen sea environments. The function of the ectosymbionts is not known (Yubuki *et al.*, 2013).

Perkinsella: the enslaved kinetoplastid

The ancestor of *Perkinsella*, which is most closely related to the fish ectoparasite *Ichthyobodo*, is thought to have diverged early in kinetoplastid evolution (Fig. 1). Extant *Perkinsella* is an obligate endosymbiont of lobose amoebae genus *Paramoeba* (phylum Amoebozoa), which are pathogenic to a variety of marine animals, including farmed fish. GC-content in *Perkinsella* is not reduced in comparison with other sampled kinetoplastids (Tanifuji *et al.*, 2017) even though the *Perkinsella*–*Paramoeba* endosymbiosis is a long time established association (Sibbald *et al.*, 2017). In contrast, gene content of *Perkinsella* is significantly reduced in comparison with free-living *B. saltans* and parasitic trypanosomatids – 5252 protein-coding genes in *Perkinsella* vs 18 943 genes in *B. saltans*; 6381 in *Phytomonas* sp.; 9068 in *T. brucei* (although this includes expansion of its critical antigenic variant surface glycoprotein gene repertoire); and 8272 genes in *Leishmania major*. This reductive evolution reflects the secondary loss of much of the cell biology that characterizes kinetoplastid

cell form (Tanifuji *et al.*, 2017). *Ichthyobodo*, in contrast, displays the biflagellate morphology typical of non-trypanosomatid kinetoplastids (Grassé, 1952).

Lost from the genome of *Perkinsela* are all the genes required for basal body/flagellum assembly and architecture, together with an absence of genes encoding homologues of trypanosomatid cytoskeletal proteins. The absence of sub-pellicular microtubules relieves the constraints on the surface siting of endocytosis and leaves the endosymbiont able to readily ingest cytoplasm from the host (Tanifuji *et al.*, 2017). Metabolism of *Perkinsela* is also minimized: glycolysis occurs but obvious metabolic routes from pyruvate to acetyl-CoA are lacking; a truncated Krebs' cycle running from α -ketoglutarate to oxaloacetate likely uses a (host-derived) glutamate carbon source and provides electrons to fuel a mitochondrial respiratory chain truncated by the loss of complex I (NADH:ubiquinone oxidoreductase). It is likely that the benign environment offered by the *Paramoeba* host, with respect to carbon provision, facilitates the reductive evolution of intermediary metabolism. An absence of sterol metabolism potentially reflects either absence of sterol from endosymbiont membranes (similar to a few other eukaryotes) or a possibility that the host provides an easy availability of the ergosta- and stigmasta-type sterols found in other amoebozoans and trypanosomatids (Raederstorff and Rohmer, 1985; Nes *et al.*, 1990; Roberts *et al.*, 2003). A lack of sugar nucleotide biosynthesis possibly indicates a reduced requirement for protein glycosylation or limited need for investment in a protective cell surface glycocalyx.

Benefits arising from an intracellular lifestyle for *Perkinsela* are clear, although this is not to suggest that the lifestyle is lazy: the kinetoplastid makes a huge investment in RNA editing, perhaps as a consequence of the neutral evolutionary ratchet discussed by Lukeš (2011), for the expression of the six (essential) respiratory chain components encoded on the mitochondrial genome (David *et al.*, 2015) and the nuclear genome hints at the presence of a sexual cycle that is perhaps integrated within that of its host (Tanifuji *et al.*, 2017).

Apart from *Paramoeba* and *Perkinsela*, all known endosymbioses involving only eukaryotes bring the provision of photosynthesis to the host partner (David *et al.*, 2015). What *Paramoeba* derives from its unusual endosymbiont is currently a mystery and a source only for speculation.

Concluding remarks

Endosymbiosis is a feature of kinetoplastid evolution. Several case examples provide tractable opportunities to understand how, at the host–endosymbiont interface, long-lasting endosymbiotic relationships become established in microbial eukaryotes and leave other questions that will likely be more challenging to address. Of the latter, until more robust culture systems for *Paramoeba* are forthcoming, it will be difficult to establish what *Perkinsela* provides for its host. Similarly, without relevant traits being revealed in continuing surveys of trypanosomatid diversity, the chronology and interplay between endosymbiont acquisition, mitochondrial hypertrophy, altered kinetoplast structure and PFR reduction cannot realistically be addressed. However, we now work in an era of easy next-generation sequencing. Thus, paralleling combined genomic, transcriptomic and proteomic studies of environmentally sourced protists such as the breviate *Lenisia limosa* (Hamann *et al.*, 2016), there is much scope to reveal what endosymbionts (and epibiotic bacteria) contribute to free-living kinetoplastid hosts. Similarly, and in contrast to many other examples of protists with endosymbionts, the tractability of trypanosomatids towards genetic manipulation (Morales *et al.*, 2016b) leaves huge opportunity to dissect at a molecular

level the regulatory influences on endosymbiont growth and division in strigomonads and *Novymonas*. Genetic tractability also provides the means to probe the extent to which protein targeting from host-to-symbiont (and perhaps *vice versa*) also eclipses the biology of trypanosomatid–endosymbiont associations.

Financial support. This work received support from the Czech Grant Agency (16-18699S to J.L.) and the European Regional Development Fund [project CePaViP (OPVVV16_019/0000759 to J. L. and V. Y.)].

Conflict of interest. None.

Ethical standards. Not applicable.

References

- Akhoundi M, Downing T, Votýpka J, Kuhls K, Lukeš J, Cannet A, Ravel C, Marty P, Delaunay P, Kasbari M, Granouillac B, Gradoni L and Sereno D (2017) *Leishmania* infections: molecular targets and diagnostics. *Molecular Aspects of Medicine* 57, 1–29.
- Alcantara CL, Vidal JC, de Souza W and Cunha-E-Silva NL (2017) The cytostome-cytopharynx complex of *Trypanosoma cruzi* epimastigotes disassembles during cell division. *Journal of Cell Science* 130, 164–176.
- Allen A, Dupont C, Oborník M, Horák A, Nunes-Nesi A, McCrow J, Zheng H, Johnson D, Hu H, Fernie A and Bowler C (2011) Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. *Nature* 473, 203–209.
- Alves JM, Serrano MG, Maia da Silva F, Voegtly LJ, Matveyev AV, Teixeira MM, Camargo EP and Buck GA (2013) Genome evolution and phylogenomic analysis of *Candidatus* Kinetoplastibacterium, the betaproteobacterial endosymbionts of *Strigomonas* and *Angomonas*. *Genome Biology and Evolution* 5, 338–350.
- Annoura T, Nara T, Makiuchi T, Hashimoto T and Aoki T (2005) The origin of dihydroorotate dehydrogenase genes of kinetoplastids, with special reference to their biological significance and adaptation to anaerobic, parasitic conditions. *Journal of Molecular Evolution* 60, 113–127.
- Aphasizhev R and Aphasizheva I (2014) Mitochondrial RNA editing in trypanosomes: small RNAs in control. *Biochimie* 100, 125–131.
- Attias M, Vommario RC and de Souza W (1996) Computer aided three-dimensional reconstruction of the free-living protozoan *Bodo* sp. (Kinetoplastida: Bodonidae). *Cell Structure and Function* 21, 297–306.
- Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, Lennard NJ, Caler E, Hamlin NE, Haas B, Böhme U, Hannick L, Aslett MA, Shallom J, Marcello L, Hou L, Wickstead B, Alsmark UC, Arrowsmith C, Atkin RJ, Barron AJ, Bringaud F, Brooks K, Carrington M, Cherevach I, Chillingworth TJ, Churcher C, Clark LN, Corton CH, Cronin A, Davies RM, Doggett J, Djikeng A, Feldblyum T, Field MC, Fraser A, Goodhead I, Hance Z, Harper D, Harris BR, Hauser H, Hostetler J, Ivens A, Jagels K, Johnson D, Johnson J, Jones K, Kerhornou AX, Koo H, Larke N, Landfear S, Larkin C, Leech V, Line A, Lord A, Macleod A, Mooney PJ, Moule S, Martin DM, Morgan GW, Mungall K, Norbertczak H, Ormond D, Pai G, Peacock CS, Peterson J, Quail MA, Rabinowitz E, Rajandream MA, Reitter C, Salzberg SL, Sanders M, Schobel S, Sharp S, Simmonds M, Simpson AJ, Tallon L, Turner CM, Tait A, Tivey AR, Van Aken S, Walker D, Wanless D, Wang S, White B, White O, Whitehead S, Woodward J, Wortman J, Adams MD, Embley TM, Gull K, Ullu E, Barry JD, Fairlamb AH, Opperdoes F, Barrell BG, Donelson JE, Hall N, Fraser CM, Melville SE and El-Sayed NM (2005) The genome of the African trypanosome *Trypanosoma brucei*. *Science* 309, 416–422.
- Bombaça ACS, Dias FA, Ennes-Vidal V, Garcia-Gomes ADS, Sorgine MHF, d'Avila-Levy CM, Menna-Barreto RFS (2017) Hydrogen peroxide resistance in *Strigomonas culicis*: effects on mitochondrial functionality and *Aedes aegypti* interaction. *Free Radical Biology and Medicine* 113, 255–266.
- Bonhivers M, Nowacki S, Landrein N and Robinson DR (2008) Biogenesis of the trypanosome exo-endocytic organelle is cytoskeleton-mediated. *PLoS Biology* 6, e105.
- Brooker BE (1971a) Fine structure of *Bodo saltans* and *Bodo caudatus* (Zoomastigophora: Protozoa) and their affinities with the Trypanosomatidae. *Bulletin of the British Museum (Natural History)* 22, 89–102.

- Brooker BE (1971b) The fine structure of *Crithidia fasciculata* with special reference to the organelles involved in the ingestion and digestion of protein. *Zeitschrift für Zellforschung und Mikroskopische Anatomie* **116**, 532–563.
- Brugerolle G, Lom J, Nohynkova E and Joyon L (1979) Comparaison et évolution des structures cellulaires chez plusieurs espèces de Bodonidés et Cryptobiidés appartenent aux genres *Bodo*, *Cryptobia* et *Trypanoplasma* (Kinetoplastida, Mastigophora). *Protistologica* **15**, 197–221.
- Burzell LA (1973) Observations on the proboscis-cytopharynx complex and flagella of *Rhynchomonas metabolite* Pshenin, 1964 (Zoomastigophorea: Bodonidae). *Journal of Protozoology* **20**, 385–393.
- Burzell LA (1975) Fine structure of *Bodo curvifilus* Griessmann (Kinetoplastida: Bodonidae). *Journal of Protozoology* **22**, 35–59.
- Carpenter KJ, Weber PK, Davisson ML, Pett-Ridge J, Haverty MI and Keeling PJ (2013) Correlated SEM, FIB-SEM, TEM, and nanoSIMS imaging of microbes from the hindgut of a lower termite: methods for in-situ functional and ecological studies of uncultivable microbes. *Microscopy and Microanalysis* **19**, 1490–1501.
- Catta-Preta CM, Nascimento MT, Garcia MC, Saraiva EM, Motta MCM and Meyer-Fernandes JR (2013) The presence of a symbiotic bacterium in *Strigomonas culicis* is related to differential ecto-phosphatase activity and influences the mosquito-protozoa interaction. *International Journal for Parasitology* **43**, 571–577.
- Catta-Preta CMC, Brum FL, da Silva CC, Zuma AA, Elias MC, de Souza W, Schenkman S and Motta MCM (2015) Endosymbiosis in trypanosomatid protozoa: the bacterium division is controlled during the host cell cycle. *Frontiers in Microbiology* **6**, 520.
- Cavalier-Smith T (2016) Higher classification and phylogeny of Euglenozoa. *European Journal of Protistology* **56**, 250–276.
- Cavalier-Smith T and Lee JJ (1985) Protozoa as hosts for endosymbiosis and the conversion of symbionts into organelles. *Journal of Protozoology* **32**, 376–379.
- Clayton CE (2014) Networks of gene expression regulation in *Trypanosoma brucei*. *Molecular and Biochemical Parasitology* **195**, 96–106.
- David V, Flegontov P, Gerasimov E, Tanifuji G, Hashimi H, Logacheva MD, Maruyama S, Onodera NT, Gray MW, Archibald JM and Lukeš J (2015) Gene loss and error-prone RNA editing in the mitochondrion of *Perkinsella*, an endosymbiotic kinetoplastid. *MBio* **6**, e01498–15.
- d'Avila-Levy CM, Silva BA, Hayashi EA, Vermelho AB, Alviano CS, Saraiva EM, Branquinha MH and Santos AL (2005) Influence of the endosymbiont of *Blastocrithidia culicis* and *Crithidia deanei* on the glycoconjugate expression and on *Aedes aegypti* interaction. *FEMS Microbiology Letters* **252**, 279–286.
- De Menezes MCND and Roitman I (1991) Nutritional requirements of *Blastocrithidia culicis*, a trypanosomatid with an endosymbiont. *Journal of Eukaryotic Microbiology* **38**, 122–123.
- de Souza W and Motta MCM (1999) Endosymbiosis in protozoa of the Trypanosomatidae family. *FEMS Microbiology Letters* **173**, 108.
- Dorrell RG and Howe CJ (2012) What makes a chloroplast? Reconstructing the establishment of photosynthetic symbioses. *Journal of Cell Science* **125**, 1865–1875.
- Du Y, Maslov DA and Chang KP (1994) Monophyletic origin of beta-division proteobacterial endosymbionts and their coevolution with insect trypanosomatid protozoa *Blastocrithidia culicis* and *Crithidia* spp. *Proceedings of the National Academy of Sciences of the United States of America* **91**, 8437–8441.
- Dwyer DM and Chang KP (1976) Surface membrane carbohydrate alterations of a flagellated protozoan mediated bacterial endosymbionts. *Proceedings of the National Academy of Sciences of the United States of America* **73**, 852–856.
- Dyková I, Fiala I, Lom J and Lukeš J (2003) *Perkinsella amoebae*-like endosymbionts of *Neoparamoeba* spp., relatives of the kinetoplastid *Ichthyobodo*. *European Journal of Protistology* **39**, 37–52.
- Edgcomb VP, Orsi W, Breiner H-W, Stock A, Filker S, Yakimov MM and Stoeb T (2011) Novel active kinetoplastids associated with hypersaline anoxic basins in the Eastern Mediterranean deep-sea. *Deep Sea Research I* **58**, 1040–1048.
- Eme L, Spang A, Lombard J, Stairs CW and Ettema TJJ (2017) Archaea and the origin of eukaryotes. *Nature Reviews Microbiology* **15**, 711–723.
- Esteban GF, Finlay BJ and Clarke KJ (2009) Sequestered organelles sustain aerobic microbial life in anoxic environments. *Environmental Microbiology* **11**, 544–550.
- Field MC and Carrington M (2009) The trypanosome flagellar pocket. *Nature Reviews Microbiology* **7**, 775–786.
- Field MC, Horn D, Fairlamb AH, Ferguson MA, Gray DW, Read KD, De Rycker M, Torrie LS, Wyatt PG, Wyllie S and Gilbert IH (2017) Anti-trypanosomatid drug discovery: an ongoing challenge and a continuing need. *Nature Reviews Microbiology* **15**, 217–231.
- Flegontova O, Flegontov P, Malviya S, Audic S, Wincker P, de Vargas C, Bowler C, Lukeš J and Horák A (2016) Extreme diversity of diplomemid eukaryotes in the ocean. *Current Biology* **26**, 3060–3065.
- Flegontova O, Flegontov P, Malviya S, Poulain J, de Vargas C, Bowler C, Lukeš J and Horák A (2018) Neobodonids are dominant kinetoplastids in the global ocean. *Environmental Microbiology* **20**, 878–889.
- Gadelha C, Rothery S, Morphew M, McIntosh JR, Severs NJ and Gull K (2009) Membrane domains and flagellar pocket boundaries are influenced by the cytoskeleton in African trypanosomes. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 17425–17430.
- Gadelha C, Wickstead B, de Souza W, Gull K and Cunha-e-Silva N (2006) Cryptic paraflagellar rod in endosymbiont-containing kinetoplastid protozoa. *Eukaryotic Cell* **4**, 516–525.
- Ginger ML (2006) Niche metabolism in parasitic protozoa. *Philosophical Transactions of the Royal Society Series B* **361**, 101–118.
- Ginger ML, Portman N and McKean PG (2008) Swimming with protists: perception, motility and flagellum assembly. *Nature Reviews Microbiology* **6**, 838–850.
- Ginger ML, Collingridge PW, Brown RW, Sproat R, Shaw MK and Gull K (2013) Calmodulin is required for paraflagellar rod assembly and flagellum-cell body attachment in trypanosomes. *Protist* **164**, 528–540.
- Giordani F, Morrison LJ, Rowan TG, de Koning HP and Barrett MP (2016) The animal trypanosomiasis and their chemotherapy: a review. *Parasitology* **143**, 1862–1889.
- Gomaa F, Garcia PA, Delaney J, Girguis PR, Buie CR, Edgcomb VP (2017) Toward establishing model organisms for marine protists: successful transfection protocols for *Parabodo caudatus* (Kinetoplastida: Excavata). *Environmental Microbiology* **19**, 3487–3499.
- Grassé PP (1952) Zooflagelles de position systematique incertaine (Flagellata incertae sedis). In Grassé P (ed.) *Traité de Zoologie. Anatomie Systematique Biologie*. Paris, France: Masson Et. Cie Editeurs, vol. 1, pp. 1011–1014.
- Haanstra JR, González-Marcano EB, Gualdrón-López M and Michels PA (2016) Biogenesis, maintenance, and dynamics of glycosomes in trypanosomatid parasites. *Biochimica Biophysica Acta* **1863**, 1038–1048.
- Hamann E, Gruber-Vodicka H, Kleiner M, Tegetmeyer HE, Riedel D, Littmann S, Chen J, Milucka J, Viehweger B, Becker KW, Dong X, Stairs CW, Hinrichs KU, Brown MW, Roger AJ and Strous M (2016) Environmental Breviatea harbour mutualistic Arcobacter epibionts. *Nature* **534**, 254–258.
- Jaskowska E, Butler C, Preston G and Kelly S (2015) *Phytomonas*: trypanosomatids adapted to plant environments. *PLoS Pathogens* **11**, e1004484.
- Kaufer A, Ellis J, Stark D and Barratt J (2017) The evolution of trypanosomatid taxonomy. *Parasites and Vectors* **10**, 287.
- Keeling PJ (2013) The number, speed, and impact of plastid endosymbioses in eukaryotic evolution. *Annual Reviews in Plant Biology* **64**, 583–607.
- Keeling PJ, McCutcheon JP and Doolittle WF (2015) Symbiosis becoming permanent: survival of the luckiest. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 10101–10103.
- Kořený L, Sobotka R, Kovářová J, Gnypová A, Flegontov P, Horváth A, Oborník M, Ayala FJ and Lukeš J (2012) Aerobic kinetoplastid flagellate *Phytomonas* does not require heme for viability. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 3808–3813.
- Kostygov AY, Butenko A, Nenarokova A, Tashyreva D, Flegontov P, Lukeš J and Yurchenko V (2017) Genome of *Ca. Pandoraea novymonadis*, an endosymbiotic bacterium of the trypanosomatid *Novymonas esmeraldas*. *Frontiers in Microbiology* **8**, 1940.
- Kostygov AY, Dobáková E, Grybchuk-Ieremenko A, Váhala D, Maslov DA, Votýpka J, Lukeš J and Yurchenko V (2016) Novel trypanosomatid-bacterium association: evolution of endosymbiosis in action. *MBio* **7**, e01985.
- Kraeva N, Butenko A, Hlaváčová J, Kostygov A, Myšková J, Grybchuk D, Leštinová T, Votýpka J, Volf P, Opperdoes F, Flegontov P, Lukeš J and Yurchenko V (2015) *Leptomonas seymouri*: adaptations to the dixenous life cycle analyzed by genome sequencing, transcriptome profiling and co-infection with *Leishmania donovani*. *PLoS Pathogens* **11**, e1005127.
- Lai DH, Hashimi H, Lun ZR, Ayala FJ and Lukeš J (2008) Adaptations of *Trypanosoma brucei* to gradual loss of kinetoplast DNA. *Trypanosoma*

- equiperdum* and *Trypanosoma evansi* are petite mutants of *T. brucei*. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 1999–2004.
- Lander N, Li ZH, Niyogi S and Docampo R (2015) CRISPR/Cas9-induced disruption of paraflagellar rod protein 1 and 2 genes in *Trypanosoma cruzi* reveals their role in flagellar attachment. *MBio* **6**, e01012.
- Loyola-Machado AC, Azevedo-Martins AC, Catta-Preta CMC, de Souza W, Galina A and Motta MCM (2017) The symbiotic bacterium fuels the energy metabolism of the host trypanosomatid *Strigomonas culicis*. *Protist* **168**, 253–269.
- Lukeš J, Flegontova O and Horák A (2015) Diplonemids. *Current Biology* **25**, R702–R704.
- Lukeš J, Archibald JM, Keeling PJ, Doolittle WF and Gray MW (2011) How a neutral evolutionary ratchet can build cellular complexity. *IUBMB Life* **63**, 528–537.
- Lukeš J, Guilbride DL, Votýpka J, Zíková A, Benne R and Englund PT (2002) Kinetoplast DNA network: evolution of an improbable structure. *Eukaryotic Cell* **1**, 495–502.
- Lukeš J, Skalický T, Týč J, Votýpka J and Yurchenko V (2014) Evolution of parasitism in kinetoplastid flagellates. *Molecular and Biochemical Parasitology* **195**, 115–122.
- Maga JA, Sherwin T, Francis S, Gull K and LeBowitz JH (1999) Genetic dissection of the *Leishmania* paraflagellar rod, a unique flagellar cytoskeleton structure. *Journal of Cell Science* **112**, 2753–2763.
- Maslov DA, Votýpka J, Yurchenko V and Lukeš J (2013) Diversity and phylogeny of insect trypanosomatids: all that is hidden shall be revealed. *Trends in Parasitology* **29**, 43–52.
- McCutcheon JP (2016) From microbiology to cell biology: when an intracellular bacterium becomes part of its host cell. *Current Opinion in Cell Biology* **41**, 132–136.
- Michels PA, Chevalier N, Opperdoes FR, Rider MH and Rigden DJ (1997) The glycosomal ATP-dependent phosphofructokinase of *Trypanosoma brucei* must have evolved from an ancestral pyrophosphate-dependent enzyme. *European Journal of Biochemistry* **250**, 698–704.
- Milder R, Camargo EP and Freymuller E (1990) Intracytoplasmic flagellum in trypanosomatids. *European Journal of Protistology* **25**, 306–309.
- Morales J, Hashimoto M, Williams TA, Hirawake-Mogi H, Makiuchi T, Tsubouchi A, Kaga N, Taka H, Fujimura T, Koike M, Mita T, Bringaud F, Concepción JL, Hashimoto T, Embley TM and Nara T (2016a) Differential remodelling of peroxisome function underpins the environmental and metabolic adaptability of diplomonads and kinetoplastids. *Proceedings of the Royal Society B, Biological Sciences* **283**, 20160520.
- Morales J, Kokkori S, Weidauer D, Chapman J, Goltsman E, Rokhsar D, Grossman AR and Nowack ECM (2016b) Development of a toolbox to dissect host-endosymbiont interactions and protein trafficking in the trypanosomatid *Angomonas deanei*. *BMC Evolutionary Biology* **16**, 247.
- Morris JC, Wang Z, Drew ME and Englund PT (2002) Glycolysis modulates trypanosome glycoprotein expression as revealed by an RNAi library. *EMBO Journal* **21**, 4429–4438.
- Motta MCM, Catta-Preta CM, Schenkman S, de Azevedo Martins AC, Miranda K, de Souza W and Elias MC (2010) The bacterium endosymbiont of *Crithidia deanei* undergoes coordinated division with the host cell nucleus. *PLoS ONE* **5**, e12415 doi: 10.1371/journal.pone.0012415.
- Motta MCM, Martins AC, de Souza SS, Catta-Preta CM, Silva R, Klein CC, de Almeida LG, de Lima Cunha O, Ciapina LP, Brocchi M, Colabardini AC, de Araujo Lima B, Machado CR, de Almeida Soares CM, Probst CM, de Menezes CB, Thompson CE, Bartholomeu DC, Gradia DF, Pavoni DP, Grisard EC, Fantinatti-Garbozzini F, Marchini FK, Rodrigues-Luiz GF, Wagner G, Goldman GH, Fietto JL, Elias MC, Goldman MH, Sagot MF, Pereira M, Stoco PH, de Mendonça-Neto RP, Teixeira SM, Maciel TE, de Oliveira Mendes TA, Ürményi TP, de Souza W, Schenkman S and de Vasconcelos AT (2013) Predicting the proteins of *Angomonas deanei*, *Strigomonas culicis* and their respective endosymbionts reveals new aspects of the trypanosomatidae family. *PLoS ONE* **8**, e60209 doi: 10.1371/journal.pone.0060209.
- Motta MCM, Monteiro-Leal LH, de Souza W, Almeida DF and Ferreira LCS (1997) Detection of penicillin-binding proteins in endosymbiosis of the trypanosomatid *Crithidia deanei*. *Journal of Eukaryotic Microbiology* **44**, 49–496 doi: 10.1111/j.1550-7408.1997.tb05729.x.
- Mukherjee I, Hodoki Y and Nakano S (2015) Kinetoplastid flagellates overlooked by universal primers dominate in the oxygenated hypolimnion of Lake Biwa, Japan. *FEMS Microbiology Ecology* **91**, fiv083.
- Mundim MH, Roitman I, Hermans MA and Kitajima EW (1974) Simple nutrition of *Crithidia deanei*, a reduviid trypanosomatid with an endosymbiont. *Journal of Protozoology* **21**, 518–521.
- Nawathean P and Maslov DA (2000) The absence of genes for cytochrome c oxidase and reductase subunits in maxicircle kinetoplast DNA of the respiratory deficient plant trypanosomatid *Phytomonas serpens*. *Current Genetics* **38**, 95–103.
- Nes WD, Norton RA, Crumley FG, Madigan SJ and Katz ER (1990) Sterol phylogenies and algal evolution. *Proceedings of the National Academy of Sciences of the United States of America* **87**, 7565–7569.
- Newton BA (1957) Nutritional requirements and biosynthetic capabilities of the parasitic flagellate *Strigomonas oncopelti*. *Journal of General Microbiology* **17**, 708–717.
- Nowack ECM and Grossman AR (2012) Trafficking of protein into the recently established photosynthetic organelles of *Paulinella chromatophora*. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 5340–5345.
- Nowack ECM and Melkonian M (2010) Endosymbiotic associations within protists. *Philosophical Transactions of the Royal Society Series B* **365**, 699–712.
- Nussbaum K, Honek J, Cadmus CM and Efferth T (2010) Trypanosomatid parasites causing neglected diseases. *Current Medicinal Chemistry* **17**, 1594–1617.
- Ogbadoyi EO, Robinson DR and Gull K (2003) A high-order transmembrane structural linkage is responsible for mitochondrial genome positioning and segregation by flagellar basal bodies in trypanosomes. *Molecular Biology of the Cell* **14**, 1769–1779.
- Ong HB, Lee WS, Patterson S, Wyllie S and Fairlamb AH (2015) Homoserine and quorum-sensing acyl homoserine lactones as alternative sources of threonine: a potential role for homoserine kinase in insect-stage *Trypanosoma brucei*. *Molecular Microbiology* **95**, 143–156.
- Opperdoes F, Butenko A, Flegontov P, Yurchenko V and Lukeš J (2016) Comparative metabolism of free-living *Bodo saltans* and parasitic trypanosomatids. *Journal of Eukaryotic Microbiology* **63**, 657–678.
- Pan Y, Birdsey RA, Fang J, Houghton R, Kauppi PE, Kurz WA, Phillips OL, Shvidenko A, Lewis SL, Canadell JG, Ciais P, Jackson RB, Pacala SW, McGuire AD, Piao S, Rautiainen A, Sitch S and Hayes D (2011) A large and persistent carbon sink in the world's forests. *Science* **333**, 988–993.
- Phillips OL and Lewis SL (2014) Evaluating the tropical forest carbon sink. *Global Change Biology* **20**, 2039–2041.
- Podlipaev SA, Sturm NR, Fiala I, Fernandes O, Westenberger SJ, Dollet M, Campbell DA and Lukes J (2004) Diversity of insect trypanosomatids assessed from the spliced leader RNA and 5S rRNA genes and intergenic regions. *Journal of Eukaryotic Microbiology* **51**, 283–290.
- Porcel BM, Denoed F, Opperdoes F, Noel B, Madoui MA, Hammarton TC, Field MC, Da Silva C, Couloux A, Poulain J, Katinka M, Jabbari K, Aury JM, Campbell DA, Cintron R, Dickens NJ, Docampo R, Sturm NR, Koumandou VL, Fabre S, Flegontov P, Lukeš J, Michaeli S, Mottram JC, Szöör B, Zilberstein D, Bringaud F, Wincker P and Dollet M (2014) The streamlined genome of *Phytomonas* spp. relative to human pathogenic kinetoplastids reveals a parasite tailored for plants. *PLoS Genetics* **10**, e1004007.
- Porto-Carreiro I, Attias M, Miranda K, De Souza W and Cunha-e-Silva N (2000) *Trypanosoma cruzi* epimastigote endocytic pathway: cargo enters the cytosome and passes through an early endosomal network before storage in reservosomes. *European Journal of Cell Biology* **79**, 858–869.
- Postell FJ and McGhee RB (1981) An ultrastructural study of *Phytomonas davidi* Lafont (Trypanosomatidae). *Journal of Protistology* **28**, 78–83.
- Priya M, Haridas A and Manilal VB (2008) Anaerobic protozoa and their growth in biomethanation systems. *Biodegradation* **19**, 179–185.
- Pullen TJ, Ginger ML, Gaskell SJ and Gull K (2004) Protein targeting of an unusual, evolutionarily conserved adenylate kinase to a eukaryotic flagellum. *Molecular Biology of the Cell* **15**, 3257–3265.
- Raederstorff D and Rohmer M (1985) Sterol biosynthesis de novo via cycloartenol by the soil amoeba *Acanthamoeba polyphaga*. *Biochemical Journal* **231**, 609–615.
- Read LK, Lukeš J and Hashimi H (2016) Trypanosome RNA editing: the complexity of getting U in and taking U out. *WIREs RNA* **7**, 33–51.
- Roberts CW, McLeod R, Rice DW, Ginger M, Chance ML and Goad LJ (2003) Fatty acid and sterol metabolism: potential antimicrobial targets in apicomplexan and trypanosomatid parasitic protozoa. *Molecular and Biochemical Parasitology* **126**, 129–142.


- Skalický T, Dobáková E, Wheeler RJ, Tesařová M, Flegontov P, Jirsová D, Votýpka J, Yurchenko V, Ayala FJ and Lukeš J (2017) Extensive flagellar remodeling during the complex life cycle of *Paratrypanosoma*, an early-branching trypanosomatid. *Proceedings of the National Academy of Sciences of the United States of America* **114**, 11757–11762.
- Sibbald SJ, Cenci U, Colp M, Eglit Y, O'Lelly CJ and Archibald JM (2017) Diversity and evolution of *Paramoeba* spp. and their kinetoplastid endosymbionts. *Journal of Eukaryotic Microbiology* **64**, 598–607.
- Silva F, Kostygov AY, Spodareva VV, Butenko A, Tossou R, Lukeš J, Yurchenko V and Alves JMP (2018) The reduced genome of *Candidatus Kinetoplastibacterium sorsogonicus*, the endosymbiont of *Kentomonas sorsogonicus* (Trypanosomatidae): loss of the heme synthesis pathway. *Parasitology*. doi: 10.1017/S003118201800046X.
- Simpson AG, Stevens JR and Lukeš J (2006) The evolution and diversity of kinetoplastid flagellates. *Trends in Parasitology* **22**, 168–174.
- Singer A, Poschmann G, Mühlich C, Valadez-Cano C, Hänsch S, Hüren V, Rensing SA, Stühler K and Nowack ECM (2017) Massive protein import into the early-evolutionary-stage photosynthetic organelle of the amoeba *Paulinella chromatophora*. *Current Biology* **27**, 2763–2773.e5.
- Sousa FL, Neukirchen S, Allen JF, Lane N and Martin WF (2016) Lokiarchaeon is hydrogen dependent. *Nature Microbiology* **1**, 16034.
- Tai V, Carpenter KJ, Weber PK, Nalepa CA, Perlman SJ and Keeling PJ (2016) Genome evolution and nitrogen fixation in bacterial ectosymbionts of a protist inhabiting wood-feeding cockroaches. *Applied and Environmental Microbiology* **82**, 4682–4695.
- Tanifuji G, Cenci U, Moog D, Dean S, Nakayama T, David V, Fiala I, Curtis BA, Sibbald SJ, Onodera NT, Colp M, Flegontov P, Johnson-MacKinnon J, McPhee M, Inagaki Y, Hashimoto T, Kelly S, Gull K, Lukeš J and Archibald JM (2017) Genome sequencing reveals metabolic and cellular interdependence in an amoeba-kinetoplastid symbiosis. *Scientific Reports* **7**, 11688.
- Tanifuji G, Kim E, Onodera NT, Gibeault R, Dlutek M, Cawthorn RJ, Fiala I, Lukeš J, Greenwood SJ and Archibald JM (2011) Genomic characterization of *Neoparamoeba pemaquidensis* (Amoebozoa) and its kinetoplastid endosymbiont. *Eukaryotic Cell* **10**, 1143–1146.
- Teixeira MM, Borghesan TC, Ferreira RC, Santos MA, Takata CS, Campaner M, Nunes VL, Milder RV, de Souza W and Camargo EP (2011) Phylogenetic validation of the genera *Angomonas* and *Strigomonas* of trypanosomatids harboring bacterial endosymbionts with the description of new species of trypanosomatids and of proteobacterial symbionts. *Protist* **162**, 503–524.
- Theissen U and Martin W (2006) The difference between endosymbionts and organelles. *Current Biology* **16**, R1016–7.
- Vassella E, Den Abbelee JV, Bütikofer P, Renggli CK, Furger A, Brun R and Roditi I (2000) A major surface glycoprotein of *Trypanosoma brucei* is expressed transiently during development and can be regulated post-transcriptionally by glycerol or hypoxia. *Genes and Development* **14**, 615–626.
- Vickerman K (1977) DNA throughout the single mitochondrion of a kinetoplastid flagellate: observations on the ultrastructure of *Cryptobia vaginalis* (Hesse, 1910). *Journal of Protozoology* **24**, 221–233.
- Vidal JC, Alcantara CL, de Souza W and Cunha-E-Silva NL (2016) Loss of the cytostome-cytopharynx and endocytic ability are late events in *Trypanosoma cruzi* metacyclogenesis. *Journal of Structural Biology* **196**, 319–328.
- von der Heyden S, Chao EE, Vickerman K and Cavalier-Smith T (2004) Ribosomal RNA phylogeny of bodonid and diplomonid flagellates and the evolution of euglenozoan. *Journal of Eukaryotic Microbiology* **51**, 402–416.
- Votýpka J, Kostygov AY, Kraeva N, Grybchuk-Ieremenko A, Tesařová M, Grybchuk D, Lukeš J and Yurchenko V (2014) *Kentomonas* gen. n., a new genus of endosymbiont-containing trypanosomatids of Strigomonadinae subfam. n. *Protist* **165**, 825–838.
- Worden AZ, Follows MJ, Giovannoni SJ, Wilken S, Zimmerman AE and Keeling PJ (2015) Environmental science. Rethinking the marine carbon cycle: factoring in the multifarious lifestyles of microbes. *Science* **347**, 1257594.
- Yazaki E, Ishikawa SA, Kume K, Kumagai A, Kamaishi T, Tanifuji G, Hashimoto T and Inagaki Y (2017) Global Kinetoplastea phylogeny inferred from a large-scale multigene alignment including parasitic species for better understanding transitions from a free-living to a parasitic lifestyle. *Genes Genetics Systems* **92**, 35–42.
- Yubuki N, Simpson AG and Leander BS (2013) Reconstruction of the feeding apparatus in *Postgaardi mariagerensis* provides evidence for character evolution within the Symbiontida (Euglenozoa). *European Journal of Protistology* **49**, 32–39.
- Zachar I and Szathmáry E (2017) Breath-giving cooperation: critical review of origin of mitochondria hypotheses: major unanswered questions point to the importance of early ecology. *Biology Direct* **12**, 19.

RESEARCH ARTICLE

Open Access



Transcriptome, proteome and draft genome of *Euglena gracilis*

ThankGod E. Ebenezer^{1,2}, Martin Zoltner¹, Alana Burrell³, Anna Nenarokova⁴, Anna M. G. Novák Vanclová⁵, Binod Prasad⁶, Petr Soukal⁵, Carlos Santana-Molina⁷, Ellis O'Neill⁸, Nerissa N. Nankisoor⁹, Nithya Vadakedath⁶, Viktor Daiker⁶, Samson Obado¹⁰, Sara Silva-Pereira¹¹, Andrew P. Jackson¹¹, Damien P. Devos⁷, Julius Lukeš⁴, Michael Lebert⁶, Sue Vaughan³, Vladimír Hampl⁵, Mark Carrington², Michael L. Ginger¹², Joel B. Dacks^{9,13*}, Steven Kelly^{8*} and Mark C. Field^{1,4*} 

Abstract

Background: Photosynthetic euglenids are major contributors to fresh water ecosystems. *Euglena gracilis* in particular has noted metabolic flexibility, reflected by an ability to thrive in a range of harsh environments. *E. gracilis* has been a popular model organism and of considerable biotechnological interest, but the absence of a gene catalogue has hampered both basic research and translational efforts.

Results: We report a detailed transcriptome and partial genome for *E. gracilis* Z1. The nuclear genome is estimated to be around 500 Mb in size, and the transcriptome encodes over 36,000 proteins and the genome possesses less than 1% coding sequence. Annotation of coding sequences indicates a highly sophisticated endomembrane system, RNA processing mechanisms and nuclear genome contributions from several photosynthetic lineages. Multiple gene families, including likely signal transduction components, have been massively expanded. Alterations in protein abundance are controlled post-transcriptionally between light and dark conditions, surprisingly similar to trypanosomatids.

Conclusions: Our data provide evidence that a range of photosynthetic eukaryotes contributed to the *Euglena* nuclear genome, evidence in support of the 'shopping bag' hypothesis for plastid acquisition. We also suggest that euglenids possess unique regulatory mechanisms for achieving extreme adaptability, through mechanisms of paralog expansion and gene acquisition.

Keywords: *Euglena gracilis*, Transcriptome, Cellular evolution, Plastid, Horizontal gene transfer, Gene architecture, Splicing, Secondary endosymbiosis, Excavata

Introduction

Euglena gracilis, a photosynthetic flagellate, was first described by van Leeuwenhoek in 1684 [1]. There are over 250 known species in the genus *Euglena*, with around 20 predominantly cosmopolitan, including *E. gracilis* [2–5]. *Euglena* spp. are facultative mixotrophs in aquatic environments [6] and many possess a green secondary plastid derived by endosymbiosis of a chlorophyte algae [7]. Amongst the many unusual features of euglenids are

a proteinaceous cell surface pellicle [8] and an eyespot [9–14]. Euglenids, together with kinetoplastids, diplomonads and symbiotids, form the Euglenozoa subgroup of the Discoba phylum [15]. Kinetoplastids are best known for the *Trypanosoma* and *Leishmania* lineages [15], important unicellular parasites, while diplomonads have been little studied, yet represent one of the most abundant and diverse eukaryotic lineages in the oceans [16].

E. gracilis is thus of importance due to evolutionary history, divergent cellular architecture, complex metabolism and biology, together with considerable potential for biotechnological exploitation [17]. However, the full complexity of euglenid biology remains to be revealed, and the absence of a complete genome sequence or annotated transcriptome has greatly hampered efforts to

* Correspondence: dacks@ualberta.ca; steven.kelly@plants.ox.ac.uk; mfield@mac.com

⁹Division of Infectious Disease, Department of Medicine, University of Alberta, Edmonton, Alberta T6G, Canada

⁸Department of Plant Sciences, University of Oxford, Oxford OX1 3RB, UK

¹School of Life Sciences, University of Dundee, Dundee DD1 5EH, UK

Full list of author information is available at the end of the article



study *E. gracilis* or to develop genetic tools [17, 18]. Two transcriptomes have been published, one derived from cells grown in light and dark conditions plus rich versus minimal media [17] and a second examining the impact of anaerobic conditions on gene expression [19]. For the most part, these studies focused on the biosynthetic properties of *E. gracilis* and not cellular systems or aspects of protein family evolution. Most recently, a study of low molecular weight RNA populations identified over 200 snoRNAs [20].

Comparisons between euglenozoans such as the free-living bodonids, early-branching trypanosomatids (*Paratrypanosoma confusum*), and parasitic forms have uncovered many genetic changes associated with parasitism [21–24]. Both the cell surface and flagellum of euglenoids are of significant importance to life cycle development, interaction with the environment and, for parasitic trypanosomes, pathogenesis and immune evasion [25, 26]. The surface macromolecules of trypanosomatids are highly lineage-specific with roles in life cycle progression [23, 27–31], but it remains to be determined to what extent *E. gracilis* shares surface proteins or other aspects of biology with the trypanosomatids or how cellular features diverge. Such information is invaluable for determining how parasitism arose in the kinetoplastids.

E. gracilis produces a wide range of secondary metabolites, and many of which are of potential commercial value [17]. Furthermore, *E. gracilis* is of considerable promise for biofuel production [32–34], and extremely resistant to conditions such as low pH and high metal ion concentrations, fueling interest as possible sentinel species or bioremediation agents [19, 35–37]. In parts of Asia, *E. gracilis* is cultivated as an important food supplement [38].

E. gracilis possesses a complex genome, with nuclear, plastid and mitochondrial components, an overall architecture known for decades. The coding potential of the mitochondrial genome is surprisingly small [39, 40], while the plastid is of more conventional structure [41]. The plastid is the result of a secondary endosymbiotic event, which is likely one of several such events occurring across eukaryotes [42]. Uncertainties concerning the origins of the plastid have remained, and not least of which has been the presence of genes from both red and green algae in the *E. gracilis* nuclear genome [19, 43]. Such a promiscuous origin for photosynthetic genes is not restricted to the euglenids and has been proposed as a general mechanism, colloquially the ‘shopping bag’ hypothesis, whereby multiple endosymbiotic events are proposed and responsible for the range of genes remaining in the nuclear genome, providing a record of such events and collecting of genes, but where earlier symbionts have been completely lost from the modern host [44].

The *E. gracilis* nuclear genome size has been estimated as in the gigabyte range [45–48] and organization and intron/exon boundaries of very few genes described [49–54]. In the kinetoplastids, unusual transcriptional mechanisms, involving the use of *trans*-splicing as a near universal mechanism for maturation of protein-coding transcripts and polycistronic transcription units, have been well described. As *E. gracilis* supports multiple splicing pathways, including conventional and non-conventional *cis*- [52, 53] and *trans*-splicing [55], there is scope for highly complex mechanisms for controlling expression, transcription and mRNA maturation [56], but how these are related to kinetoplastids is unclear.

We undertook a polyomic analysis of the Z1 strain of *E. gracilis* to provide a platform for improved understanding of the evolution and functional capabilities of euglenids. Using a combination of genome sequencing, together with pre-existing [17] and new RNA-seq analysis, proteomics and expert annotation, we provide an improved view of *E. gracilis* coding potential and gene expression for greater understanding of the biology of this organism.

Results and discussion

Genome sequencing of *Euglena gracilis*

We initiated sequencing of the *E. gracilis* genome using Roche 454 technology. The early assemblies from these data indicated a large genome in excess of 250 Mb and that data coverage was low. We turned to the Illumina platform and generated data from multiple-sized libraries, as well as a full lane of 150 bp paired-end sequences. These data were assembled as described in methods and as previously [48] and latterly supplemented with PacBio data generously donated by colleagues (Purificación López-García, David Moreira and Peter Myler, with thanks). The PacBio data however failed to improve the assembly quality significantly, presumably due to low coverage.

Our final draft genome assembly has 2,066,288 sequences with N_{50} of 955 (Table 1), indicating significant fragmentation. The estimated size of the single-copy proportion of the genome is 140–160 mb and the estimated size of the whole haploid genome is 332–500 mb. This is consistent with several estimates from earlier work (e.g. [57]), albeit based here on molecular sequence data rather than estimates of total DNA content. Using the core eukaryotic genes mapping approach (CEGMA) [58], we estimate that the genome assembly, or at least the coding sequence proportion, is ~20% complete. Hence, this assembly could only support an initial analysis of genome structure and is unable to provide a full or near full open reading frame catalog (Table 2). The heterozygosity, size and frequency of low complexity

Table 1 Statistics of genome assembly

Parameter	
Number of sequences	2,066,288
Median sequence length	457
Mean sequence length	694
Max sequence length	166,587
Min sequence length	106
No. sequence > 1kbp	373,610
No. sequence > 10kbp	1459
No. sequence > 100kbp	2
No. gaps	0
Bases in gaps	0
N50	955
Combined sequence length	1,435,499,417

Following the assembly process, over two million sequences were retained, with a median sequence length of 457 bp

sequence hampered our ability to assemble this dataset (see the “Materials and Methods” section for more details). The size and frequency of low-complexity sequence clearly precluded assembly of our dataset from Illumina reads, and significantly, PacBio data had no significant impact on assembly quality. Due to the large proportion of low-complexity sequence, any estimate for the size of the genome is very much an approximation.

Restricting analysis to contigs > 10 kb, where some features of overall gene architecture could be inferred, we identified several unusual aspects of genome structure (Table 3, Fig. 1, Additional file 1: Figure S1). These contigs encompassed about 22 Mb of sequence, but with

Table 3 Characteristics of contigs assembled with length exceeding 10 kb

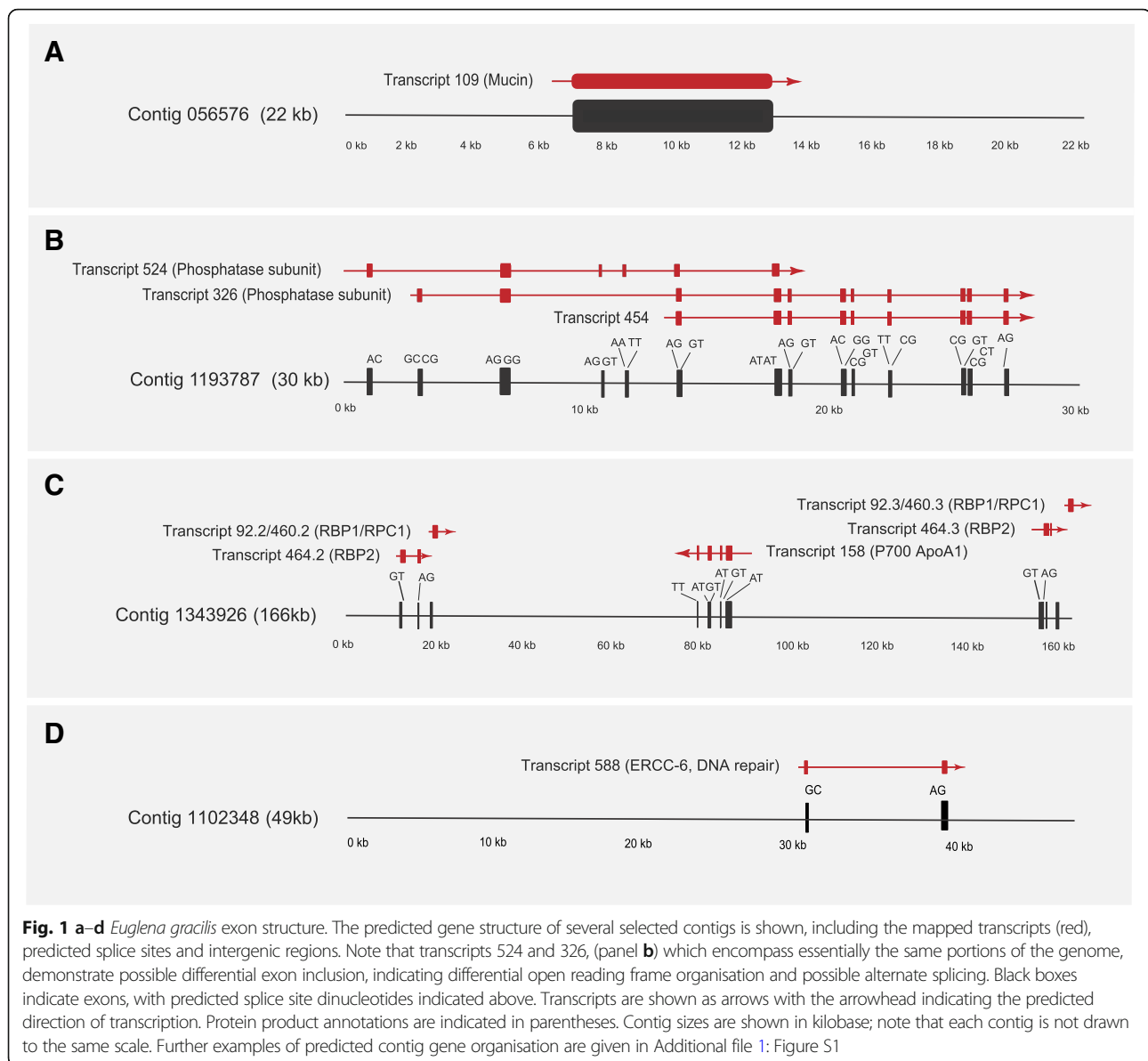
Contigs	Total contigs analysed > 10 kb	1459
	Total nucs in contigs analysed	22 Mb
	Contigs with CDS	53
	Percent contigs with CDS	3.6
CDS	Number analysed	135
	Average length	3790
	Total length	481,369
Exons	Number of exons analysed	421
	Average Length	174.54
	Median Length	112
	Total Length	73,482
	Average per predicted CDS	3.85
Introns	Total introns analysed	271
	Average length	1027.14
	Median length	598
	Total length	278,354
	Introns per predicted CDS	2.01
	Number/percent conventional	218/80.1
	Number/percent intermediate	30/11.1
	Number/percent non-conventional	23/8.5
	Percent nucleotides in CDS (exon)	0

The contigs were ranked by size and those exceeding 10 kbp extracted and analyzed for length, coding sequence, exon structure and other features

Table 2 CEGMA analysis of selected datasets

Assembly	Organism	Gene status	Prots	%Completeness	Total	Average	%Ortho
Genome	<i>E. gracilis</i>	Complete	22	8.87	37	1.68	54.55
		Partial	50	20.16	89	1.78	56
	<i>T. brucei</i>	Complete	196	79.03	259	1.32	24.49
		Partial	205	82.66	282	1.38	28.29
	<i>L. major</i>	Complete	194	78.23	220	1.13	11.34
		Partial	204	82.26	245	1.2	15.69
Transcriptome	<i>E. gracilis</i>	Complete	187	75.4	390	2.09	65.78
		Partial	218	87.9	506	2.32	69.72
	<i>T. brucei</i>	Complete	190	76.61	393	2.07	60
		Partial	205	82.66	448	2.19	63.41
	<i>L. major</i>	Complete	133	53.63	275	2.07	64.66
		Partial	194	78.23	405	2.1	64.43

Comparisons for CEGMA scores between *E. gracilis*, *T. brucei* and *L. major* as an estimate of ‘completeness’ based on 248 CEGs. *Prots* number of 248 ultra-conserved CEGs present in genome, *%Completeness* percentage of 248 ultra-conserved CEGs present, *Total* total number of CEGs present including putative orthologs, *Average* average number of orthologs per CEG, *%Ortho* percentage of detected CEGs that have more than 1 ortholog, *Complete* those predicted proteins in the set of 248 CEGs that when aligned to the HMM for the KOG for that protein family, give an alignment length that is 70% of the protein length. i.e. if CEGMA produces a 100 amino acid protein, and the alignment length to the HMM to which that protein should belong is 110, then we would say that the protein is “complete” (91% aligned), *Partial* those predicted proteins in the 248 sets that are incomplete, but still exceeds a pre-computed minimum alignment score. Keys are as described [58]



only 135 genes predicted based on Exonerate [59], this suggests an extremely low gene density of <1%, similar to that in *Homo sapiens*. In those contigs that possess predicted coding sequence, there was frequently more than one open reading frame (ORF), suggesting gene clusters present within large expanses of non-coding sequence (e.g. Contig11343926, Fig. 1c), but with the caveat that we have sampled a very small proportion of total ORFs (Table 3). It is also possible that some genes were not predicted due to absence of expression under the conditions we used for RNA-seq, though we consider this likely a minor contribution as multiple culturing conditions were included within the final RNA-seq dataset (see below). Most identified genes are predicted to be *cis*-spliced and most introns are conventional, with

a smaller proportion of intermediate and non-conventional splice sites (consistent with [57]). Some introns appear very large compared to the coding sequence contained between them (Contig 1102348, Transcript 588, Fig. 1d). Furthermore, some genes are apparently unspliced (Fig. 1a; Contig 056576, Transcript 109) and there is evidence for alternate splicing (Fig. 1b; Contig 1193787, Transcripts 326, 454 and 524). Evidence for alternate splicing was described earlier [19], but it was based on RNA-seq data without a genomic context, unlike here. The near complete absence of *cis*-splicing from bodonids and trypanosomatids clearly reflects loss post-speciation of these lineages from euglenids and removed a considerable mechanism for generation of proteome diversity [60]. The biological basis for the

extreme genome streamlining in the trypanosomatids versus *Euglena* is unclear.

We also sequenced and assembled an *E. gracilis* transcriptome using a combination of in-house generated sequence and publicly available data [17]. This strategy had the advantage of focusing on coding sequence, as well as including data from multiple environmental conditions (see [17], which used dark, light conditions and rich or minimal media and data from here that used distinct media and also light and dark conditions), to increase the likelihood of capturing transcripts, and represents a third analysis, albeit incorporating raw reads from previous work [17].

Over 32,000 unique coding transcripts were predicted by [17], which compares well with this new assembly and which accounted for 14 Mb of sequence overall. Of these transcripts, approximately 50% were annotatable using UniRef, and over 12,000 were associated with a GO term. In a second report, Yoshida et al. [19], assembled 22 Mb of coding sequence within 26,479 likely unique components, with about 40% having assignable function based on sequence similarity to Swiss-Prot.

The total number of coding sequence nucleotides in our new assembly was >38 Mb, with a mean length of 869 bases and 36,526 unique coding sequences (Table 4). This is a significant improvement over 391 bases reported by [17], and comparable to [19], albeit with a significant increase in total sequence assembled. Transcriptome coverage of ORFs was, as expected, significantly superior to the genome, and CEGMA indicated 87.9% recovery (the *Trypanosoma brucei* genome is 82.66%) (Tables 2 and 4).

We also compared the completeness of our transcriptome with the two published transcriptomes of *E. gracilis* [17, 19]. We used TransDecoder (v2.0.1) [61] to translate nucleotide transcripts to proteins and then excluded duplicated proteins with CD-HIT utility (v4.6) with standard parameters [62]. The final comparison,

made by BUSCO (v2.0.1) [63] with the eukaryotic database, is shown as Additional file 1: Figure S12. Note that all three studies report similar statistics, including concordance in the cohort of BUSCOs not found; these may have failed to be detected or genuinely be absent. Given that 19 BUSCOs were not found in concatenated data (i.e. all three assemblies), with between four to eight missing BUSCOs specific to individual assemblies, it is highly likely that these datasets are robust while also indicating saturation in terms of achieving ‘completeness’, together with possible limitations with BUSCO for divergent species such as *E. gracilis*.

Comparisons between genome and transcriptome assembly sizes confirmed the very small coding component, with genome contigs containing significantly less than 1% coding sequence, despite the total number of *E. gracilis* ORFs (36526) being two to three times greater than *Bodo saltans* (18963), *T. brucei* (9068) or *Naegleria gruberi* (15727) [64–66]. This is in full agreement with earlier estimates of genome versus transcriptome size [17] as well as estimates of the proportion of coding and total genomic sequence discussed above. This is also similar to other large genomes and, specifically, *Homo sapiens*. Blast2GO and InterProScan annotated over 19,000 sequences with GO terms, a proportion similar to previous reports (Additional file 1: Figure S2, [17, 19]).

In addition to the formal analysis and calculation of the numbers of unique sequences, our annotation of the transcriptome adds additional confidence that the dataset is a good resource:

- (i) Most expected metabolic pathways could be reconstructed, with very few exceptions,
- (ii) Major known differences between kinetoplastids and *Euglena* were identified, supporting sampling to a deep level,

Table 4 Assembly statistics for the transcriptome

Transcripts		Coding sequence (CDS)		Proteins	
Number of sequences	72,509	Number of sequences	36,526	Number of proteins	36,526
Median sequence length	540	Median sequence length	765	Median protein length	254
Mean sequence length	869	Mean sequence length	1041	Mean protein length	346
Max sequence length	25,763	Max sequence length	25,218	Max protein length	8406
Min sequence length	202	Min sequence length	297	Min protein length	98
No. sequence > 1kbp	19,765	No. sequence > 1kbp	13,991	No. proteins > 1kaa	1290
No. sequence > 10kbp	25	No. sequence > 10kbp	24	N50	471
No. sequence > 100kbp	0	N50	1413		
No. gaps	0	Combined sequence length	38,030,668		
Bases in gaps	0				
N50	1242				
Combined sequence length	63,050,794				

(iii) For most analyzed protein complexes, all subunits or none were identified, indicating that partial coverage of components is likely rare.

Overall, we conclude that the transcriptome is of sufficient quality for robust annotation and prediction and encompasses more than previous datasets.

Post-transcriptional control of protein expression

Trypanosomatids exploit post-transcriptional mechanisms for control of protein abundance, where essentially all genes are produced from polycistronic transcripts via *trans*-splicing. To improve annotation and investigate gene expression in *E. gracilis*, we conducted comparative proteomic analysis between light and dark-adapted *E. gracilis* but retained in the same media and temperature. Previous work suggested that control of protein abundance may be post-transcriptional [67, 68], but analysis was limited and did not consider the entire proteome, while a separate study identified some changes to mRNA abundance

under low oxygen tension [19]. Under these well-controlled conditions, however, significant changes to the proteome were expected. We confirmed by UV/VIS spectroscopy and SDS-PAGE that photosynthetic pigments were lost following dark adaptation and that ensuing ultrastructural changes, i.e. loss of plastid contents, were as expected (Additional file 1: Figure S3). Total protein extracts were separated by SDS-PAGE with 8661 distinct protein groups (representing peptides mapping to distinct predicted ORFs, but which may not distinguish closely related paralogs) identified. Ratios for 4681 protein groups were quantified (Additional file 2: Table S1) including 384 that were observed in only one state (232 in light and 152 in dark). In parallel, we extracted RNA for RNA-seq analysis; comparing transcript hits with protein groups identified 4287 gene products with robust information for both protein and RNA abundance.

Correlations between changes to transcript and protein abundance were remarkably poor (Fig. 2, Additional file 1: Figure S3, Additional file 2: Table S1), consistent with

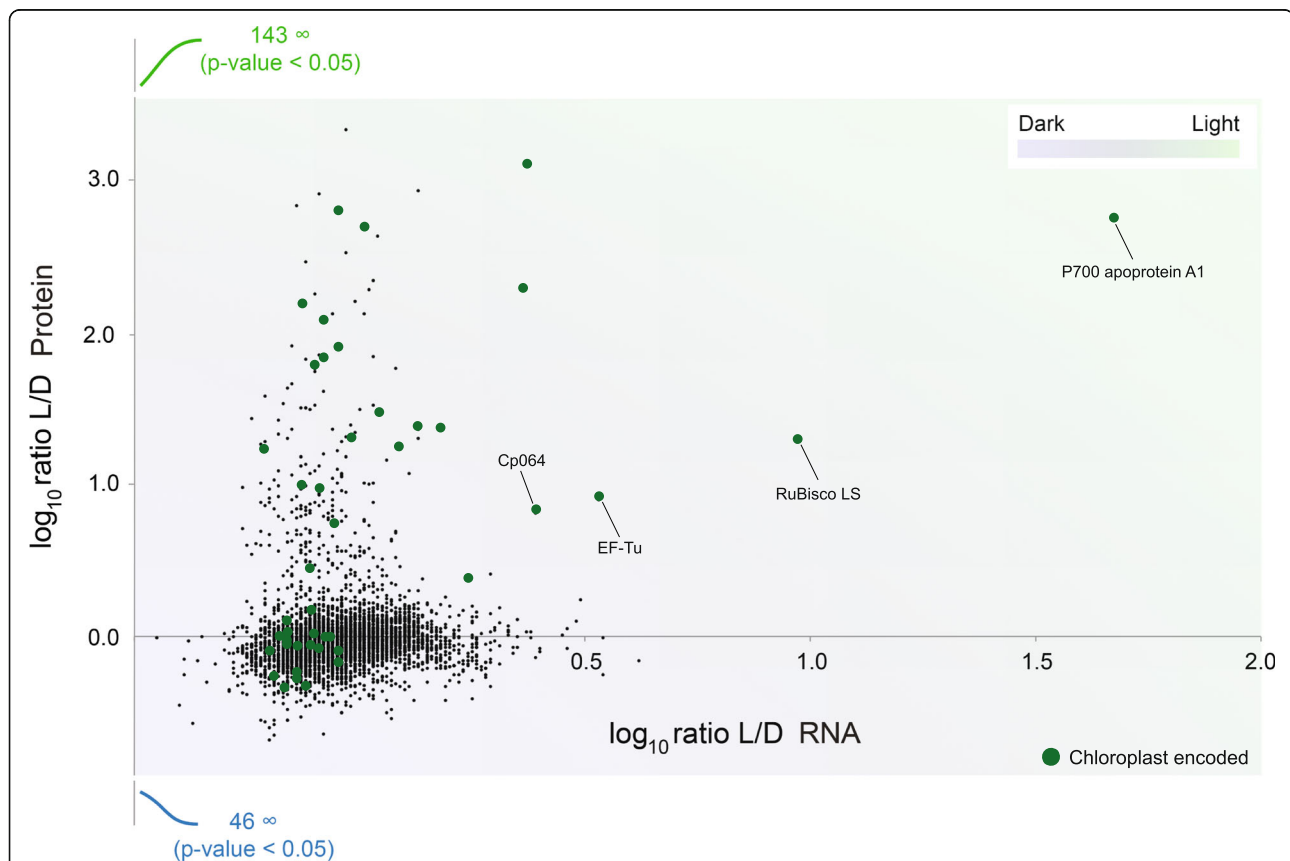


Fig. 2 Expression level changes induced by light are mainly post-transcriptional. Alterations to the transcriptome and proteome in response to ambient light or complete darkness were analysed using RNA-seq and SILAC/LCMS² proteomics respectively. Data are plotted for individual transcripts/polypeptides as the \log_{10} ratio between the two conditions, light (L) and dark (D), with protein on the y-axis and RNA on the x-axis. The presence of a number of proteins that were detected exclusively under one or other condition (hence infinite ratio) are indicated in green (for light) and blue (for dark). With the exception of a few transcripts, which are plastid encoded (green dots), there is little alteration to RNA abundance, but considerable changes to protein levels. Raw data for transcriptome/proteome analysis are provided in Additional file 3

some much smaller earlier studies [67, 68] and broadly with the more extensive study reported in [19]. BLAST analysis revealed that those transcripts where differential abundance did correlate with protein abundance are encoded by the chloroplast genome, including several photosystem I proteins, i.e. P₇₀₀ chlorophyll apoprotein A₁, the large subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO) and chloroplast encoded EF-Tu. Nuclear elongation factors are not influenced by switching growth conditions from dark to light [69], consistent with our finding of no differential expression of nuclear EF-1 α , while both the chloroplast EF-Tu protein and corresponding transcript (EG_transcript_1495) are highly upregulated by light. This absence of transcriptional control for proteome changes between these two conditions is highly similar to that reported for the kinetoplastids, despite the presence of widespread *cis*-splicing and a sparse genome that likely precludes extensive polycistronic transcription. It remains to be determined if this is a general feature for *E. gracilis* or only for certain environmental cues; a cohort of genes are strongly impacted at the RNA level when comparing aerobic to anaerobic transcripts for example, but in that instance none of these transcripts were plastid-encoded nor was a protein analysis performed [19].

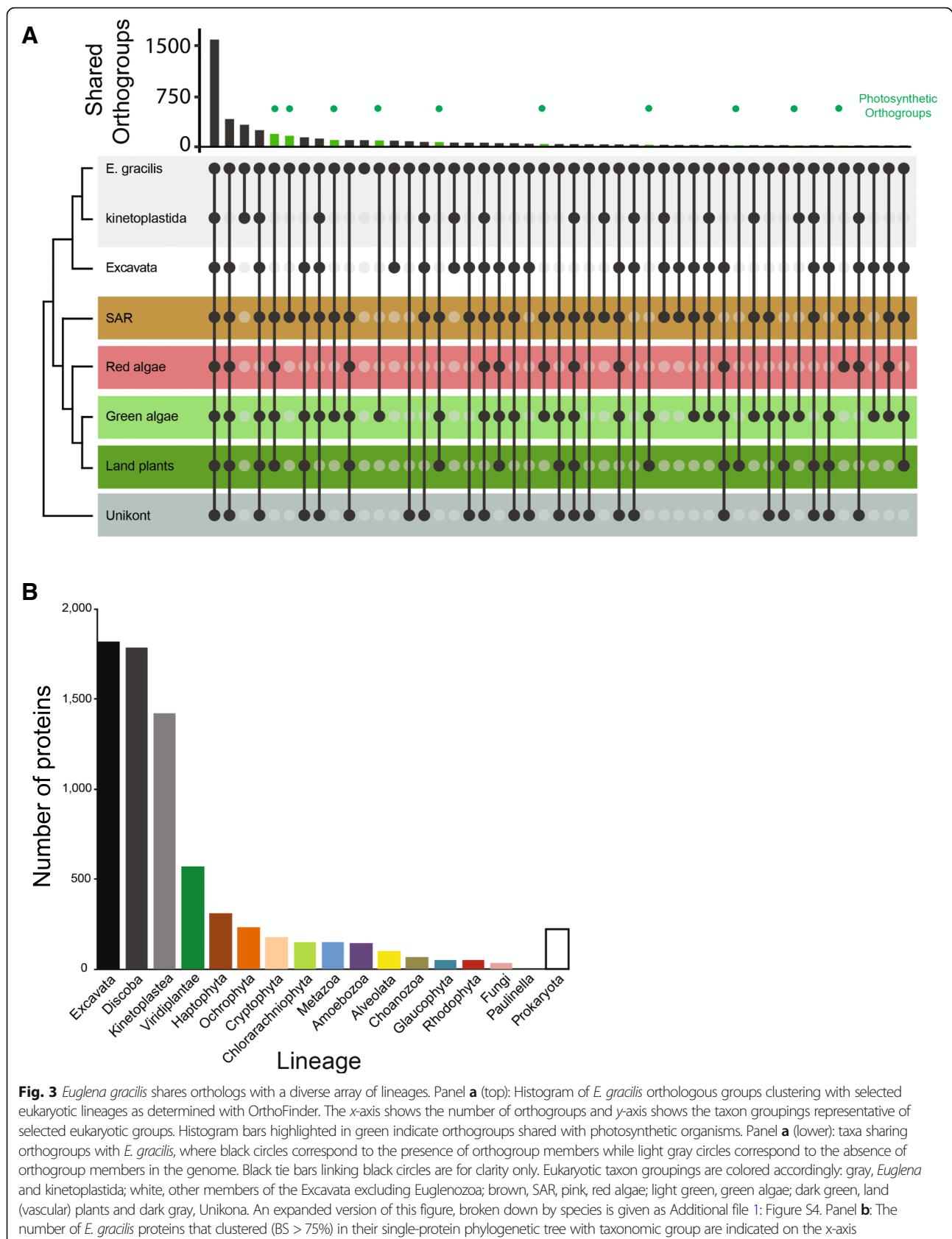
Ancestry of *Euglena gracilis* genes

We used two different approaches to analyze the evolutionary origin of genes predicted from the *E. gracilis* transcriptome. Firstly, we used OrthoFinder [70] to identify *E. gracilis* ortholog gene families shared across eukaryotes and those restricted to specific taxonomic groupings (Fig. 3a, Additional file 1: Figure S4). As expected, the largest proportion was represented by all supergroups and dominated by core metabolic, structural and informational processes, consistent with previous work [19]. A second cohort is shared between *E. gracilis* and other excavates. These classes are broadly within the relative frequencies of previous analyses of excavate genomes [19, 71]. A third cohort represents nuclear transfer of endosymbiotic genes from acquisition of the plastid, and consequently, the genome is a complex mosaic as all eukaryotic genomes also harbour genes driven from the mitochondrial endosymbiont. GO terms associated with orthogroups indicated increased frequency of regulatory function genes in green/secondary plastid orthogroups (Additional file 1: Figure S2). Previous transcriptome studies reported the presence of pan-eukaryotic genes and cohorts shared with kinetoplastids and plants [17, 19], but these were not analyzed in detail, and specifically did not determine which plant taxa were acting as potential gene donors. This is important in terms of understanding the origins of the *Euglena* plastid and where earlier data suggested the

presence of a diverse set of genes from at least green, red and brown algae ([43, 72]). Particularly relevant here is that plastid acquisition in euglenoids is relatively recent [73].

To address this question, we employed a second approach, in which we performed exhaustive analysis to establish phylogenetic ancestry of individual proteins from the predicted *Euglena* proteome by generating single-protein phylogenies. Unlike the analyses of orthogroup sharing, this second approach can be used only for a subset of proteins with a sufficiently robust phylogenetic signal, but also allows determination of the gene ancestry; moreover, this is applicable for members of complex gene families. From all predicted *E. gracilis* proteins only 18,108 formed reliable alignment (> 75 positions) with more than two sequences from our custom database, which comprised 207 taxa in total (Additional file 3 Table S2) and was used for tree construction. In 4087 trees, *E. gracilis* formed a robust (bootstrap support $\geq 75\%$) sister relationship with a taxonomically homogeneous clade (Fig. 3b). Of these, 1816 (44%) were related to one of the lineages of Excavata and 1420 (35%) were related specifically to kinetoplastids. This major fraction represents mostly the vertically inherited component of the genome. The largest non-vertical component forms a group of 572 (14%) proteins related to green plants and green algae, likely representing genes acquired by endosymbiotic gene transfer from the *Euglena* secondary chloroplast, but it should be noted that the direction of transfer cannot be objectively determined. This category is followed by four groups related to the algal groups: haptophytes, cryptophytes, ochrophytes and chlorarachniophytes. While many proteins within the chlorarachniophyte group may represent mis-assigned genes related to green algae, these relatively large numbers related to the three brown-algal groups (723 in total) suggests that these algae contributed considerably to the *E. gracilis* genome and that the process of chloroplast endosymbiosis was complex (see below). On the other hand, the number of proteins related to red algae and glaucophytes (50 and 53) is near negligible. Proteins in groups shared with prokaryotes (220) and non-photosynthetic eukaryotes, e.g. Metazoa (149) and Amoebozoa (145), are most probably the result of horizontal gene transfers, differential gene losses or artifacts caused by biased phylogenetic reconstructions or contaminations in the data sets used to construct the custom database. The robust nature of our analysis, being restricted to phylogenetically well-resolved trees, provides an additional level of confidence to the concept of multiple origins for LGT genes.

It was initially thought that plastid-possessing organisms would overwhelmingly possess nuclear genes derived by transfer from the endosymbiont corresponding



to the plastid currently present, but this has been challenged [74, 75]. While contributions from multiple algal lineages could be explained by incomplete phylogenetic sampling, this is also consistent with the ‘shopping bag’ hypothesis, which proposes an extended process of transient endosymbiosis and gene acquisition by the host prior to the present configuration [44, 75] and which is likely a quite general phenomenon and occurs in many lineages. Our analysis strongly supports the concept of sequential endosymbiotic events.

Expansive paralog families

Several orthogroups consist of an expansive cohort of *E. gracilis* sequences, and a selected few were analyzed phylogenetically and annotated for protein architectural/domain features (Additional file 1: Figure S5, Additional file 4: Table S3). Firstly, highly significant in terms of size and evolutionary history is a family of nucleotidylcyclase III (NCIII)-domain-containing proteins widely distributed across eukaryotes. In African trypanosomes, adenylate cyclases are mediators of immune modulation in the mammalian host [71]. One nucleotidylcyclase subfamily is restricted to kinetoplastids and organisms with secondary plastids and contains photosensor adenylate cyclases [12] that possess one or two BLUF domains (blue light sensor) with a double NCIII domain (Fig. 4). These nucleotidylcyclases are phylogenetically similar to the NCIII-family of *N. gruberi* [66]. A second subfamily is pan-eukaryotic and possesses one NCIII domain and several *trans*-membrane domains, a HAMP (histidine kinases, adenylate cyclases, methyl-accepting proteins and phosphatases) domain as well as cache 1 (calcium channel and chemotaxis receptor) domains. These domains are associated with proteins involved, as their name implies, in signal transduction, particularly chemotaxis [76, 77]. Again, this subfamily is closely related to NCIII-family genes from *N. gruberi*. The third subfamily represents a kinetoplastid cluster with *trans*-membrane proteins and frequently also HAMP and cache1 domains. This complexity indicates considerable flexibility in nucleotidylcyclase evolution and that many lineage-specific paralogs have arisen, with implications for signal transduction, suggesting an extensive regulatory and sensory capacity in *E. gracilis*.

A second example is a large protein kinase C-domain containing a group of protein kinases, which also exhibit extensive lineage-specific expansions in *E. gracilis* (several orthogroups contained a very large number of *E. gracilis* sequences, and a few selected were analysed phylogenetically and annotated for architecture (Additional file 1: Figure S5)). A third orthogroup possess a signal receiver domain (REC) with clear lineage-specific *E. gracilis* paralogs present (Additional file 1: Figure S5). The *E. gracilis* members possess an H-ATPase domain, which is distinct from the Per-Arnt-Sim (PAS) domain

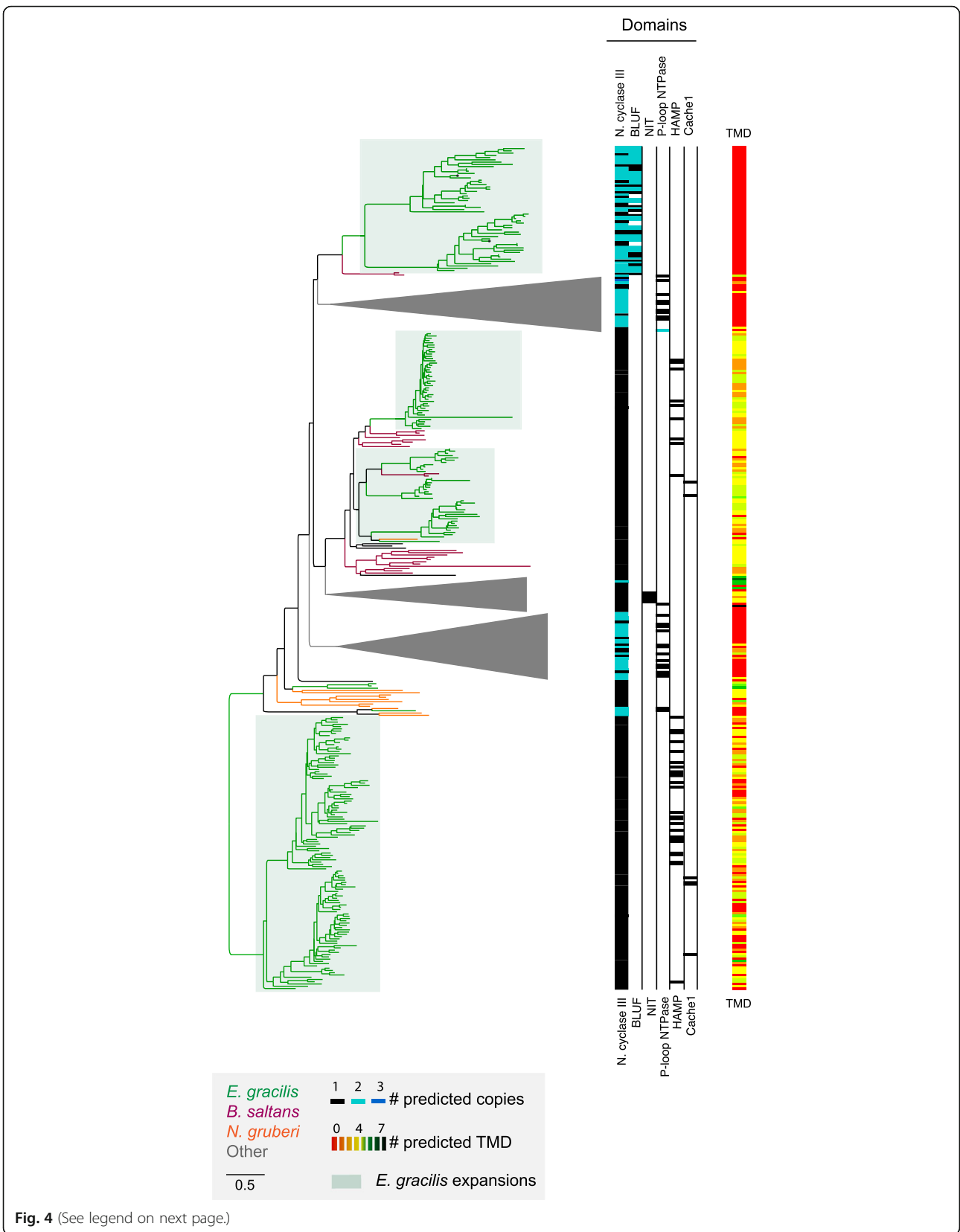
present in many orthologs from other lineages. The presence of independently expanded signaling protein families in *E. gracilis* suggests both highly complex and divergent pathways. These very large families likely partly explain the expanded coding potential in *E. gracilis*, as well as provide some indication of how sensing and adaptation to diverse environments is achieved.

Conservation and divergence of systems between *E. gracilis* and kinetoplastids

To better understand the evolution of *Euglena* and its relationship to free living and parasitic relatives, we selected multiple cellular systems for detailed annotation. These were selected based on documented divergence between kinetoplastids and other eukaryotic lineages and encompass features of metabolism, the cytoskeleton, the endomembrane system and others (Additional file 5: Table S4). Additional annotations of systems not discussed here are available in Additional file 5: Table S4 and provided in Additional file 6: Supplementary analysis.

A unique feature of energy metabolism in kinetoplastids is compartmentalisation of several glycolytic enzymes within peroxisome-derived glycosomes and the presence of additional enzymes for metabolism of the glycolytic intermediate phospho-enolpyruvate to succinate [78]. Glycosomes have been recently reported in diplomonads, the second major euglenozoan group, suggesting an origin predating kinetoplastida [79]. Using 159 query protein sequences for experimentally supported glycosomal *T. brucei* proteins [80], we found candidate orthologs for the majority, but based on the absence of detectable PTS-1 or PTS-2 targeting signals, no evidence that enzymes linked to carbohydrate metabolism are (glyco)peroxisomal. Of the 159 queries, 49 are annotated as hypothetical or trypanosomatid-specific and none had a detectable ortholog in *E. gracilis* (Additional file 5: Table S4). Collectively, this suggests that peroxisomes in *E. gracilis* most likely function in diverse aspects of lipid metabolism rather than glycolysis or other aspects of carbohydrate metabolism and distinct from kinetoplastids.

The surface membrane of *E. gracilis* is in close association with a microtubule corset, and with some structural similarity to the subpellicular array of trypanosomatids, but with very unique architecture [81]. While the plasma membrane composition of kinetoplastids is lineage-specific, in terms of many major surface proteins and a major contributor to host-parasite interactions [82], transporters and some additional surface protein families are more conserved. To compare with *E. gracilis*, we predicted membrane proteins using the signal peptide together with orthogroup clustering, which will encompass both surface and endomembrane compartment constituents. Many genes have significant similarity to



(See figure on previous page.)

Fig. 4 Large paralog gene families are present in the *Euglena gracilis* genome. Several orthogroups contain many *E. gracilis* paralogs. The phylogenetic distribution of one large orthogroup, the nucleotidylcyclase III domain-containing proteins, is shown. Lineage groupings are colour coded: gray, all eukaryotes (and collapsed for clarity); red, *N. gruberi*; amber, *B. saltans*; and green, *E. gracilis*. Clades containing only *Euglena* sequences are boxed in green. Each sequence has been assigned a domain composition (colour gradient black to teal to blue), number of predicted trans-membrane domains (colour coded red to orange to black gradient). To obtain this phylogenetic tree, sequences with likely low coverage (less than 30% of the length of the overall alignment) were removed during alignment to avoid conflicting homology or artefact generation. Domain compositions identified are nucleotidylcyclase III, BLUF, NIT, P-loopNTPase, HAMP and Cache1

kinetoplastids (1103), *B. saltans* (32) or non-kinetoplastida (487) (Additional file 7: Table S5). About 698 proteins with a signal peptide appear to be *E. gracilis* specific, and most of these are a single copy (87.5%), while there are clear large families that possess conserved features (see above). Notably, we were unable to identify a rhodopsin homolog, in contrast to several biochemical analyses suggesting the presence of retinal, the rhodopsin cofactor, which has been interpreted as evidence for a rhodopsin-like light sensor. It remains possible that the euglenid rhodopsin was not represented in the transcriptome or is too divergent to detect [83].

In common with *B. saltans*, *E. gracilis* has a distinct class of amastin, a major kinetoplastid surface protein and which arose from a single ancestor shared with the last euglenozoan common ancestor (Additional file 1: Figure S6). *E. gracilis* also possesses enzymes for the synthesis of lipophosphoglycan (LPG), a glycoconjugate first described in *Leishmania* and implicated in defense and disease mechanisms, together with the pathways for synthesis of GPI protein anchors and free lipids. These data suggest that LPG predates the evolution of parasitism and that the ancestral role was possibly more general, for example, a defense against proteases or predation, or in cell-cell/cell-substrate interactions. Significantly, gp63, a major surface protein present in the vast majority of eukaryotes and also involved in *Leishmania* pathogenesis, is absent and represents a secondary loss following separation from the kinetoplastid lineage.

The endomembrane system is responsible for biosynthesis, degradation and targeting of proteins and lipids and can be considered as a proxy for intracellular complexity. Compartments and transport routes can be predicted with accuracy based on the presence of genes encoding proteins mediating these routes. Using such an analysis, it has been predicted that the complexity of endomembrane compartments in trypanosomatids is decreased compared with free-living bodonids [23, 84]. *E. gracilis* possesses a relatively complete set of membrane-trafficking proteins, extending this trend further (Additional file 1: Figure S7). Two key adaptin family complexes involved in vesicle coat formation and post-Golgi transport, AP5 and TSET, are absent from kinetoplastids, and while AP5 is also absent from *E. gracilis*, a near complete TSET is present. Significantly,

endosomal pathways are predicted as more complex than kinetoplastids, with multiple Rab7 (late endosome/lysosome) and Rab11 (recycling endosome) paralogs, together with ER-associated paralogs for Rab1 (early anterograde transport) and Rab32, respectively. Rab32 may also be associated with the contractile vacuole, an endolysosomal organelle responsible for osmoregulation in many freshwater protists, but these aspects of *E. gracilis* biology remain to be explored.

In kinetoplastids, an unusual cytoskeletal element, the bilobe, plays a central role in Golgi, flagellar pocket collar and flagellum attachment zone biogenesis [74]. All of the structural proteins (MORN1, RRP1, BILBO1, Centrin-2 and Centrin-4) were found [85–90] (Additional file 5: Table S4). Therefore, the potential for the synthesis of a bilobe-like structure in *E. gracilis* is supported, although clearly experimental evidence is needed for the presence of such a structure, but which suggests an origin predating the kinetoplastids.

The considerable size of the *E. gracilis* genome and complex splicing patterns suggests the presence of sophisticated mechanisms for organizing chromatin, mRNA processing and transcription [53, 57]. Furthermore, the *E. gracilis* nucleus has somewhat unusual heterochromatin morphology, with electron-dense regions appearing as numerous foci throughout the nucleoplasm (Additional file 1: Figure S8). Nucleoskeletal proteins related to lamins, NMCPs of plants or kinetoplastid-specific NUP-1/2 are all absent from *E. gracilis*, suggesting that anchoring of chromatin to the nuclear envelope exploits a distinct mechanism [91]. Further, while much of the nuclear pore complex (NPC) is well conserved across most lineages, orthologs for DBP5 and Gle1, two proteins involved in mRNA export in mammalian, yeast and plant NPCs, but absent from trypanosomes, are present. This is consistent with an earlier proposal that the absence of DBP5/Gle1 is connected to the loss of *cis*-splicing in kinetoplastids, but indicates that this is not due to the presence of *trans*-splicing per se as this is common to *E. gracilis* and the kinetoplastids [92]. Finally, kinetochores, required for engagement of chromosomes with the mitotic spindle, are also highly divergent in trypanosomes (Additional file 1: Figure S8) [93, 94]. Of the trypanosomatid kinetochore proteins, only KKT19 and KKT10 are obviously present in *E. gracilis*; as these are a kinase and phosphatase,

respectively, they may not be bona fide kinetochore proteins in *E. gracilis*. Further, very few canonical kinetochore proteins were found, suggesting possible divergence from both higher eukaryote and trypanosome configurations. Overall, these observations suggest unique mechanisms operate in the *E. gracilis* nucleus, which may reflect transitions between conventional kinetochores, lamins and nuclear pores into the more radical configuration present in kinetoplastids. Additional systems are discussed in supplementary material (Additional file 6).

The *Euglena* mitochondrion

In kinetoplastids, unique mitochondrial genome structures are present [95]. Typically, kinetoplastid mitochondrial genomes comprise ~40 copies of a maxicircle encoding several mitochondrial proteins and several thousand minicircles encoding guide RNAs for editing maxicircle transcripts [40, 95]. In trypanosomatids, this structure is attached to the flagellum basal body via a complex cytoskeletal element, the tri-partite attachment complex (TAC) [95]. We find no evidence for RNA editing in *E. gracilis*, nor for the TAC, both of which are consistent with the presence of a mitochondrial genome composed of only short linear DNA molecules and a conventional mitochondrial mRNA transcription system [39]. Specifically, only 16 of 51 proteins involved in RNA editing in *T. brucei* [96] had reciprocal best BLAST hits, and only one predicted protein contained a mitochondrial targeting signal. No homologs to TAC proteins were found (Additional file 5: Table S4).

The *E. gracilis* mitochondrial proteome is predicted to exceed 1000 proteins and encompasses 16 functional categories (Additional file 1: Figure S9A). The kinetoplastid mitochondrion possesses a non-canonical outer mitochondrial membrane translocase (A)TOM (archaic translocase of the outer membrane). The major component is (A)TOM40, a conserved beta-barrel protein that forms the conducting pore, but which is highly diverged in kinetoplastids [97–99]. We identified homologs of two specific receptor subunits of (A)TOM, namely ATOM46 and ATOM69 [100], and two TOM40-like proteins; both these latter are highly divergent and could not be assigned unequivocally as TOM40 orthologs.

We also identified canonical subunits of respiratory chain complexes I–V and 27 homologs of kinetoplastid-specific proteins, together with the widely represented alternative oxidase, consistent with earlier work [101]. Moreover, an ortholog of *T. brucei* alternative type II NADH dehydrogenase (NDH2) was detected. We found only 38 of 133 canonical and only three of 56 kinetoplastid-specific mitoribosomal proteins, which suggests considerable divergence. Hence, the *E. gracilis* mitochondrion has unique features, representing an intermediate between the mitochondria familiar from

yeast or mammals and the atypical organelle present in kinetoplastids (Fig. 5).

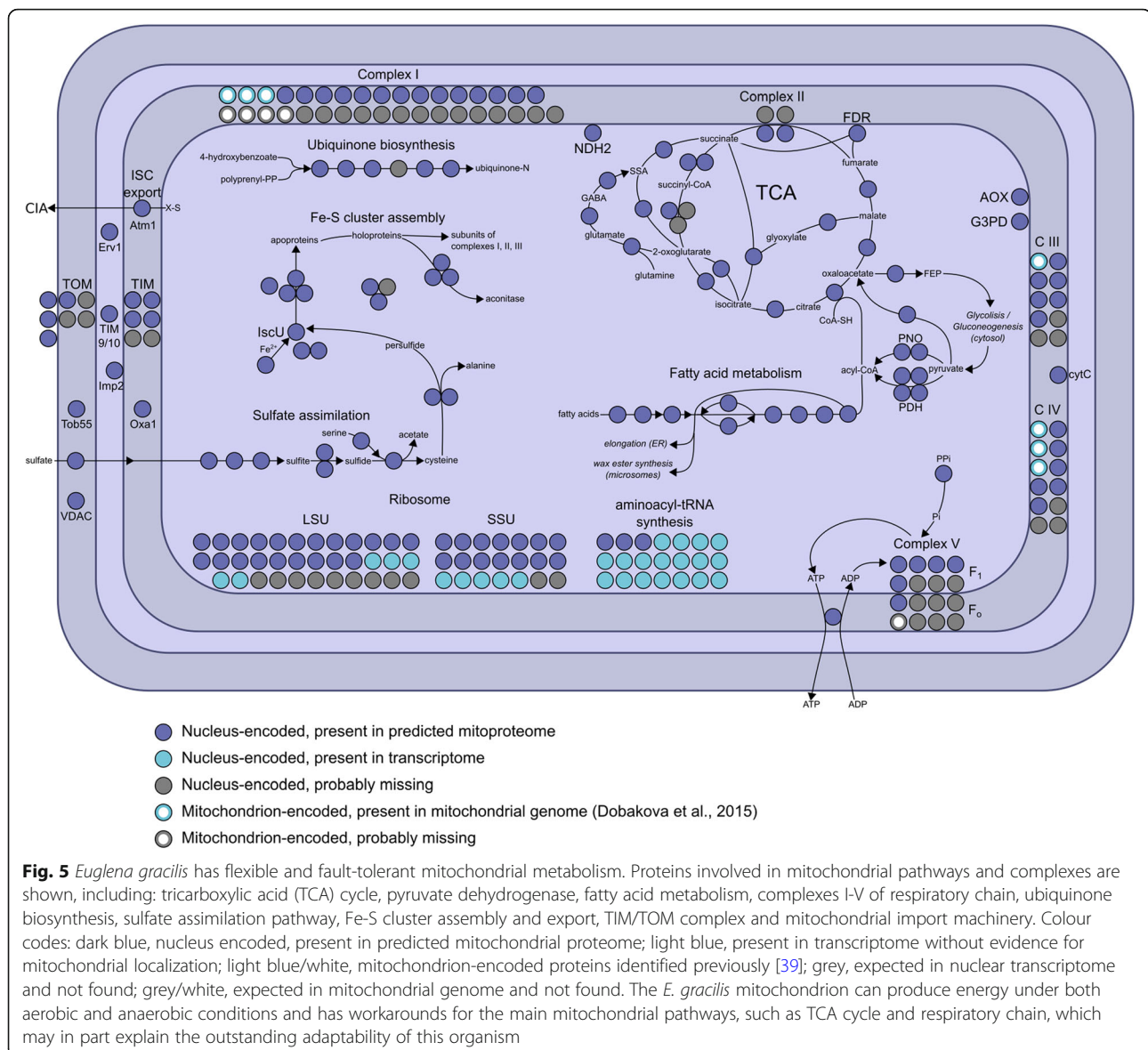
The *Euglena* plastid

The *Euglena* chloroplast, as a secondary acquisition, represents a near unique configuration for studying fundamental aspects of organelle origins and evolution. The predicted *E. gracilis* plastid proteome contains 1902 proteins (Fig. 6, Additional file 1: Figure S9B; Additional file 8: Table S6). Typical plastid metabolic pathways and enzymes are present, including 70 proteins involved in the chloroplast electron transport chain and light harvesting antennae. A few expected genes were absent, such as glycolytic glucose-6-phosphate isomerase and carotenoid synthesis 15-*cis*-phytoene desaturase; as both pathways are known to be present, these likely arise from incomplete sequence data [41]. The C₅ tetrapyrrole pathway was completely reconstructed, while the C₄ pathway for aminolevulinic acid synthesis is absent, consistent with previous findings [102]. Enzymes connecting the cytosolic/mitochondrial mevalonate and plastid methyl-D-erythritol pathway (MEP/DOXP) pathways of terpenoid synthesis were not found, in accordance with separate plastid and cytosolic pools of geranylgeranyl pyrophosphate. Carotenoid and non-plastid isoprenoid (e.g. sterols, dolichols) biosynthetic pathways appear unconnected [103]. Significantly, over 50% of the predicted plastid proteome represent proteins with no homology in the databases, suggesting considerable novel metabolic potential.

Protein targeting to the *E. gracilis* plastid involves trafficking via the Golgi complex. Since the plastid was newly established in the euglenoid lineage, this implies that at least two novel membrane-trafficking pathways should be present, one anterograde *trans*-Golgi to plastid and a retrograde pathway operating in reverse. The relevant machinery for such pathways could be produced via either gene transfer from the green algal host or duplication of host membrane-trafficking machinery. We found no reliable evidence for contributions to the endomembrane protein complement by endosymbiotic gene transfer, but there are extensive gene duplications within the endomembrane machinery. Specifically, additional paralogs of key factors involved in post-Golgi to endosome transport, e.g. AP1 and Rab14, are present, as are expansions in retromer and syntaxin16 that specifically serve to retrieve material from endosomes to the *trans*-Golgi network. Overall, we suggest both a period of kleptoplasty prior to stable establishment of the secondary green plastid and a model whereby novel transport pathways were established by gene duplication, as proposed by the organelle paralogy hypothesis [44].

Conclusions

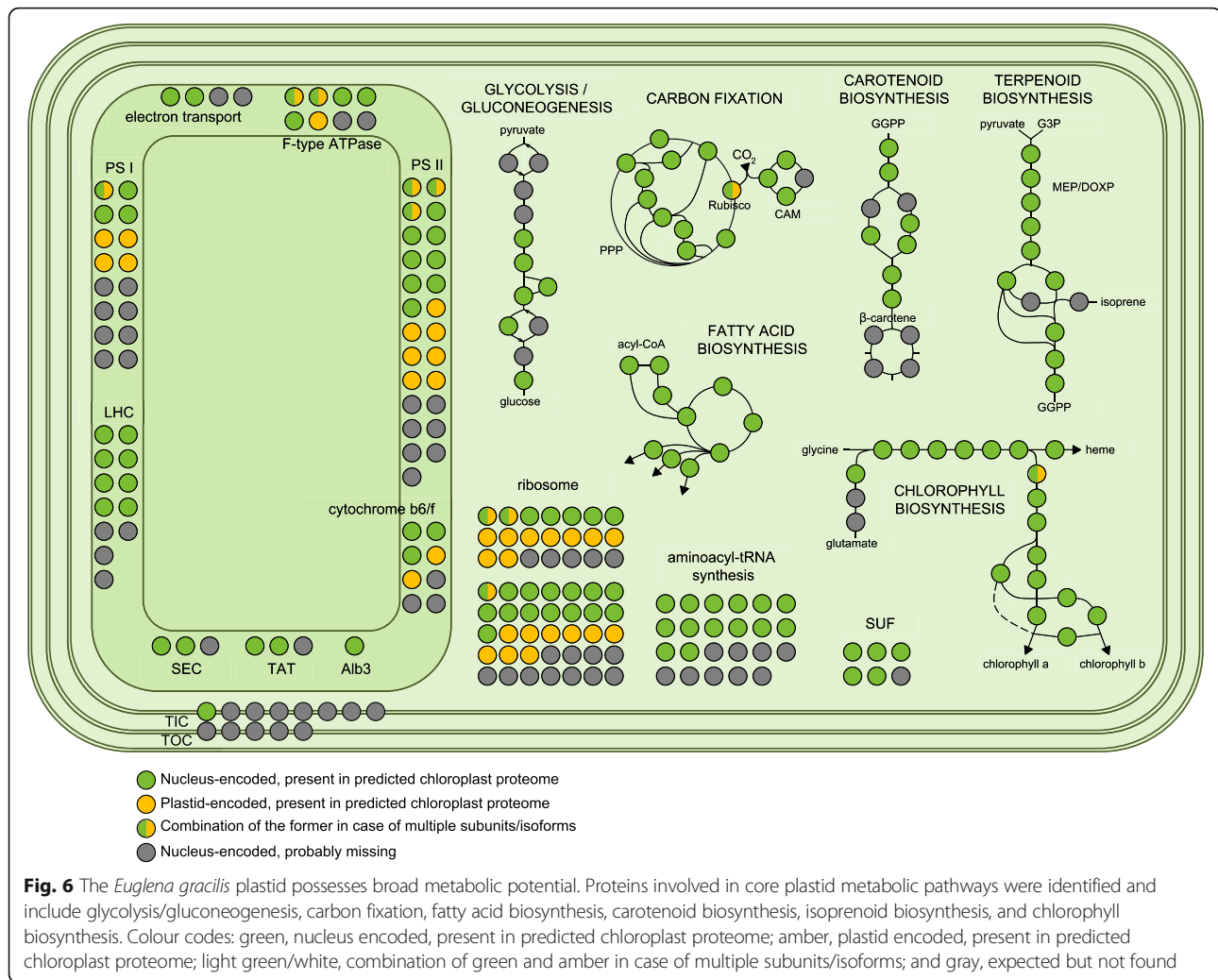
We present here a detailed analysis of the protein-coding complement of *E. gracilis*, together with insights into genome organization. The genome is very



large for a unicellular organism, consistent with many earlier estimates and has exceptionally low coding content, similar to large metazoan genomes. BUSCO, CEGMA and also annotation of many metabolic pathways, complexes and systems indicate that both our data and that from previous work attained very high coverage of the transcriptome. Significant concatenation of all three datasets resulted in essentially negligible improvement to BUSCO scores, suggesting that the data approach a complete sampling.

We predict a highly divergent surface proteome with expanded signal transduction capabilities likely present at the plasma membrane. *E. gracilis* possesses machinery for synthesis of lipophosphoglycan, suggesting the presence of a defensive phosphoglycan sheath [104]. Significantly, we find evidence for gradual loss of conventional

kinetochores, *cis*-splicing and complex RNA processing at the NPC during Euglenozoa evolution. Unexpectedly, there is little evidence for transcriptional control, highly similar to kinetoplastids. Reliance on post-transcriptional processes has been recognized as a feature of *E. gracilis* [105] with mounting evidence that translational and degradative processes are crucial determinants of protein abundance and in agreement with this work [106]. An extensive endomembrane system indicates complex internal organization and multiple endosomal routes representing mechanisms for the sorting, uptake and digestion of material from a range of sources. We also find evidence for novel trafficking pathways between the endomembrane system and the chloroplast; this, together with analysis of the nuclear genome and likely origins of many genes, provides insights into the processes by which secondary



plastids become enslaved, and is consistent with a protracted period of plastid acquisition.

Materials and methods

Cultivation

E. gracilis strain Z1 was provided by William Martin (Düsseldorf). Cells were cultivated at ambient temperature under continuous illumination from a 60-W tungsten filament bulb at 20 cm from the culture vessel, in Hutner's media [107]. Cells were collected in exponential growth phase at $\sim 9 \times 10^5$ cells/ml, measured using a haemocytometer. For light and dark adaptation, cells were adapted to Hutner heterotrophic medium [107] for 16 days prior to the initiation of a light or dark growth period. Cultures were subcultured and dark-adapted cultures transferred to a light proof box adjacent to the light cultures. Subculturing was done under low light conditions periodically and cultures maintained for up to 2 weeks prior to harvesting. The impact of a prolonged period under dark conditions was assessed by microscopy (Zeiss LSM 700 confocal

microscope; $\times 40$ Plan-Neofuar NA1.3 lens under phase contrast, by UV/VIS spectroscopy using a Shimadzu UV-2450, wavelength scan of 190–800 nm and SDS-PAGE).

Isolation of RNA and proteins for gene expression studies

Equivalent numbers (1×10^7 cells) of dark or light cultured cells were harvested by centrifugation at 25 °C, 1000g for 10 mins. RNA extraction was performed using the Qiagen RNeasy Mini Kit (Cat. No. 74104). Genomic DNA contamination was eliminated by performing on-column DNase digestion. Extracted RNA was preserved at -80 °C for RNA sequencing. For proteomics, cells were washed with PBS containing complete protease inhibitors (Roche), extracted with NuPAGE sample buffer (3X), sonicated and lysates containing 1×10^7 cells fractionated on a NuPAGE Bis-Tris 4–12% gradient polyacrylamide gel (Thermo Scientific, Waltham, MA, USA) under reducing conditions. The sample lane was

divided into eight slices that were subjected to tryptic digestion and reductive alkylation.

Proteomics analysis for gene expression studies

Liquid chromatography tandem mass spectrometry (LC-MS²) was performed in house at the University of Dundee, UK. Samples were analyzed on a Dionex UltiMate 3000 RSLCnano System (Thermo Scientific, Waltham, MA, USA) coupled to an Orbitrap Q-exactive mass spectrometer (Thermo Scientific) at the University of Dundee proteomics facility. Protein mass spectra were analyzed using MaxQuant version 1.5 [108] searching the predicted *E. gracilis* proteome from the de novo transcriptome assembly reported here. Minimum peptide length was set at six amino acids, isoleucine and leucine were considered indistinguishable and false discovery rates (FDR) of 0.01 were calculated at the levels of peptides, proteins and modification sites based on the number of hits against the reversed sequence database. Ratios were calculated from label-free quantification intensities using only peptides that could be uniquely mapped to a given protein. If the identified peptide sequence set of one protein contained the peptide set of another protein, these two proteins were assigned to the same protein group. *P* values were calculated applying *t* test-based statistics using Perseus [109]. There were 8661 distinct protein groups identified by MaxQuant analysis. For further analyses, data were reduced to 4297 protein groups by rejecting those groups not identified at the peptide level in each of the three replicates for one state. Additionally, a cohort of 384 protein groups was extracted that were observed in only one state (232 light and 152 dark).

Ultrastructure of *E. gracilis* cells in light and dark conditions

Two populations of *E. gracilis* cells cultured in either light or dark conditions were initially fixed using 2.5% glutaraldehyde and 2% paraformaldehyde in 0.1 M sodium cacodylate buffer pH 7.2. Both samples were post-fixed for an hour in buffered 1% (*w/v*) OsO₄ and embedded in molten agarose prior to incubating overnight in 2% (*w/v*) uranyl acetate. Agarose pellets were dehydrated through a graded acetone series and slowly embedded in Low Viscosity resin (TAAB Ltd.) over 4 days. Following polymerization, 70–90-nm-thin sections were cut by ultramicrotome, post-stained using 2% (*w/v*) uranyl acetate and Reynolds lead citrate [110] and imaged with a Hitachi H-7650 transmission electron microscope. Image resolution varied between 20 and 0.3 nm per pixel, depending on the magnification.

Transcriptome analysis for gene expression studies

Extracted RNA was sequenced at the Beijing Genomics Institute (<https://www.bgi.com/global/>). Analysis and comparisons of the data were performed using standard

pipelines. An estimated 62 M clean reads were generated which were subject to quality filtering using Trimmomatic [111], to remove low-quality bases and read pairs as well as contaminating adaptor sequences, prior to assembly. Sequences were searched for all common Illumina adaptors and settings for read processing by Trimmomatic were LEADING:10 TRAILING:10 SLIDINGWINDOW:5:15 MINLEN:50. The trimmed filtered reads were then used to quantify the de novo-assembled transcriptome using Salmon [112] with the bias-correction option operating. Expected counts were integerised before being subject to differential expression testing using DESeq2 [113] using default parameters. In the transcriptomics analysis, 66,542 distinct sequence classes were detected and the data was reduced to 41,045 applying the same rejection criteria as the proteome (minimum three replicates).

Nucleic acid isolation and purification for genomic and transcriptomic studies

E. gracilis genomic DNA was isolated using the Qiagen DNA purification system to obtain low and high molecular weight DNA for Illumina paired-end and mate-pair read libraries (100-bp paired-end libraries with insert sizes of 170 bp, 500 bp and 800 bp, and mate-pair libraries with insert sizes of 2 kbp, 5 kbp and 40 kbp). For the shorter length libraries (≤ 5 kbp), cells were harvested by centrifugation for 10 mins at 1000 g and DNA extracted using the Qiagen DNAeasy blood and tissue kit (Qiagen Inc., Cat.No. 69504). The cultured animal cell protocol was modified and involved firstly, using 1×10^7 cells, and secondly, prior to adding Buffer AL, 200 μ l of RNase A was added to eliminate RNA contamination. Immediately after the washing step with Buffer AW2, centrifugation was performed for 1 min at 20,000g to eliminate traces of ethanol. To obtain high molecular weight DNA fragments for the ≥ 40 kb insert size library, the Qiagen Genomic-DNA isolation kit (blood and cell culture DNA kit - Maxi, Cat. No. 13362) was used. In this case, 1×10^8 cells were harvested. Prior to adding Buffer C1, samples were ground in liquid nitrogen using a planetary ball mill (Retsch) [114] at 300 rpm for 3 min (the grinding was limited to two cycles to minimize DNA shearing). Four wash steps were performed to remove contaminants including traces of RNA. To determine molecular weight, 400 ng of DNA was loaded onto a 0.45% agarose gel in TAE buffer, stained with Thermo Scientific 6X Orange Loading Dye, and electrophoresed at 80 V for 2 h. A NanoDrop spectrophotometer (DeNovix DS-11+) was used to determine concentration and purity. Total RNA from *E. gracilis* was isolated using the Qiagen RNeasy Mini kit (Cat. No. 74104), and the protocol for the purification of total RNA from animal cells using spin technology was employed as above.

Library preparation and sequencing for genomic and transcriptomic studies

Genome and transcriptome library preparation and sequencing were performed at the Beijing Genomic Institute, using Illumina Genome Analyzer HiSeq2000 and MiSeq. In the former case, paired-end genomic sequence of multiple read lengths (49 bp and 100 bp) corresponding to eight insert size libraries (170 bp, 250 bp, 500 bp, 540 bp, 800 bp, 2 kbp, 5 kbp, and 40 kbp) were generated with a combined length of ~57 Gbp. Additional PacBio libraries were generated at the University of Seattle (5.5 Gbp combined length) and Université Paris-Sud (3.3 Gbp combined length), and the data were kind gifts. A combined total of 305,447 PacBio circular consensus reads (CCS) were generated with estimated average length of 8870 bases and estimated coverage of ~1X.

Genome and transcriptome assembly

Multiple routes were explored for the generation of an acceptable assembly [48]. The most successful strategy, as assessed by core eukaryotic gene mapping analysis (CEGMA) and the proportion of RNAseq reads that mapped to the genome assembly [115, 116], utilised *Platanus* [117], SSPACE [118] and String Graph Assembler (SGA) [119]. Here, the two MiSeq paired-end read libraries (150 bp paired-end and 300 bp paired-end libraries) and 100 bp (170 bp insert size) paired-end HiSeq read libraries were used for the *Platanus* assembly. Each of the paired-end read libraries was subject to overlapping paired-end read joining using the ErrorCorrectReads.pl algorithm of the ALLPATHS assembly package [120]. This step in ALLPATHS reduces the complexity of the input data by combining overlapping paired-end reads into single larger reads and performs well on independent benchmark tests of real and simulated data [120]. No other steps in the ALLPATHS assembly algorithm were used. These joined paired-end reads were provided to *Platanus* as single-end reads. The 500 bp and 800 bp insert size read libraries, which could not be subject to read joining as their insert sizes were too large, were included as single-end reads. This collective set of reads was provided to *Platanus*, and the method was run using its default parameters. The combined Illumina read data provided an estimated 25x coverage of the single-copy component of the genome by k-mer spectrum analysis using ALLPATHS (Additional file 1: Fig. S11). The resulting contigs from the *Platanus* [117] assembly were subject to six rounds of scaffolding and gap filling using the SSPACE [118] and SGA [119] algorithms. SSPACE was run with the following settings `-a 0.7 -m 30 -n 50 -o 20` using the 500 bp and 800 bp insert size paired-end read libraries and the 2000 bp, 5000 bp and 40,000 bp insert size mate pair read libraries. Following each round of scaffolding, SGA was run on the

scaffolds in gap filling mode (“-gapfill”) using the same combined input read library as *Platanus* above. This resulted in a de novo assembly with an N_{50} of 955 bp, comprising 2,066,288 scaffolds (Table S1).

A k-mer spectrum for the genome was calculated from the highest coverage read library (150 bp paired-end read library). It generated a single peak at 8.8x coverage, corresponding to the homozygous single-copy portion of the genome (Additional file 1: Figure S11A). Assuming a Poisson distribution that would be observed if all regions of the genome were single copy and homozygous, the estimated genome size of the single-copy proportion of genome is 487.2 Mb and the estimated size of the whole genome 2.33 Gb. The discrepancy between the Poisson model and the observed corresponds to multi-copy sequences, with a large proportion of low to medium copy number sequences represented at high frequency. There are more than 80,000 unique k-mers of length 31 that appear more than 10,000 times. These high copy number repeat sequences are those we refer to in the results and are most likely responsible for the difficulty with progressing an assembly further than we have been able to achieve.

To estimate the genome size and the proportion of the genome that is comprised of repetitive unique sequence a k-mer spectrum analysis was conducted (Additional file 1: Figure S11A). The largest Illumina paired-end read library (150-bp paired-end) was used for this analysis. Canonical k-mers were counted using jellyfish (Marçais et al. *Bioinformatics* 27(6): 764–770) at a range of different k-mer sizes (19, 21, 27 and 31). The resulting k-mer count histograms were analysed using GenomeScope [121]. Using these methods the haploid genome size was estimated to be between 330 mb and 500 mb (Additional file 1: Figure S11A). The repetitive component of the genome was estimated to be between 191 and 339 mb, and the unique component of the genome was estimated to be 141 mb to 160 mb (Additional file 1: Figure S11A). Heterozygosity was estimated to be between 2.2 and 2.6%.

The transcriptome assembly was generated by combining multiple different read libraries into a single transcriptome assembly. These included two 100 bp paired-end read libraries generated on an Illumina HiSeq2500 (200 bp insert size) that were previously published in [17]. *Euglena* transcriptome (PRJEB10085, 17) and the six 100-bp paired-end read libraries (200 bp insert size) were generated on an Illumina HiSeq2000 generated in this study (Additional file 2: Table S1, PRJNA310762). These read libraries were combined to give a total of 2.05×10^8 paired-end reads that were provided as input for transcriptome assembly. Illumina adaptors and low-quality bases were trimmed from the reads using Trimmomatic. Ribosomal RNA sequence was removed using SortMeRNA [122] using default

settings, before read error correction using BayesHammer [123] with default settings. Reads were normalized using khmer [124] with settings $-C\ 20\ -k\ 21\ -M\ 8e9$, and overlapping paired-end reads joined using ALLPATHS-LG [120] and all reads subject to de novo assembly using SGA, minimum overlap size of 80 nucleotides, no mismatches. These filtered, normalized, and joined reads were then mapped to this assembly using Bowtie2 [125]. Reads that were absent from the assembly were identified and placed with the assembled contigs into a new input file. This file containing the unassembled reads and assembled contigs was subject to assembly using SGA with an overlap size of 70. This process of identifying unmapped reads and reassembling with SGA was repeated each time, decreasing the overlap size by 10 nucleotides until a minimum overlap size of 40 was reached. This strategy was taken to minimize the occurrence of assembly errors that are commonly obtained when a default small k-mer size is used in de Bruijn graph assembly. Contigs were then subject to scaffolding using SSPACE and the full set of non-ribosomal, corrected, normalized paired-end reads using the settings $-k\ 10, -a\ 0.7, -n\ 50, -o\ 20$. Scaffolds were subject to gap filling using the SGA gap filling function. Finally, the assembled contigs were subject to base-error correction using Pilon [126] with the default settings. CEGMA [58] suggests ~88% completeness in terms of representation of coding sequence.

Genome and transcriptome structural and functional automatic annotation

In silico analysis such as open reading frame (ORF) determination, gene predictions, gene ontology (GO) and KEGG (biological pathways) and taxa distribution were performed as part of an automatic functional annotation previously described [127] with minor modifications. Six frame translation and ORF determination of assembled transcriptome sequences were predicted using TransDecoder prediction tool [61] and Gene MarkS-T [128], and the longest ORF with coding characteristics, BLAST homology, and PFAM domain information extracted [129]. The predicted ORF was queried against the NCBI non-redundant protein database using BLASTp homology searches, and the top hit for each protein with an E value cutoff $< 1e^{-10}$ retained. Using the Blast2GO automatic functional annotation tool [130], the GO annotations of the best BLAST results with an E value cutoff $< 1e^{-10}$ were generated from the GO database. The protein domain, biological pathway analyses, and top species distributions were determined using InterPro, BLAST, enzyme code and KEGG [131]. To greatly reduce run times, BLASTp and Interpro scans were processed locally prior to uploading to Blast2GO in .xml file formats.

Assembling sequence data, data mining and phylogenetic inference

Homology searches for orthologs and paralogs of specific biological annotations were performed against the predicted proteome for *E. gracilis* using BLASTp. Clustering at 100% identity was performed for the predicted *E. gracilis* proteins using the Cluster Database at High Identity (CD-HIT) [62] algorithm to remove gapped/incomplete and redundant sequences. Sequences with significant BLASTp top hit search (E value = $1e^{-10}$) were subjected to both Reversed Position Specific BLAST RPS-BLAST and InterProScan [132]. The annotated sequences with domain and/or protein signature matches were extracted using a combination of custom UNIX commands and Bio-Perl scripts and clustered to 99% identity using CD-HIT. CD-HIT outputs a set of 'non-redundant' (nr) protein representative sequences which were aligned to known eukaryotic protein reference sequences using ClustalX2 [133] and MAFFT [134]. Poorly aligned positions or gaps were removed using the gap deletion command prior to alignment, and the final alignments processed locally for phylogenetic inference with the PhyML Command Line Interface (CLI) using default settings [135], RAxML [136], FastTree [137] and MrBayes [138]. Annotations of the trees were performed using TreeGraph2 [139] and Adobe Illustrator (Adobe Inc.).

Contigs > 10 kbp in the *E. gracilis* genome

For an initial insight into the architecture of the genome contigs > 10 kbp were analyzed. These contigs were interrogated using tBLASTn with the *E. gracilis* proteome predicted from the transcriptome. Sequences with hits were further interrogated using the Exonerate algorithm [59] for insights into splicing mechanisms and coding regions using the `--protein2genome` and `--showquerygff` and `--showtargetgff` options. Sequences, and their respective splicing coordinates in gff3, were uploaded to the Artemis genome viewer [140] for visualization. Coding regions in gff formats were extracted and translated using a combination of BEDtools getfasta [141] and the EMBOSS getorf [142] tools.

Orthologous group clustering

To identify orthologous genes in *E. gracilis* shared across eukaryotic taxa, we clustered the *E. gracilis* predicted proteome with 30 selected eukaryotic taxa using OrthoFinder [70] with taxa distribution including kinetoplastids, other members of the excavates, unikonts, bikonts, green algae, land plants and red algae.

Phylogenetic analyses of ancestry of *Euglena* genes

All 36,526 predicted nucleus-encoded proteins were searched (BLASTp 2.2.29) against a custom database containing 207 organisms (Additional file 3: Table S2).

Homologues with E value $< 10^{-2}$ were retrieved. Since an unrooted phylogenetic tree can be calculated only for three or more organisms, all proteins with less than three recovered homologues (16,636 proteins) were excluded. The remaining (19,890 proteins) were aligned (MAFFT 7.273; default parameters) and trimmed (trimAl 1.2 [143], default parameters). Alignments longer than 74 amino acid residues and with all sequences determined, i.e. there was no sequence containing only undetermined characters, (18,108 alignments) were used for tree reconstruction. The trees were calculated with RAxML [136] (v8.1.17; 100 rapid bootstraps) in Metacentrum (The National Grid Infrastructure in the Czech Republic). Custom scripts (Python 3.4) were used to sort the trees into bins based on the taxonomic affiliation of the clan in which *E. gracilis* branched. The tree was included in a bin if a bipartition supported by bootstrap 75% and higher comprised of *E. gracilis* and members of one defined taxonomic group only. In 34 cases, in which *E. gracilis* was contained in two such bipartitions containing taxa from different defined group, the tree was assigned to the two respective bins.

Mitochondrial proteome prediction

The predicted proteins were subjected to Blast2GO [130] and KEGG automatic annotation server (KAAS [144]) automatic annotation, BLASTp searches against the *T. brucei*, *Homo sapiens*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana* reference mitoproteomes and, finally, targeting signal prediction using TargetP [145]. *E. gracilis* protein was predicted as mitochondrial if (i) TargetP mitochondrial score was higher than 0.9 (607 proteins), or (ii) there was an ortholog in at least one reference mitoproteome, not associated with non-mitochondrial functions (343 proteins), or (iii) assigned mitochondrial by Blast2GO (with the exception of the MTERF family) (62 proteins). The missing members of the found mitochondrial pathways and modules were identified by a manual search (81 proteins). To streamline the final annotated output and to ensure retention of only the most reliable predictions, we chose the most confident annotation between Blast2GO, BLASTp and KAAS for each protein. The final mitochondrial dataset includes 1093 proteins.

Plastid proteome prediction

The translated *E. gracilis* transcriptome (predicted proteome) was subjected to signal prediction pipeline using a combination of SignalP [146] and PrediSI [147] while chloroplast transit peptide prediction was performed using ChloroP [148]. The sequences which scored positive by either SignalP (2551 sequences) or PrediSI (4857 sequences) were cut at the predicted signal peptide cleavage site. The sequences were then truncated to

maximum length of 200 amino acid residues for faster calculation and analyzed by ChloroP. The preliminary dataset of *E. gracilis* plastid targeted proteins (1679 sequences) consisted of transcripts which scored positive in SignalP + ChloroP (59 sequences), PrediSI + ChloroP (1002 sequences) and SignalP + PrediSI + ChloroP (618 sequences) analysis. In the second step, model dataset of 920 sequences of *Arabidopsis thaliana* proteins localized to the plastid envelope, stroma, thylakoid, grana and lamellae obtained from the public AT_CHLORO proteomic database [149] were searched by BLAST against the whole translated *E. gracilis* transcriptome and the identified orthologs were then combined with the results of orthogroup clustering performed by OrthoFinder (see above). Based on these searches, an additional 144 sequences representing orthologs of *A. thaliana* chloroplast proteins were added to the dataset of *E. gracilis*-predicted plastid proteome regardless of their targeting sequences. This enriched dataset of 1823 proteins was annotated automatically using BLAST at NCBI, KOBAS [150] and KAAS [144] independently. All automatic annotations including KO and EC numbers were then revised and edited or corrected manually and used for metabolic map reconstruction. The missing enzymes and subunits of otherwise chloroplast pathways and complexes were investigated and eventually added manually to the set regardless of their targeting sequences during the manual annotation and pathway reconstruction. This approach resulted in inclusion of another 79 sequences. The final set of predicted *E. gracilis* chloroplast proteins consisted of 1902 entries.

Additional files

Additional file 1: Figure S1. Organisation of open reading frames in the *E. gracilis* genome. **Figure S2.** Functional analysis of *E. gracilis* coding capacity by Gene Ontology. **Figure S3.** Dark adapted cells have altered proteomes and transcriptomes. **Figure S4.** Orthogroup clusters in *E. gracilis* and selected eukaryotes. **Figure S5.** Phylogeny of selected shared large paralog families. **Figure S6.** Surface families of *E. gracilis*. **Figure S7.** The *E. gracilis* endomembrane system. **Figure S8.** The *E. gracilis* nuclear pore and kinetochore complexes. **Figure S9.** The predicted proteomes of *E. gracilis* organelles. **Figure S10.** Metabolism in *E. gracilis*. **Figure S11.** Additional assembly features. **Figure S12.** BUSCO comparisons between the present work and prior transcriptomes. (PDF 10993 kb)

Additional file 2: Table S1. Raw data for proteomics and transcriptomics of *E. gracilis* under adaptive conditions. Cells were grown under dark or light conditions as described in methods and subjected to protein or RNA extraction and analysed by mass spectrometry or RNAseq. Each condition was analysed in triplicate ($n = 3$) and data for individual samples together with the merged data are provided (Transcripts, Proteome), together with BLAST annotation of altered transcripts (additional tabs). (XLSX 19876 kb)

Additional file 3: Table S2. Analysis of phylogenetic relationships of *E. gracilis* proteins. The sheet contains three tables. First table summarizes the taxon composition of the custom database used for the search of homologues of *E. gracilis* proteins. Second table summarizes the number of items in each step and the pipeline. The third table gives exact numbers of trees that fell into defined taxonomic bins. (XLSX 16396 kb)

Additional file 4: Table S3. Analysis of GO term frequency, domains and large orthogroup architecture. Sheet 1: GO terms in orthogroups. The sheet has two subtables. In one the GO terms represented above 5% in each orthogroup are shown - all other GO terms with less than 5% frequency have been omitted as the numbers of sequences included are very small. The second shows the number of annotated and non-annotated sequences of each taxonomic group selected. Yellow highlight shows the GO terms of interest belonging to *molecular process* that are analyzed in this study. Sheet 2: Conserved domains from NCBI database (CDD) detected in those sequences with the GO terms of interest highlighted in sheet 1. Output provided by CDD searches. For the sequence identifiers, note that first field separated with " _ ", represents the taxonomic group to which it belongs. Sheet 3: Incidence of conserved domains detected in CDD searches and orthogroups. This table summarizes the output of the CDD searches. Gray highlight represents the conserved domains in parallel with the respective orthogroup (OG number) of the sequences for which we provide phylogenetic analyses. Sheet 4: Data for annotation of NCIII tree. *Trans*-membrane domains and conserved domains. Sheet 5: Data for annotation of REC tree. *Trans*-membrane domains and conserved domains. (XLSX 127 kb)

Additional file 5: Table S4. Accessions of genes associated with specific cellular functions. Each worksheet contains details of the orthologs and their accession numbers for a specific subset of predicted ORFs associated with an indicated cellular function, metabolic process or organelle. The first two sheets show the overall predictions (all annotated transcripts) and a summary graphic (Distributions). (XLSX 870 kb)

Additional file 6: Supplementary analyses. (DOCX 17 kb)

Additional file 7: Table S5. Surface/endomembrane proteome predictions. Panel A: Predicted numbers of ORFs encoded in the *E. gracilis* predicted proteome that contain a signal sequence (SS) together with additional determinants for stable membrane attachment (i.e. a glycosylphosphatidylinositol anchor (GPI) or trans-membrane domain (TMD)). Panel B: Frequency distribution of predicted *Euglena*-specific surface gene families, shown as number of families according to size. 608 (87.5%), *Euglena*-specific surface genes are predicted to be single-copy, whereas five families are predicted to have more than seven members. Panel C: PHYRE 2.0 summary results for an element of each multi-copy family ($n > 4$) of *E. gracilis*, including family size, residues matching the model and correspondent coverage of the sequence, percentage identity, confidence of prediction, and description of top template model. (XLSX 44 kb)

Additional file 8: Table S6. Predicted proteomes for the *E. gracilis* plastid and the mitochondrion. Panels include summaries for each organelle for numbers of genes in functional categories found, annotations for transcripts predicted as mitochondrial or chloroplast and finally a reconstruction of major mitochondrial complexes and pathways. (DOCX 141 kb)

Acknowledgements

We are most grateful to Purificación Lopéz-García, David Moreira and Peter Myler for the most generous donation of PacBio sequence data and also to Robert Field for permission to reutilize transcriptome data. We thank Douglas Lamont and the Fingerprints proteomics facility at the University of Dundee for excellent mass spectrometric analysis. Some computational resources were provided by the CESNET LM2015042 and the CERIT Scientific Cloud LM2015085, provided under the program "Projects of large research, development, and innovations infrastructures".

Funding

This work was supported by the Yousef Jameel Academic Program (through the Yousef Jameel PhD Scholarship), the Cambridge Commonwealth, European and International Trust, the Cambridge University Student Registry, the Cambridge Philosophical Society (all to TEE), the Medical Research Council (Grant #: P009018/1 to MCF), and German Aerospace Center - DLR, Cologne, on the behalf of Federal Ministry of Education and Research (BMBF), Germany (Grant no: 50WB1128 and 50WB1528 to ML), the European Research Council CZ LL1601 BFU2013-40866-P (to DPD) and the Czech Ministry of Education, Youth and Sports - National Sustainability Program II (Project BIOCEV-FAR) LQ 1604, by

the project BIOCEV (CZ.1.05/1.1.00/02.0109), by the Centre for research of pathogenicity and virulence of parasites CZ.02.1.01/0.0/0.0/16_019/0000759 and by the Czech Science Foundation project nr. 16-25280S (to VH, AV and PS).

Availability of data and materials

Assembled transcripts and predicted proteome (PRJNA298469), light/dark adapted transcriptomes (PRJNA310762). Genome data and assembly data are available from the European Nucleotide Archive under the project accession ERP109500. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD009998. Supporting analyses of several annotated systems are available in Additional file 5: Table S4 and Additional file 6: Supplementary analysis.

Authors' contributions

MCF, MLG, SK and ML conceived the study. TEE, MZ and AB carried out the experimental. TEE, MZ, AB, AN, AMGNV, MG, BP, PS, CS-M, EO'N, NNN, SSP, NV, VD, SO and MCF analyzed the data. MCF, MG, SK, ML, MZ, APJ, DD, JL, JBD, ML, SV and VH supervised the research. TEE, MZ, AB, AN, AMGNV, BP, PS, CS-M, EO'N, NNN, SSP, JBD, APJ and MCF drafted the manuscript. MCF, VH, TEE and SK edited the final draft. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹School of Life Sciences, University of Dundee, Dundee DD1 5EH, UK. ²Department of Biochemistry, University of Cambridge, Cambridge CB2 1QW, UK. ³Department of Biological and Medical Sciences, Faculty of Health and Life Sciences, Oxford Brookes University, Oxford OX3 0BP, UK. ⁴Biology Centre, Institute of Parasitology, Czech Academy of Sciences, and Faculty of Sciences, University of South Bohemia, 37005 České Budějovice, Czech Republic. ⁵Department of Parasitology, Faculty of Science, Charles University, BIOCEV, 252 50 Vestec, Czech Republic. ⁶Cell Biology Division, Department of Biology, University of Erlangen-Nuremberg, 91058 Erlangen, Germany. ⁷Centro Andaluz de Biología del Desarrollo (CABD)-CSIC, Pablo de Olavide University, Seville, Spain. ⁸Department of Plant Sciences, University of Oxford, Oxford OX1 3RB, UK. ⁹Division of Infectious Disease, Department of Medicine, University of Alberta, Edmonton, Alberta T6G, Canada. ¹⁰Laboratory of Cellular and Structural Biology, The Rockefeller University, New York, NY 10065, USA. ¹¹Department of Infection Biology, Institute of Infection and Global Health, University of Liverpool, Liverpool, UK. ¹²Department of Biological and Geographical Sciences, School of Applied Sciences, University of Huddersfield, Queensgate, Huddersfield HD1 3DH, UK. ¹³Department of Life Sciences, The Natural History Museum, Cromwell Road, London SW7 5BD, UK.

Received: 9 February 2018 Accepted: 8 January 2019

Published online: 07 February 2019

References

1. Dobell C. Antony van Leeuwenhoek and his "Little Animals." 1932. doi: <https://doi.org/10.1038/130679a0>.
2. Kim JT, Boo SM, Zakrýs B. Floristic and taxonomic accounts of the genus *Euglena* (Euglenophyceae) from Korean fresh waters. *Algae*. 1998;13:173–97.
3. Gojdic M. The genus *Euglena*. *American Association for the Advancement of Science*; 1953. doi:<https://doi.org/10.1126/science.120.3124.799-a>.
4. Zakrýs B, Walne PL. Floristic, taxonomic and phytogeographic studies of green Euglenophyta from the Southeastern United States, with emphasis

- on new and rare species. *Algal Stud für Hydrobiol Suppl Vol.* 1994;72:71–114.
5. Zakrýs B. The nuclear DNA level as a potential taxonomic character in *Euglena* Ehr. (Euglenophyceae). *Algal Stud für Hydrobiol Suppl Vol.* 1988; 483–504.
 6. Buetow DE. The biology of *Euglena*: Academic Press; 1968;49.
 7. McFadden GI. Primary and secondary endosymbiosis and the origin of plastids. *J Phycol.* 2001;37:951–9. <https://doi.org/10.1046/j.1529-8817.2001.01126.x>.
 8. Dragoş N, Péterfi LŞ, Popescu C. Comparative fine structure of pellicular cytoskeleton in *Euglena* Ehrenberg. *Arch Protistenkd.* 1997;148:277–85. [https://doi.org/10.1016/S0003-9365\(97\)80008-5](https://doi.org/10.1016/S0003-9365(97)80008-5).
 9. Daiker V, Lebert M, Richter P, Häder D-P. Molecular characterization of a calmodulin involved in the signal transduction chain of gravitaxis in *Euglena gracilis*. *Planta.* 2010;231:1229–36. <https://doi.org/10.1007/s00425-010-1126-9>.
 10. van der Horst MA, Hellingwerf KJ. Photoreceptor proteins, “star actors of modern times”: a review of the functional dynamics in the structure of representative members of six different photoreceptor families. *Acc Chem Res.* 2004;37:13–20. <https://doi.org/10.1021/ar020219d>.
 11. Heijde M, Ulm R. UV-B photoreceptor-mediated signalling in plants. *Trends Plant Sci.* 2012;17:230–7. <https://doi.org/10.1016/j.tplants.2012.01.007>.
 12. Iseki M, Matsunaga S, Murakami A, Ohno K, Shiga K, Yoshida K, et al. A blue-light-activated adenylyl cyclase mediates photoavoidance in *Euglena gracilis*. *Nature.* 2002;415:1047–51. <https://doi.org/10.1038/4151047a>.
 13. Masuda S. Light detection and signal transduction in the BLUF photoreceptors. *Plant Cell Physiol.* 2013;54:171–9. <https://doi.org/10.1093/pcp/pcs173>.
 14. Richter PR, Schuster M, Lebert M, Streb C, Häder D-P. Gravitaxis of *Euglena gracilis* depends only partially on passive buoyancy. *Adv Sp Res.* 2007;39: 1218–24. <https://doi.org/10.1016/J.ASR.2006.11.024>.
 15. Adl SM, Simpson AGB, Lane CE, Lukeš J, Bass D, Bowser SS, et al. The revised classification of eukaryotes. *J Eukaryot Microbiol.* 2012;59:429–93. <https://doi.org/10.1111/j.1550-7408.2012.00644.x>.
 16. Flegontova O, Flegontov P, Malviya S, Audic S, Wincker P, de Vargas C, et al. Extreme diversity of diplomonid eukaryotes in the ocean. *Curr Biol.* 2016;26: 3060–5.
 17. O'Neill EC, Trick M, Hill L, Rejzek M, Dusi RG, Hamilton CJ, et al. The transcriptome of *Euglena gracilis* reveals unexpected metabolic capabilities for carbohydrate and natural product biochemistry. *Mol Biosyst.* 2015;11: 2808–20. <https://doi.org/10.1039/C5MB00319A>.
 18. O'Neill EC, Trick M, Henrissat B, Field RA. *Euglena* in time: evolution, control of central metabolic processes and multi-domain proteins in carbohydrate and natural product biochemistry. *Perspect Sci.* 2015;6:84–93. <https://doi.org/10.1016/J.PISC.2015.07.002>.
 19. Yoshida Y, Tomiyama T, Maruta T, Tomita M, Ishikawa T, Arakawa K. De novo assembly and comparative transcriptome analysis of *Euglena gracilis* in response to anaerobic conditions. *BMC Genomics.* 2016;17:182. <https://doi.org/10.1186/s12864-016-2540-6>.
 20. Moore AN, McWatters DC, Hudson AJ, Russell AG. RNA-Seq employing a novel rRNA depletion strategy reveals a rich repertoire of snoRNAs in *Euglena gracilis* including box C/D and Ψ-guide RNAs targeting the modification of rRNA extremities. *RNA Biol.* 2018;15:1309–18. <https://doi.org/10.1080/15476286.2018.1526561>.
 21. Lukeš J, Skalický T, Týč J, Votýpka J, Yurchenko V. Evolution of parasitism in kinetoplastid flagellates. *Mol Biochem Parasitol.* 2014;195:115–22. <https://doi.org/10.1016/j.molbiopara.2014.05.007>.
 22. Flegontov P, Votýpka J, Skalický T, Logacheva MDD, Penin AAA, Tanifuji G, et al. Paratrypanosoma is a novel early-branching trypanosomatid. *Curr Biol.* 2013;23:1787–93.
 23. Jackson AP, Otto TD, Aslett M, Armstrong SD, Bringaud F, Schlacht A, et al. Kinetoplastid phylogenomics reveals the evolutionary innovations associated with the origins of parasitism. *Curr Biol.* 2016;26:161–72. <https://doi.org/10.1016/j.cub.2015.11.055>.
 24. Jackson AP. Gene family phylogeny and the evolution of parasite cell surfaces. *Mol Biochem Parasitol.* 2016;209:64–75. <https://doi.org/10.1016/j.molbiopara.2016.03.007>.
 25. Langousis G, Hill KL. Motility and more: the flagellum of *Trypanosoma brucei*. *Nat Rev Microbiol.* 2014;12:505–18.
 26. Perdomo D, Bonhivers M, Robinson D. The trypanosome flagellar pocket collar and its ring forming protein—TbBILBO1. *Cell.* 2016;5:9. <https://doi.org/10.3390/cells5010009>.
 27. Kalb LC, Frederico YCA, Boehm C, Moreira CM do N, Soares MJ, Field MC. Conservation and divergence within the clathrin interactome of *Trypanosoma cruzi*. *Sci Rep.* 2016;6:31212. <https://doi.org/10.1038/srep31212>.
 28. Zoltner M, Horn D, de Koning HP, Field MC. Exploiting the Achilles' heel of membrane trafficking in trypanosomes. *Curr Opin Microbiol.* 2016;34:97–103. <https://doi.org/10.1016/j.mib.2016.08.005>.
 29. Hovel-Miner G, Mugnier MR, Goldwater B, Cross GAM, Papavasiliou FN. A conserved DNA repeat promotes selection of a diverse repertoire of *Trypanosoma brucei* surface antigens from the genomic archive. *PLoS Genet.* 2016;12:e1005994. <https://doi.org/10.1371/journal.pgen.1005994>.
 30. Devault A, Bañuls A-L. The promastigote surface antigen gene family of the *Leishmania* parasite: differential evolution by positive selection and recombination. *BMC Evol Biol.* 2008;8:292. <https://doi.org/10.1186/1471-2148-8-292>.
 31. Chamakh-Ayari R, Bras-Gonçalves R, Bahi-Jaber N, Petitdidier E, Markikou-Ouni W, Aoun K, et al. In vitro evaluation of a soluble *Leishmania* promastigote surface antigen as a potential vaccine candidate against human leishmaniasis. *PLoS One.* 2014;9:e92708. <https://doi.org/10.1371/journal.pone.0092708>.
 32. Mahapatra DM, Chanakya HN, Ramachandra TV. *Euglena* sp. as a suitable source of lipids for potential use as biofuel and sustainable wastewater treatment. *J Appl Phycol.* 2013;25:855–65. <https://doi.org/10.1007/s10811-013-9979-5>.
 33. Furuhashi T, Ogawa T, Nakai R, Nakazawa M, Okazawa A, Padermschoke A, et al. Wax ester and lipophilic compound profiling of *Euglena gracilis* by gas chromatography-mass spectrometry: toward understanding of wax ester fermentation under hypoxia. *Metabolomics.* 2015;11:175–83. <https://doi.org/10.1007/s11306-014-0687-1>.
 34. Yamada K, Suzuki H, Takeuchi T, Kazama Y, Mitra S, Abe T, et al. Efficient selective breeding of live oil-rich *Euglena gracilis* with fluorescence-activated cell sorting. *Sci Rep.* 2016;6:26327. <https://doi.org/10.1038/srep26327>.
 35. Miazek K, Iwanek W, Remacle C, Richel A, Goffin D. Effect of metals, metalloids and metallic nanoparticles on microalgae growth and industrial product biosynthesis: a review. *Int J Mol Sci.* 2015;16:23299–69. <https://doi.org/10.3390/ijms161023929>.
 36. Rodríguez-Zavala JS, García-García JD, Ortiz-Cruz MA, Moreno-Sánchez R. Molecular mechanisms of resistance to heavy metals in the protist *Euglena gracilis*. *J Environ Sci Heal Part A.* 2007;42:1365–78. <https://doi.org/10.1080/10934520701480326>.
 37. dos Santos Ferreira V, Rocchetta I, Conforti V, Bench S, Feldman R, Levin MJ, et al. Gene expression patterns in *Euglena gracilis*: insights into the cellular response to environmental stress. *Gene.* 2007;389:136–45.
 38. Zeng M, Hao W, Zou Y, Shi M, Jiang Y, Xiao P, et al. Fatty acid and metabolomic profiling approaches differentiate heterotrophic and mixotrophic culture conditions in a microalgal food supplement “*Euglena*”. *BMC Biotechnol.* 2016;16:49. <https://doi.org/10.1186/s12896-016-0279-4>.
 39. Dobáková E, Flegontov P, Skalický T, Lukeš J. Unexpectedly streamlined mitochondrial genome of the euglenozoan *Euglena gracilis*. *Genome Biol Evol.* 2015;7:3358–67. <https://doi.org/10.1093/gbe/evw229>.
 40. Faktorová D, Dobáková E, Peña-Díaz P, Lukeš J. From simple to supercomplex: mitochondrial genomes of euglenozoan protists. *F1000Research.* 2016;5:392. doi:<https://doi.org/10.12688/f1000research.8040.1>.
 41. Hallick RB, Hong L, Drager RG, Favreau MR, Monfort A, Orsat B, et al. Complete sequence of *Euglena gracilis* chloroplast DNA. *Nucleic Acids Res.* 1993;21:3537–44.
 42. Rogers MB, Gilson PR, Su V, McFadden GI, Keeling PJ. The complete chloroplast genome of the chlorarachniophyte *Bigelowiella natans*: evidence for independent origins of chlorarachniophyte and euglenid secondary endosymbionts. *Mol Biol Evol.* 2007;24:54–62. <https://doi.org/10.1093/molbev/msl129>.
 43. Maruyama S, Suzuki T, Weber AP, Archibald JM, Nozaki H. Eukaryote-to-eukaryote gene transfer gives rise to genome mosaicism in euglenids. *BMC Evol Biol.* 2011;11:105. <https://doi.org/10.1186/1471-2148-11-105>.
 44. Howe CJ, Barbrook AC, Nisbet RER, Lockhart PJ, Larkum AWD. The origin of plastids. *Philos Trans R Soc Lond Ser B Biol Sci.* 2008;363:2675–85. <https://doi.org/10.1098/rstb.2008.0050>.
 45. Dooijes D, Chaves I, Kieft R, Dirks-Mulder A, Martin W, Borst P. Base J originally found in kinetoplastida is also a minor constituent of nuclear DNA of *Euglena gracilis*. *Nucleic Acids Res.* 2000;28:3017–21.
 46. Stankiewicz AJ, Falchuk KH, Vallee BL. Composition and structure of zinc-deficient *Euglena gracilis* chromatin. *Biochemistry.* 1983;22:5150–6.

47. Mazur B, Falchuk KH, Vallee BL. Histone formation, gene expression, and zinc deficiency in *Euglena gracilis*. *Biochemistry*. 1984;23:42–7.
48. Ebenezer TE, Carrington M, Lebert M, Kelly S, Field MC. *Euglena gracilis* genome and transcriptome: organelles, nuclear genome assembly strategies and initial features. In: *Advances in experimental medicine and biology*; 2017. p. 125–40. https://doi.org/10.1007/978-3-319-54910-1_7.
49. Schantz ML, Schantz R. Sequence of a cDNA clone encoding beta tubulin from *Euglena gracilis*. *Nucleic Acids Res*. 1989;17:6727.
50. Jackson AP, Vaughan S, Gull K. Evolution of tubulin gene arrays in trypanosomatid parasites: genomic restructuring in *Leishmania*. *BMC Genomics*. 2006;7:261. <https://doi.org/10.1186/1471-2164-7-261>.
51. Levasseur PJ, Meng Q, Bouck GB. Tubulin genes in the algal protist *Euglena gracilis*. *J Eukaryot Microbiol*. 1994;41:468–77.
52. Milanowski R, Karnkowska A, Ishikawa T, Zakryś B. Distribution of conventional and nonconventional introns in tubulin (α and β) genes of euglenids. *Mol Biol Evol*. 2014;31:584–93. <https://doi.org/10.1093/molbev/mst227>.
53. Milanowski R, Gumińska N, Karnkowska A, Ishikawa T, Zakryś B. Intermediate introns in nuclear genes of euglenids – are they a distinct type? *BMC Evol Biol*. 2016;16:49.
54. Canaday J, Tessier LH, Imbault P, Paulus F. Analysis of *Euglena gracilis* alpha-, beta- and gamma-tubulin genes: introns and pre-mRNA maturation. *Mol Gen Genomics*. 2001;265:153–60.
55. Tessier L, Keller M, Chan RL, Fournier R, Weil J. Short leader sequences may be transferred from small RNAs to pre-mature mRNAs by trans-splicing in *Euglena*. *EMBO J*. 1991;10:2621–5.
56. Keller M, Chan RL, Tessier L-H, Weil J-H, Imbault P. Post-transcriptional regulation by light of the biosynthesis of *Euglena* ribulose-1,5-bisphosphate carboxylase/oxygenase small subunit. *Plant Mol Biol*. 1991;17:73–82. <https://doi.org/10.1007/BF00036807>.
57. Rawson JR. The characterization of *Euglena gracilis* DNA by its reassociation kinetics. *Biochim Biophys Acta*. 1975;402:171–8.
58. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*. 2007;23:1061–7. <https://doi.org/10.1093/bioinformatics/btm071>.
59. Slater GSC, Birney E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*. 2005;6:31. <https://doi.org/10.1186/1471-2105-6-31>.
60. Mair G, Shi H, Li H, Djikeng A, Aviles HO, Bishop JR, et al. A new twist in trypanosome RNA metabolism: cis-splicing of pre-mRNA. *RNA*. 2000;6:163–9.
61. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc*. 2013;8:1494–512.
62. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*. 2012;28:3150–2.
63. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 2015;31(19):3210–12.
64. Jackson AP, Quail MA, Berriman M. Insights into the genome sequence of a free-living kinetoplastid: *Bodo saltans* (Kinetoplastida: Euglenozoa). *BMC Genomics*. 2008;9:594. <https://doi.org/10.1186/1471-2164-9-594>.
65. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, et al. The genome of the African trypanosome *Trypanosoma brucei*. *Science*. 2005;309:416–22.
66. Fritz-Laylin LK, Prochnik SE, Ginger ML, Dacks JB, Carpenter ML, Field MC, et al. The genome of *Naegleria gruberi* illuminates early eukaryotic versatility. *Cell*. 2010;140:631–42. <https://doi.org/10.1016/j.cell.2010.01.032>.
67. Van Assche E, Van Puyvelde S, Vanderleyden J, Steenackers HP. RNA-binding proteins involved in post-transcriptional regulation in bacteria. *Front Microbiol*. 2015;6:141. <https://doi.org/10.3389/fmicb.2015.00141>.
68. Araujo PR, Teixeira SM. Regulatory elements involved in the post-transcriptional control of stage-specific gene expression in *Trypanosoma cruzi*: a review. *Mem Inst Oswaldo Cruz*. 2011;106:257–66.
69. Montandon PE, Stutz E. Structure and expression of the *Euglena* nuclear gene coding for the translation elongation factor EF-1 alpha. *Nucleic Acids Res*. 1990;18:75–82.
70. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol*. 2015;16:157. <https://doi.org/10.1186/s13059-015-0721-2>.
71. Salmon D, Vanwalleghem G, Morias Y, Denoel J, Krumbholz C, Lhomme F, et al. Adenylate cyclases of *Trypanosoma brucei* inhibit the innate immune response of the host. *Science*. 2012;337:463–6. <https://doi.org/10.1126/science.1222753>.
72. Ponce-Toledo RI, Moreira D, López-García P, Deschamps P. Secondary plastids of euglenids and chlorarachniophytes function with a mix of genes of red and green algal ancestry. *Mol Biol Evol*. 2018;35:2198–204. <https://doi.org/10.1093/molbev/msy121>.
73. Jackson C, Knoll AH, Chan CX, Verbruggen H. Plastid phylogenomics with broad taxon sampling further elucidates the distinct evolutionary origins and timing of secondary green plastids. *Sci Rep*. 2018;8:1523. <https://doi.org/10.1038/s41598-017-18805-w>.
74. 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. <https://doi.org/10.1038/nature11681>.
75. Dorrell RG, Gile G, McCallum G, Méheust R, Bapteste EP, Klinger CM, et al. Chimeric origins of ochrophytes and haptophytes revealed through an ancient plastid proteome. *elife*. 2017;6. <https://doi.org/10.7554/eLife.23717>.
76. Dunin-Horkawicz S, Lupas AN. Comprehensive analysis of HAMP domains: implications for transmembrane signal transduction. *J Mol Biol*. 2010;397:1156–74. <https://doi.org/10.1016/j.jmb.2010.02.031>.
77. Anantharaman V, Aravind L. Cache – a signaling domain common to animal Ca(2+)-channel subunits and a class of prokaryotic chemotaxis receptors. *Trends Biochem Sci*. 2000;25:535–7.
78. Szöör B, Haanstra JR, Gualdrón-López M, Michels PA. Evolution, dynamics and specialized functions of glycosomes in metabolism and development of trypanosomatids. *Curr Opin Microbiol*. 2014;22:79–87. <https://doi.org/10.1016/j.cmi.2014.09.006>.
79. Morales J, Hashimoto M, Williams TA, Hirawake-mogi H, Makiuchi T, Tsubouchi A, et al. Differential remodelling of peroxisome function underpins the environmental and metabolic adaptability of diplomonads and kinetoplastids. *Proc R Soc B*. 2016;283:20160520.
80. Güther MLS, Urbaniak MD, Tavendale A, Prescott A, Ferguson MAJ. High-confidence glycosome proteome for procyclic form *Trypanosoma brucei* by epitope-tag organelle enrichment and SILAC proteomics. *J Proteome Res*. 2014;13:2796–806. <https://doi.org/10.1021/pr401209w>.
81. Loneragan TA. Regulation of cell shape in *Euglena*. IV. Localization of actin, myosin and calmodulin. *J Cell Sci*. 1985;77:197–208.
82. Gadelha C, Zhang W, Chamberlain JW, Chait BT, Wickstead B, Field MC. Architecture of a host-parasite interface: complex targeting mechanisms revealed through proteomics. *Mol Cell Proteomics*. 2015;14:1911–26. <https://doi.org/10.1074/mcp.M114.047647>.
83. Barsanti L, Passarelli V, Walne PL, Gualtieri P. The photoreceptor protein of *Euglena*. *FEBS Lett*. 2000;482:247–51.
84. Venkatesh D, Boehm C, Barlow LD, Nankisoor NN, O'Reilly A, Kelly S, et al. Evolution of the endomembrane systems of trypanosomatids – conservation and specialisation. *J Cell Sci*. 2017;130:1421–34. <https://doi.org/10.1242/jcs.197640>.
85. Zhou Q, Gheiratmand L, Chen Y, Lim TK, Zhang J, Li S, et al. A comparative proteomic analysis reveals a new bi-lobe protein required for bi-lobe duplication and cell division in *Trypanosoma brucei*. *PLoS One*. 2010;5:e9660. <https://doi.org/10.1371/journal.pone.0009660>.
86. Esson HJ, Morriswood B, Yavuz S, Vidilaseris K, Dong G, Warren G. Morphology of the trypanosome bilobe, a novel cytoskeletal structure. *Eukaryot Cell*. 2012;11:761–72. <https://doi.org/10.1128/EC.05287-11>.
87. Morriswood B, Havlicek K, Demmel L, Yavuz S, Sealey-Cardona M, Vidilaseris K, et al. Novel bilobe components in *Trypanosoma brucei* identified using proximity-dependent biotinylation. *Eukaryot Cell*. 2013;12:356–67. <https://doi.org/10.1128/EC.00326-12>.
88. McAllister MR, Ikeda KN, Lozano-Núñez A, Anrather D, Unterwurzacher V, Gossenreiter T, et al. Proteomic identification of novel cytoskeletal proteins associated with TbPLK, an essential regulator of cell morphogenesis in *Trypanosoma brucei*. *Mol Biol Cell*. 2015;26:3013–29. <https://doi.org/10.1091/mbc.E15-04-0219>.
89. Aslett M, Aurecochea C, Berriman M, Brestelli J, Brunk BP, Carrington M, et al. TriTrypDB: a functional genomic resource for the Trypanosomatidae. *Nucleic Acids Res*. 2010;38(Database issue):D457–62. <https://doi.org/10.1093/nar/gkp851>.
90. Bugreev DV, Pezza RJ, Mazina OM, Voloshin ON, Camerini-Otero RD, Mazin AV. The resistance of DMC1 D-loops to dissociation may account for the DMC1 requirement in meiosis. *Nat Struct Mol Biol*. 2011;18:56–60. <https://doi.org/10.1038/nsmb.1946>.
91. Koreny L, Field MC. Ancient eukaryotic origin and evolutionary plasticity of nuclear lamina. *Genome Biol Evol*. 2016;8:2663–71.

92. Obado SO, Brillantes M, Uryu K, Zhang W, Ketaren NE, Chait BT, et al. Interactome mapping reveals the evolutionary history of the nuclear pore complex. *PLoS Biol.* 2016;14:e1002365. <https://doi.org/10.1371/journal.pbio.1002365>.
93. Akiyoshi B, Gull K. Discovery of unconventional kinetochores in kinetoplastids. *Cell.* 2014;156:1247–58. <https://doi.org/10.1016/j.cell.2014.01.049>.
94. D'Archivio S, Wickstead B. Trypanosome outer kinetochore proteins suggest conservation of chromosome segregation machinery across eukaryotes. *J Cell Biol.* 2017;216:379–91. <https://doi.org/10.1083/jcb.201608043>.
95. Lukeš J, Guilbride DL, Votýpka J, Žilková A, Benne R, Englund PT. Kinetoplast DNA network: evolution of an improbable structure. *Eukaryot Cell.* 2002;1:495–502.
96. David V, Flegontov P, Gerasimov E, Tanifuji G, Hashimi H, Logacheva MD, et al. Gene loss and error-prone RNA editing in the mitochondrion of *Perkinsela*, an endosymbiotic kinetoplastid. *MBio.* 2015;6:1–12.
97. Pusnik M, Schmidt O, Perry AJ, Oeljeklaus S, Niemann M, Warscheid B, et al. Mitochondrial preprotein translocase of trypanosomatids has a bacterial origin. *Curr Biol.* 2011;21:1738–43.
98. Zarsky V, Tachezy J, Dolezal P. Tom40 is likely common to all mitochondria. *Curr Biol.* 2012;22:R479–81.
99. Pusnik M, Schmidt O, Perry AJ, Oeljeklaus S, Niemann M, Warscheid B, et al. Response to Zarsky et al. *Curr Biol.* 2012;22:R481–2.
100. Mani J, Meisinger C, Schneider A. Peeping at TOMs — diverse entry gates to mitochondria provide insights into the evolution of eukaryotes. *Mol Biol Evol.* 2016;33:337–51.
101. Perez E, Lapaille M, Degand H, Cilibrasi L, Villavicencio-Queijeiro A, Morsomme P, et al. The mitochondrial respiratory chain of the secondary green alga *Euglena* shares many additional subunits with parasitic Trypanosomatidae. *Mitochondrion.* 2014;19:338–49.
102. Gomez-Silva B, Timko MP, Schiff JA. Chlorophyll biosynthesis from glutamate or 5-aminolevulinic acid in intact *Euglena* chloroplasts. *Planta.* 1985;165:12–22. <https://doi.org/10.1007/BF00392206>.
103. Kim D, Filtz MR, Proteau PJ. The methylerythritol phosphate pathway contributes to carotenoid but not phytol biosynthesis in *Euglena*. *J Nat Prod.* 2004;67:1067–9. <https://doi.org/10.1021/np049892x>.
104. Eggimann G, Sweeney K, Bolt H, Rozatian N, Cobb S, Denny P. The role of phosphoglycans in the susceptibility of *Leishmania mexicana* to the temporin family of anti-microbial peptides. *Molecules.* 2015;20:2775–85. <https://doi.org/10.3390/molecules20022775>.
105. Saint-Guilay A, Schantz ML, Schantz R. Structure and expression of a cDNA encoding a histone H2A from *Euglena*. *Plant Mol Biol.* 1994;24:941–8.
106. Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet.* 2012;13:227–32. <https://doi.org/10.1038/nrg3185>.
107. Hutner SH, Zahalsky AC, Aaronson S, Baker H, Frank O. Culture media for *Euglena*. In: *Methods in Cell Biology*. Academic Press; 1966. p. 217–28. [https://doi.org/10.1016/S0091-679X\(08\)62140-8](https://doi.org/10.1016/S0091-679X(08)62140-8).
108. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol.* 2008;26:1367–72. <https://doi.org/10.1038/nbt.1511>.
109. Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat Methods.* 2016;13:731–40. <https://doi.org/10.1038/nmeth.3901>.
110. Reynolds ES. The use of lead citrate at high pH as an electron-opaque stain in electron microscopy. *J Cell Biol.* 1963;17:208–12. <http://www.ncbi.nlm.nih.gov/pubmed/13986422>.
111. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 2014;30:2114–20. <https://doi.org/10.1093/bioinformatics/btu170>.
112. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods.* 2017;14:417–9. <https://doi.org/10.1038/nmeth.4197>.
113. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014;15:550. <https://doi.org/10.1186/s13059-014-0550-8>.
114. Obado S, Field MC, Chait BT, Rout MP. High-efficiency isolation of nuclear envelope protein complexes from trypanosomes. *Methods Mol Biol.* 2016;1411:67–80.
115. Hornett EA, Wheat CW. Quantitative RNA-Seq analysis in non-model species: assessing transcriptome assemblies as a scaffold and the utility of evolutionary divergent genomic reference species. *BMC Genomics.* 2012;13:361. <https://doi.org/10.1186/1471-2164-13-361>.
116. O'Neil ST, Emrich SJ. Assessing de novo transcriptome assembly metrics for consistency and utility. *BMC Genomics.* 2013;14:465. <https://doi.org/10.1186/1471-2164-14-465>.
117. Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, et al. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* 2014;24:1384–95. <https://doi.org/10.1101/gr.170720.113>.
118. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics.* 2011;27:578–9. <https://doi.org/10.1093/bioinformatics/btq683>.
119. Simpson JT, Durbin R. Efficient de novo assembly of large genomes using compressed data structures. *Genome Res.* 2012;22:549–56. <https://doi.org/10.1101/gr.126953.111>.
120. Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci U S A.* 2011;108:1513–8. <https://doi.org/10.1073/pnas.1017351108>.
121. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, et al. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics.* 2017;33:2202–4. <https://doi.org/10.1093/bioinformatics/btx153>.
122. Kopylova E, Noé L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics.* 2012;28:3211–7. <https://doi.org/10.1093/bioinformatics/bts611>.
123. Nikolenko SI, Korobeynikov AI, Alekseyev MA. BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC Genomics.* 2013;14(Suppl 1):S7. <https://doi.org/10.1186/1471-2164-14-S1-S7>.
124. Crusoe MR, Alameldin HF, Awad S, Boucher E, Caldwell A, Cartwright R, et al. The khmer software package: enabling efficient nucleotide sequence analysis. *F1000Research.* 2015;4. doi:<https://doi.org/10.12688/f1000research.6924.1>.
125. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 2012;9:357–9. <https://doi.org/10.1038/nmeth.1923>.
126. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One.* 2014;9:e112963. <https://doi.org/10.1371/journal.pone.0112963>.
127. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 2016;44:D457–62.
128. Tang S, Lomsadze A, Borodovsky M. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res.* 2015;43:e78. <https://doi.org/10.1093/nar/gkv227>.
129. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. *Nucleic Acids Res.* 2018. <https://doi.org/10.1093/nar/gky995>.
130. Conesa A, Götz S, García-gómez JM, Terol J, Jalón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics.* 2005;21:3674–6. <https://doi.org/10.1093/bioinformatics/bti610>.
131. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 2017;45:D353–61. <https://doi.org/10.1093/nar/gkw1092>.
132. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics.* 2014;30:1236–40. <https://doi.org/10.1093/bioinformatics/btu031>.
133. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and Clustal X version 2.0. *Bioinformatics.* 2007;23:2947–8. <https://doi.org/10.1093/bioinformatics/btm404>.
134. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 2013;30:772–80. <https://doi.org/10.1093/molbev/mst010>.
135. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol.* 2010;59:307–21. <https://doi.org/10.1093/sysbio/syq010>.
136. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics.* 2014;30:1312–3. <https://doi.org/10.1093/bioinformatics/btu033>.
137. Price MN, Dehal PS, Arkin AP. FastTree 2 - approximately maximum-likelihood trees for large alignments. *PLoS One.* 2010;5:e9490. <https://doi.org/10.1371/journal.pone.0009490>.
138. Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics.* 2001;17:754–5.

139. Stöver BC, Müller KF. TreeGraph 2: combining and visualizing evidence from different phylogenetic analyses. *BMC Bioinformatics*. 2010;11:7. <https://doi.org/10.1186/1471-2105-11-7>.
140. Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics*. 2012;28:464–9. <https://doi.org/10.1093/bioinformatics/btr703>.
141. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26:841–2. <https://doi.org/10.1093/bioinformatics/btq033>.
142. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*. 2000;16:276–7. <http://www.ncbi.nlm.nih.gov/pubmed/10827456>.
143. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinform Appl NOTE*. 2009;25:1972–3. <https://doi.org/10.1093/bioinformatics/btp348>.
144. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res*. 2007;35(Web Server issue):W182–5. <https://doi.org/10.1093/nar/gkm321>.
145. Emanuelsson O, Brunak S, von Heijne G, Nielsen H. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc*. 2007;2:953–71. <https://doi.org/10.1038/nprot.2007.131>.
146. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods*. 2011;8:785–6. <https://doi.org/10.1038/nmeth.1701>.
147. Hiller K, Grote A, Scheer M, Münch R, Jahn D. PrediSi: prediction of signal peptides and their cleavage positions. *Nucleic Acids Res*. 2004;32(WEB SERVER ISS):W375–9. <https://doi.org/10.1093/nar/gkh378>.
148. Emanuelsson O, Nielsen H, von Heijne G. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Sci*. 1999;8:978–84. <https://doi.org/10.1110/ps.8.5.978>.
149. Bruley C, Dupierris V, Salvi D, Rolland N, Ferro M. AT_CHLORO: a chloroplast protein database dedicated to sub-plastidial localization. *Front Plant Sci*. 2012;3:205. <https://doi.org/10.3389/fpls.2012.00205>.
150. Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, et al. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Res*. 2011;39(Web Server issue):W316–22. <https://doi.org/10.1093/nar/gkr483>.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year






At BMC, research is always in progress.

Learn more biomedcentral.com/submissions





Causes and Effects of Loss of Classical Nonhomologous End Joining Pathway in Parasitic Eukaryotes

 Anna Nenarokova,^{a,b}  Kristína Záhonová,^{a,c}  Marija Krasilnikova,^d  Ondřej Gahura,^a  Richard McCulloch,^d  Alena Zíková,^{a,b} Vyacheslav Yurchenko,^{e,f}  Julius Lukeš^{a,b}

^aInstitute of Parasitology, Biology Centre, Czech Academy of Sciences, České Budějovice, Czech Republic

^bFaculty of Science, University of South Bohemia, České Budějovice, Czech Republic

^cDepartment of Parasitology, Faculty of Science, Charles University, BIOCEV, Prague, Czech Republic

^dWellcome Centre for Molecular Parasitology, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, Scotland

^eMartsinovskiy Institute of Medical Parasitology, Sechenov University, Moscow, Russia

^fLife Science Research Centre and Institute of Environmental Technologies, Faculty of Science, University of Ostrava, Ostrava, Czech Republic

ABSTRACT We report frequent losses of components of the classical nonhomologous end joining pathway (C-NHEJ), one of the main eukaryotic tools for end joining repair of DNA double-strand breaks, in several lineages of parasitic protists. Moreover, we have identified a single lineage among trypanosomatid flagellates that has lost Ku70 and Ku80, the core C-NHEJ components, and accumulated numerous insertions in many protein-coding genes. We propose a correlation between these two phenomena and discuss the possible impact of the C-NHEJ loss on genome evolution and transition to the parasitic lifestyle.

IMPORTANCE Parasites tend to evolve small and compact genomes, generally endowed with a high mutation rate, compared with those of their free-living relatives. However, the mechanisms by which they achieve these features, independently in unrelated lineages, remain largely unknown. We argue that the loss of the classical nonhomologous end joining pathway components may be one of the crucial steps responsible for characteristic features of parasite genomes.

KEYWORDS DNA repair, genome size, parasite

While DNA integrity and genome stability are crucial for all living organisms, they are permanently challenged by various factors causing DNA damage. The most deleterious DNA lesions are double-strand breaks (DSBs), since accurate repair of one strand using the other one as a template, as occurs in other types of DNA damage, is not possible in this case. To fix such an extreme type of damage, cells have evolved repair mechanisms known as homologous recombination (HR) and nonhomologous end joining (NHEJ).

HR, which relies on the presence of a homologous intact template, starts with 5'-to-3' resection at the DSB, producing 3' overhangs usually longer than 100 nucleotides. At least one of the single strand ends invades the homologous region of an intact chromosome, preferentially the sister chromatid (1). This strand invasion of single-stranded DNA into a template sequence produces a displacement loop (D-loop) and is mediated by recombinases of the RecA/Rad51/RadA family, found in all three domains of life (2). Upon invasion, the free 3' end of the strand is then extended by DNA polymerase(s). Subsequent steps diverge into one of the three pathways with various mutagenic potentials: (i) the double Holliday junction (dHJ) pathway engages both ends of the DSB and can lead to sequence crossover between the broken and intact molecules, (ii) synthesis-dependent strand annealing initially involves only one

Citation Nenarokova A, Záhonová K, Krasilnikova M, Gahura O, McCulloch R, Zíková A, Yurchenko V, Lukeš J. 2019. Causes and effects of loss of classical nonhomologous end joining pathway in parasitic eukaryotes. *mBio* 10:e01541-19. <https://doi.org/10.1128/mBio.01541-19>.

Editor Joseph Heitman, Duke University

Copyright © 2019 Nenarokova et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Julius Lukeš, jula@paru.cas.cz.

This article is a direct contribution from a Fellow of the American Academy of Microbiology. Solicited external reviewers: Greg Matlashewski, McGill University; Marc Ouellette, Centre of Research in Infectious Disease, Laval University.

Received 13 June 2019

Accepted 18 June 2019

Published 16 July 2019

DSB end, and (iii) break-induced replication employs only one end of the break and can copy many kilobases from the donor sequence (3–5).

In contrast to HR, NHEJ repairs a DSB by religating the broken ends without engaging an unbroken homologous template. It is divided into two main types, classical (C-NHEJ) and alternative (A-NHEJ) NHEJ. Unlike A-NHEJ, C-NHEJ has no enzymatic overlap with HR and in mammals is directed by five core components: Ku70/Ku80 heterodimer (Ku), DNA-dependent protein kinase catalytic subunit (DNA-PKcs), DNA ligase IV (Lig4), and the XRCC4 and XLF proteins (6–8). The Ku heterodimer first recognizes and binds a DSB in a sequence-independent manner, preventing extensive DSB end resection and serving as a scaffold on which other components of the C-NHEJ machinery are subsequently assembled.

Ku recruits DNA-PKcs, with which it forms a stable complex, tethers the broken DNA ends, and blocks access of other proteins. The lesion is processed, and DNA ends are sealed by the Lig4-XRCC4-XLF complex. Depending on the type of DNA end (overhang or blunt end), other factors (such as the endonuclease Artemis and DNA polymerases) and processes (end resection and DNA synthesis) may also be involved in this repair mechanism (6, 7).

The C-NHEJ machinery is conserved from bacteria to higher eukaryotes, although the levels of conservation of its components differ. In eukaryotes, the Ku heterodimer and Lig4 represent its core. Other components are less conserved and may even be absent. While retained in animals (9, 10), DNA-PKcs is absent in the yeast *Saccharomyces cerevisiae*, in which its roles are carried out by the MRX complex (11). Whether the absence of DNA-PKcs results in a reduced use of C-NHEJ is unclear, though yeasts certainly use HR as the main mechanism for DSB repair (12). Bacterial C-NHEJ employs a reduced enzymatic machinery, which comprises a Ku homodimer, homologous to eukaryotic Ku70 and Ku80, and a DNA ligase often fused to other functional domains (13–16). C-NHEJ in Archaea also utilizes a Ku homodimer, but with a different DNA ligase, DNA polymerase, and phosphodiesterase, all of which nonetheless appear closely related to their bacterial homologues (17).

Although the C-NHEJ pathway is often considered more error-prone than the HR pathway, this view has been challenged recently by emerging evidence that the latter can often be erroneous as well, especially in large and repetitive genomes (3, 18), whereas the C-NHEJ is often robust and accurate (19). However, such fidelity does not apply to the A-NHEJ pathways, named microhomology-mediated end joining (MMEJ) and single-strand annealing (SSA). Both are frequently associated with deletions, since they rely on short regions of homology around a DSB, revealed by more extensive DSB processing than in the case of C-NHEJ. The SSA pathway is independent of Rad51 but operates by annealing 25- to 400-bp-long stretches of high sequence homology in a Rad52-dependent reaction, suggesting at least some functional overlap with the HR machinery (3–5). Since such long stretches of homology are relatively rare, SSA normally generates large deletions around the DSB and is often associated with tandem repeats. MMEJ also results in deletions (20), but the shorter lengths of homology needed for strand annealing, allied to the reaction's tolerance of mismatches, ensure that deletions are normally less extensive. However, the same substrate requirements also imply that MMEJ can cause translocations, as well as complex deletions/insertions, where insertions are usually 2- to 30-bp-long, reiterating either adjacent or distant sequences (21, 22).

In metazoans, MMEJ is facilitated by poly(ADP-ribose) polymerase 1 (23), while DSB recognition requires additional proteins. Six- to 20-bp-long microhomologies are used to allow annealing around the processed DSB (24, 25), the overhangs are cleaved off, and single-stranded gaps are filled in and ligated by DNA ligases I and/or III (26, 27). Another key component of metazoan MMEJ is DNA polymerase theta (Pol θ), which possesses both polymerase and helicase domains, tethers DSB ends, anneals the broken ends at microhomology sites, and synthesizes DNA in template-dependent and -independent manners to allow DSB religation (21, 28–31). Despite this central role in

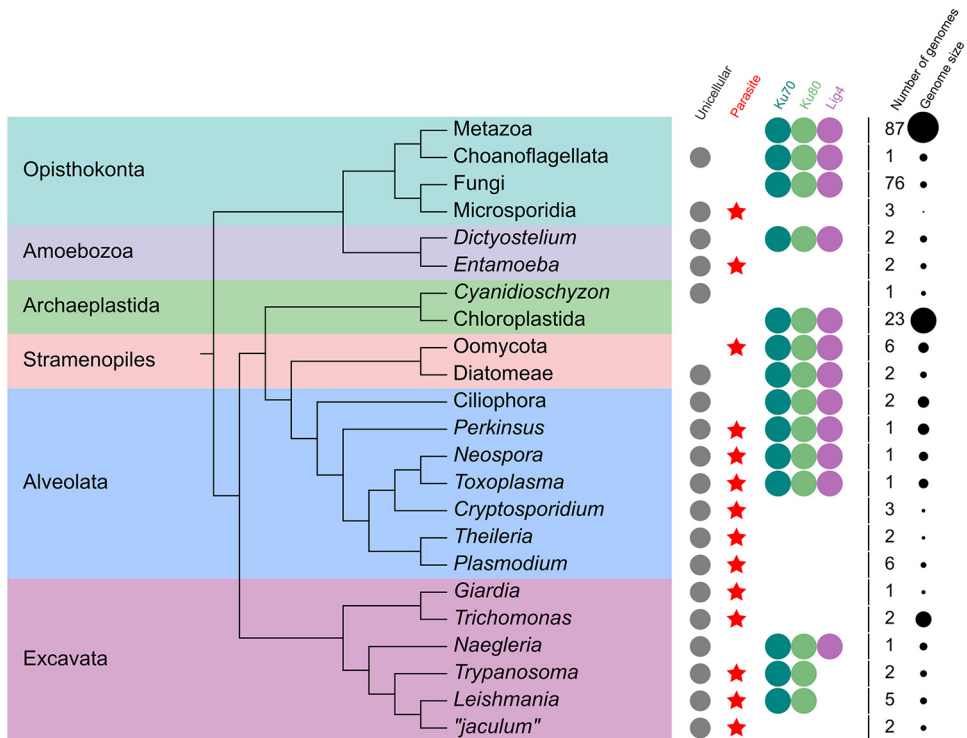


FIG 1 Distribution of main C-NHEJ components across eukaryotes. Median genome size is represented as black circles of corresponding size.

MMEJ, Pol θ is not present in all organisms. For example, yeasts employ other polymerases for this purpose (32).

The HR pathway predominates in the S and G₂ phases of the cell cycle, when newly replicated, homologous sister chromatids are present. In contrast, Ku-dependent C-NHEJ operates during the whole cell cycle, being the major DSB repair mechanism in multicellular eukaryotes (12, 33, 34). Whether MMEJ or SSA is limited to specific parts of the cell cycle is unclear.

Parasites tend to lose C-NHEJ. Perhaps because C-NHEJ is not the sole mechanism of end joining in eukaryotes, the pathway has been lost in several lineages (32, 35, 36). Prominent among the organisms lacking C-NHEJ are parasites. The absence of C-NHEJ components has been documented for the human parasitic protists *Trypanosoma* spp. (37), *Plasmodium* spp. (38), and *Encephalitozoon cuniculi* (39). Experimental analysis of DSB repair has shown that only A-NHEJ and not C-NHEJ is used in at least two of these genera (40–45).

To understand the phylogenetic distribution of C-NHEJ across eukaryotes, we searched for the orthologues of Ku70, Ku80, and Lig4, since these are the main widely conserved factors (Fig. 1).

From 230 eukaryotic genomes present in the EggNOG database (the genome of *Aspergillus oryzae*, in which Ku70 and Ku80 were artificially deleted to make HR more effective, was not included), 181, 26, and 3 genomes encoded all three, two, and one component, respectively, and in 20 genomes, all three components were missing (Table S1). The analysis revealed an overall trend of parasitic protists to lack the C-NHEJ pathway. For example, C-NHEJ is lost in microsporidia and *Entamoeba* spp., yet it is retained in free-living fungi (46) and *Dictyostelium* spp. that form their sister clades, respectively. Nonetheless, this rule is not without exceptions. Among apicomplexan parasites, all C-NHEJ components were retained in the genera *Toxoplasma* and *Neospora* yet lost in *Plasmodium*, *Cryptosporidium*, and *Theileria*. Moreover, C-NHEJ is absent in the red alga *Cyanidioschyzon merolae*, the only known free-living protist

lacking it (Fig. 1). Such a sporadic absence of C-NHEJ is most readily explained by multiple independent losses during eukaryotic evolution.

Why parasites? Two important questions arise from the observation that multiple eukaryotic lineages have discarded C-NHEJ. What processes and forces triggered the loss of such an important DNA repair pathway? What consequences might it have for genome stability and structure?

It has been suggested that the distribution of C-NHEJ in bacteria is connected with their life cycle, with the pathway present in species with a prolonged stationary phase (47, 48), during which there is no available sister chromatid to perform HR. This is also consistent with the observed predominance of C-NHEJ in the haploid cells of eukaryotes, as well as in the G₁ or G₀ phase of the cell cycle, when HR cannot be implemented and the cell has to rely on the nonhomologous DSB repair pathways (49, 50). Vice versa, the organisms that divide often and spend long time in the diploid state tend to rely on HR and lose C-NHEJ.

Alternatively, the loss of C-NHEJ may be triggered by an attempt to limit or even eradicate transposons that rely on it for their movement (51). Finally, the patchy distribution of different DSB repair pathways may reflect their relative impact on genome changes. For example, C-NHEJ can be mutagenic, contributing to sequence diversity during maturation of vertebrate immune genes (52). Consequently, the balance between the beneficial and detrimental aspects of C-NHEJ-associated mutagenesis (53) may dictate the need for its loss, facilitating use of the more faithful HR. However, the absence of C-NHEJ also results in a higher dependence on the A-NHEJ pathway, as appears to be the case during DSB repair in trypanosomatids and other organisms without C-NHEJ (40–45, 54, 55). Such prominence of A-NHEJ may become useful because of additional functions that C-NHEJ cannot perform, such as enhanced genome rearrangement, due to the reliance of A-NHEJ on annealing short, imperfect regions of homology. However, at least in the case of trypanosomatids, the extensive synteny of the *Trypanosoma brucei*, *Trypanosoma cruzi*, and *Leishmania* genomes (56) argues against the function of A-NHEJ in genome rearrangements, although we cannot exclude its reclusive role in localized genome variation, such as in multigene families (57–59).

Instead, loss of C-NHEJ can be better correlated with reduced genome size. For instance, the chordate *Oikopleura* (54), the red alga *Cyanidioschyzon* (60), and the prokaryote *Mycobacterium leprae* (61) have undergone a process of genome compaction and, unlike their relatives, notably lack C-NHEJ. Similarly, the size range from 8 to 23 Mb of the C-NHEJ-lacking genomes of the apicomplexans *Theileria parva* (62), *Cryptosporidium* spp. (63), and *Plasmodium* spp. (64) is significantly smaller than the 80-Mb genome of the related *Toxoplasma gondii* (65) (Fig. 1). The loss of C-NHEJ and subsequent gradual compaction of the genome were also observed in the evolution of microsporidians (46, 66). Importantly, Deng and colleagues associated the genome compaction in *Oikopleura* with the loss of C-NHEJ machinery (54). Consistent with this suggestion, our comparative analysis of eukaryotic genomes lacking and containing C-NHEJ machinery revealed a mean size of 29.2 Mb for the former and 667.9 Mb for the latter, a remarkable difference of >20 times ($P = 1.0 \times 10^{-8}$). While this cannot be the sole explanation of size differences, since the ~165-Mb genome of *Trichomonas vaginalis* (67) also lacks C-NHEJ machinery (although its close relative *Trichomonas tenax* has a genome of only 46 Mb [68]), it is highly plausible that when genome streamlining is advantageous, C-NHEJ tends to be discarded, either due to its dispensability or because this step further accelerates sequence loss.

Selective pressure makes parasites fast, concise, and economic, preferably exceeding their hosts in these parameters. Moreover, compared with their free-living relatives, parasites typically have smaller and streamlined genomes and are more susceptible to gene loss. All this is beneficial, since smaller genomes allow parasites to multiply faster and with lower metabolic costs (69, 70). In this context, we posit that the observed multiple independent losses of the C-NHEJ components in parasitic

lineages provide evidence that loss of this DSB repair mechanism leads to genome compaction and, in turn, provides parasites with a number of selective advantages detailed below.

At a DSB, the Ku heterodimer binds promptly to the broken DNA ends (71), protecting them from further degradation and resection by nucleases, which would lead to deleterious deletions (72). In the absence of C-NHEJ, the organism uses A-NHEJ pathways, such as MMEJ and SSA, which inevitably triggers sequence deletions (20). Moreover, the HR-based break-induced replication and SSA pathways can also produce deletions at the breakpoint flanks (73, 74). Thus, following the loss of C-NHEJ, a eukaryotic genome undergoes chromosome aberrations, including deletions and translocations, leading to loss of genetic material and consequent genome shrinkage (75–77). For instance, it has been experimentally demonstrated that A-NHEJ causes novel indel mutations in *Oikopleura*, and this process was implicated in the mechanism of genome shrinkage (54).

We may speculate about the potential mechanisms behind the genome shrinkage. Keeling and Slamovits considered two principal ways leading to the shrinkage of a genome, which are not mutually exclusive: reduction and compaction (78). Reduction is a process of elimination of some functional elements, such as protein-coding genes, whereas compaction is a process of rearranging the existing functional elements in a denser way, for instance, by removing the parts of the noncoding sequences. Both processes operate in the eukaryote genomes: they can occur together or separately. The smallest known nuclear genomes are those of parasitic microsporidia (2.5 Mb) and nucleomorphs (0.373 Mb). They represent extreme cases of both processes, having the highest gene density and the smallest number of genes among eukaryotes (78).

The physical mechanism of genome shrinkage is the loss of whole chromosomes (aneuploidy) or their parts (deletion mutations). Aneuploidy occurs due to the erroneous cell division when the chromosomes do not distribute correctly between the daughter cells. Large deletions originate as a result of DSB without rejoining, translocation of mobile elements, or erroneous, unequal, and ectopic recombination, such as between repeated regions. It is probable that this recombination is more likely to occur in the noncoding parts of genome, which have more repeated elements than protein-coding sequences, causing genome compaction (79). Small deletions occur as a result of DNA polymerase errors, such as slipping on repeats (80). Comparative studies of various animal genomes showed that on the level of small (<400-bp) indels, deletions prevail over insertions both in the protein-coding genes (81) and in the noncoding sequences (82), which may also lead to gradual loss of genetic material.

Still, we cannot exclude the possibility that loss of the C-NHEJ pathway is not the cause but rather the consequence of genome shrinkage. Even though HR occurs in mammals, C-NHEJ acts as their main DSB repair pathway (12, 33, 34). However, eukaryotes with smaller genomes and functional C-NHEJ, such as *S. cerevisiae*, preferably employ HR for DSB repair (12). There is at least one reason for C-NHEJ being the main DSB repair pathway in large eukaryotic genomes. The search for a homologous sequence during HR occurs across the entire genome, raising the risk of invading homologous ectopic sequences, which is especially high given the abundance of almost identical retrotransposon repeats in such genomes (3, 18, 83). In contrast, HR may be the mechanism of choice in small, nonrepetitive genomes, such as those of most bacteria and some unicellular eukaryotes, including parasites. The dependence of HR on the presence of homologous chromatids implies that during haploid cell cycle stages, organisms without C-NHEJ must rely on other repair pathways, such as MMEJ and/or SSA. However, as mentioned above, these pathways are highly error-prone, with a tendency to generate indel mutations (20, 75, 84–86). While deleterious for free-living eukaryotes, this sloppiness in repair mechanisms may be beneficial for parasites. By depending on these mutagenic pathways, they increase their mutation rate, thus benefiting in the arms race with the host's immune system (69, 70).

The nonrandom loss of the Ku proteins in parasitic lineages might be also associated with function(s) of the heterodimer in telomere maintenance. Ku is known to protect

telomeres from abnormal fusions and has an inhibitory effect on the recombination of normal telomeres. The Ku heterodimer also controls telomere length by recruiting telomerase and is involved in the telomere silencing effect (87–89). Furthermore, chromosomal ends and adjacent subtelomeric regions are of particular importance for parasites, as this is where factors involved in host cell interaction and immune escape mechanisms are frequently located (90, 91). Genes specifying variant surface antigens that allow parasites to evade the hosts' immune response are often found in the (sub)telomeric regions. Such surface variation systems are known for *Plasmodium* and *Babesia* spp. (64, 92), *T. brucei* (93, 94), and the fungus *Pneumocystis carinii* (95). Similar strategies have also been described for several prokaryotic pathogens, such as *Neisseria* spp. (96), *Haemophilus influenzae* (97), and *Borrelia* spp. (98). Importantly, variation of these polymorphic and fast-evolving surface proteins is promoted by DSBs, at least in the case of *T. brucei* (99). In the (sub)telomeric regions of *P. falciparum*, antigenic variation occurs via homologous and ectopic recombination (100–102), which is inhibited by Ku in the organisms that have it (10, 103). In this regard, the retention of Ku in *T. brucei* and other trypanosomatids, in the absence of other C-NHEJ components, is a notable anomaly.

Why is Ku retained in trypanosomatids? The human parasites *Trypanosoma* and *Leishmania* (Trypanosomatida, Kinetoplastida) retain Ku70 and Ku80 (104, 105) but have lost Lig4. This is an unusual combination, since other organisms lacking Lig4 usually also do not possess the Ku proteins (Fig. 1). Recently, we have sequenced and annotated the genomes of two unnamed insect flagellates belonging to the “*jaculum*” clade, a novel trypanosomatid lineage (106, 107); the raw sequencing data and the draft assembly are available at NCBI (www.ncbi.nlm.nih.gov) under BioProject PRJNA543408. Their genome sizes are 19.8 Mb and 24.9 Mb in the draft genome assemblies, and the numbers of predicted proteins are 6,163 and 7,571, correspondingly. Unexpectedly, unlike for other trypanosomatids, both Ku genes were ablated from these genomes, proving that the Ku heterodimer is not indispensable for these organisms. Interestingly, a detailed inspection of the genomes of both “*jaculum*” species revealed a high frequency of specific insertions in protein-coding genes, while deletions were rare (Fig. 2; see also Fig. S1 in the supplemental material). Since “*jaculum*” is not a basal trypanosomatid clade, but rather one from the crown (106, 107), and the insertions are specific for this group, the most parsimonious scenario is that the insertions appeared *de novo* in the common ancestor of “*jaculum*.”

Insertions were present in the majority of examined coding sequences, although they were underrepresented or completely absent from the most conserved genes, such as ribosomal proteins and glycolytic enzymes (Table S2). In 247 analyzed alignments in the two “*jaculum*” species, inserted sequences constituted 14.9% and 17.4% of the alignments, whereas in *T. brucei* only 8.9% of the alignment were represented by insertions ($P_1=4.3 \times 10^{-11}$; $P_2=1.4 \times 10^{-13}$) (Table S2). We compared the amino acid compositions of insertions and sequences without insertions, and we found that some amino acids were overrepresented or underrepresented in the inserted sequences; however, this pattern was similar in all the analyzed species (Table S3). Mass spectrometry confirmed that the insertions were indeed retained in mature proteins (Fig. 2 and Fig. S1).

Next, we investigated whether the observed insertions are neutral with respect to the function of the affected proteins. For that purpose, we mapped the insertions in selected conserved “*jaculum*” proteins on experimentally determined structures of their orthologues in *T. brucei* (Fig. 3). The inspected insertions either formed terminal extensions or were located to the external loops, but they never occurred in regions involved in ligand binding, ion coordination, or interaction with other molecules. This observation is fully consistent with the hypothesis that all insertions are functionally neutral.

We propose that the observed features are a consequence of the loss of the Ku heterodimer. Moreover, our data suggest an additional, so far unexplored, role(s) of Ku



FIG 2 Multiple insertions are present in “*jaculum*” proteins. The N-terminal part of the poly(A)-binding protein alignment of chosen trypanosomatids is shown (full-length alignment is available in Fig. S1). Insertions present in “*jaculum*” proteins are highlighted by yellow background. Peptides identified by mass spectrometry are underlined in black. Two dots represent regions of the sequence alignment that are conserved among the species and were omitted for simplicity.

in trypanosomatid parasites. In all examined species, with the sole exception of the “*jaculum*” lineage, Lig4 is absent but both Ku70 and Ku80 are retained (Fig. 1). Data available from *Trypanosoma cruzi*, *T. brucei*, and *Leishmania* spp. indicate that the Ku heterodimer does not participate in C-NHEJ and that in the corresponding genomes DSBs are predominantly repaired via HR and MMEJ (37, 43–45, 108). However, it is possible that the Ku70/80 complex plays a role in DSB repair even without its partner Lig4, because it may act as “first aid,” binding within seconds to the disrupted DNA ends (71), holding them together and protecting them from further damage until the slower HR or A-NHEJ proteins come to serve. Such a role may be important in *Leishmania* spp. and *T. brucei*, in which pronounced levels of genome rearrangements are observed, either genome-wide or in the subtelomeric region for immune evasion, and might involve DNA DSBs (109, 110). Alternatively, Ku70 and Ku80 are involved in other DNA repair pathways, such as base excision and DNA alkylation repair (111), although a role for Ku in these processes has so far not been examined in trypanosomatids. Moreover, together with the MRN complex, the Ku heterodimer may serve as a signaling molecule, modulating activity of the ATM kinase, which phosphorylates other factors and initiates a signaling cascade in the DNA damage response pathway (10). Again, the function of the ATM kinase has not yet been scrutinized in trypanosomatids. Finally, the Ku proteins play an important role in telomere maintenance (104, 105, 112). Data obtained from the analysis of the “*jaculum*” genomes may shed light on the genome-wide roles of these conserved and multifunctional proteins not only in trypanosomatids but also in other eukaryotes.

Taking the alternative end joining pathways into consideration may give us a hint regarding the origin of the insertions that are prominent in “*jaculum*.” In metazoan MMEJ, DNA polymerase θ uses only one to four complementary nucleotides to initiate polymerization, frequently producing short templated and nontemplated insertions (113, 114), reminiscent of those pervading the “*jaculum*” genome. We consider as highly plausible a hypothesis that in the “*jaculum*” trypanosomatids, the insertions may result from the erroneous A-NHEJ and HR repair processes, unconstrained by the Ku proteins. Similarly, in tunicate *Oikopleura dioica*, which lacks Ku70/80 and other components of C-NHEJ, DSB repair by A-NHEJ results in acquisition of multiple novel insertions (54).

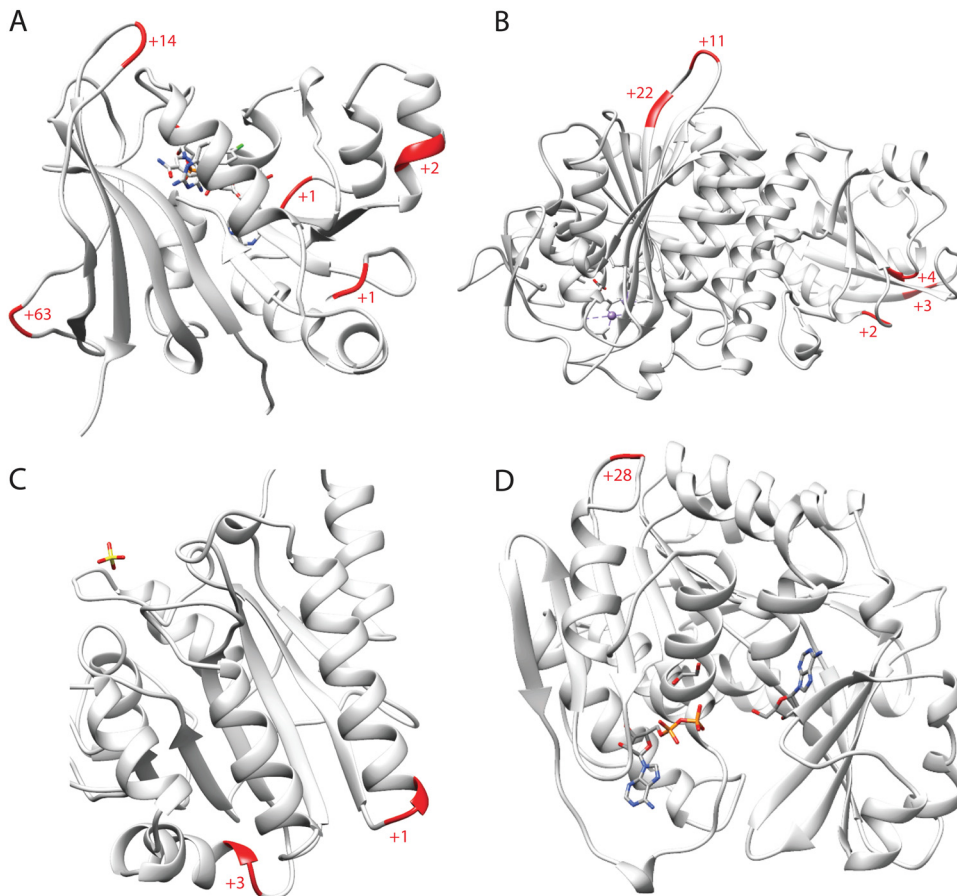


FIG 3 Mapping of insertions in the “*jaculum*” proteins onto structures of dihydrofolate reductase in complex with pyrimethamine (118) (A), leucyl aminopeptidase (119) (B), the phosphatase domain of phosphoglycerate mutase (120) (C), and adenosine kinase in complex with adenosine and AMPPNP (121) (D) from *T. brucei*. The positions and lengths of insertions in the “*jaculum*” proteins are shown in red.

An interesting question is why the observed insertions in “*jaculum*” and other trypanosomatids were significantly prevalent over deletions (Fig. 2 and Fig. S1). It is known that insertions in protein-coding sequences are usually several times more frequent than deletions, apparently because the latter are generally more deleterious and more susceptible to purifying selection (115). We also noticed that amino acids are predominantly altered in the flanking regions of the insertions and may represent remnants of the deletions, rendering these parts of the alignment to be inaccurately aligned. Moreover, the lengths of the inserted region are often variable in different species, which may be explained by consequent insertions and deletions (Fig. 2 and Fig. S1).

A comparably high incidence of indel mutations, accompanied by loss of all main C-NHEJ components, has been reported for the causative agent of human malaria, *Plasmodium falciparum* (42) (Fig. 1). In this protist, the occurrence of indels is over 10-fold higher than that of base substitutions (116). It is therefore worth pointing out that in most other organisms, base substitutions are much more frequent than indels. For example, the substitution-to-indel ratios are approximately 10:1 in primates and 20:1 in bacteria (117). While *P. falciparum* is known to be a highly polymorphic and fast-evolving parasite (116), these features are so far not associated with the absence of C-NHEJ. The above-described circumstantial evidence makes the putative connection between the DNA repair pathways and the unique features of the *Plasmodium* genomes worth exploring.

Concluding remarks. We have found that the C-NHEJ pathway, which is a highly conserved key eukaryotic DNA repair pathway, has been independently lost multiple

times in several parasitic protist lineages. We provide several alternative explanations for these seemingly nonrandom losses. Moreover, we raise the question of whether parasites benefit from this repair mechanism or, unlike their free-living kin, try to free themselves from its constraints.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at <https://doi.org/10.1128/mBio.01541-19>.

FIG S1, DOCX file, 0.4 MB.

TABLE S1, XLSX file, 0.6 MB.

TABLE S2, XLSX file, 0.1 MB.

TABLE S3, XLSX file, 0.03 MB.

ACKNOWLEDGMENTS

Support from the Czech Grant Agency 18-15962S and the ERD Funds, project OPVVV 16_019/0000759, to J.L. and V.Y., ERC CZ LL1601 to J.L., and the Czech Grant Agency 18-17529S to A.Z. is acknowledged. Work in R.M.'s lab is supported by the BBSRC (BB/K006495/1, BB/M028909/1, and BB/N016165/1) and the Wellcome Trust (206815 and 104111).

None of the funding agencies acknowledged had any role in the decision to publish.

We have no competing interests to declare.

REFERENCES

- Chang HHY, Pannunzio NR, Adachi N, Lieber MR. 2017. Non-homologous DNA end joining and alternative pathways to double-strand break repair. *Nat Rev Mol Cell Biol* 18:495–506. <https://doi.org/10.1038/nrm.2017.48>.
- Jain S, Sugawara N, Haber JE. 2016. Role of double-strand break end-tethering during gene conversion in *Saccharomyces cerevisiae*. *PLoS Genet* 12:e1005976. <https://doi.org/10.1371/journal.pgen.1005976>.
- Rodgers K, McVey M. 2016. Error-prone repair of DNA double-strand breaks. *J Cell Physiol* 231:15–24. <https://doi.org/10.1002/jcp.25053>.
- Haber JE. 2018. DNA repair: the search for homology. *Bioessays* 40:1700229. <https://doi.org/10.1002/bies.201700229>.
- Kramara J, Osia B, Malkova A. 2018. Break-induced replication: the where, the why, and the how. *Trends Genet* 34:518–531. <https://doi.org/10.1016/j.tig.2018.04.002>.
- Waters CA, Strande NT, Wyatt DW, Pryor JM, Ramsden DA. 2014. Nonhomologous end joining: a good solution for bad ends. *DNA Repair (Amst)* 17:39–51. <https://doi.org/10.1016/j.dnarep.2014.02.008>.
- Williams GJ, Hammel M, Radhakrishnan SK, Ramsden D, Lees-Miller SP, Tainer JA. 2014. Structural insights into NHEJ: building up an integrated picture of the dynamic DSB repair super complex, one component and interaction at a time. *DNA Repair (Amst)* 17:110–120. <https://doi.org/10.1016/j.dnarep.2014.02.009>.
- Her J, Bunting SF. 2018. How cells ensure correct repair of DNA double-strand breaks. *J Biol Chem* 293:10502–10511. <https://doi.org/10.1074/jbc.TM118.000371>.
- Doré AS, Drake AC, Brewerton SC, Blundell TL. 2004. Identification of DNA-PK in the arthropods: evidence for the ancient ancestry of vertebrate non-homologous end-joining. *DNA Repair (Amst)* 3:33–41. <https://doi.org/10.1016/j.dnarep.2003.09.003>.
- Fell VL, Schild-Poulter C. 2015. The Ku heterodimer: function in DNA repair and beyond. *Mutat Res Rev Mutat Res* 763:15–29. <https://doi.org/10.1016/j.mrrev.2014.06.002>.
- Chen L, Trujillo K, Ramos W, Sung P, Tomkinson AE. 2001. Promotion of Dnl4-catalyzed DNA end-joining by the Rad50/Mre11/Xrs2 and Hdf1/Hdf2 complexes. *Mol Cell* 8:1105–1115. [https://doi.org/10.1016/S1097-2765\(01\)00388-4](https://doi.org/10.1016/S1097-2765(01)00388-4).
- Emerson CH, Bertuch AA. 2016. Consider the workhorse: nonhomologous end-joining in budding yeast. *Biochem Cell Biol* 94:396–406. <https://doi.org/10.1139/bcb-2016-0001>.
- Aravind L, Koonin EV. 2001. Prokaryotic homologs of the eukaryotic DNA-end-binding protein Ku, novel domains in the Ku protein and prediction of a prokaryotic double-strand break repair system. *Genome Res* 11:1365–1374. <https://doi.org/10.1101/gr.181001>.
- Della M, Palmbo PL, Tseng H-M, Tonkin LM, Daley JM, Topper LM, Pitcher RS, Tomkinson AE, Wilson TE, Doherty AJ. 2004. Mycobacterial Ku and ligase proteins constitute a two-component NHEJ repair machine. *Science* 306:683–685. <https://doi.org/10.1126/science.1099824>.
- Shuman S, Glickman MS. 2007. Bacterial DNA repair by non-homologous end joining. *Nat Rev Microbiol* 5:852–861. <https://doi.org/10.1038/nrmicro1768>.
- Pitcher RS, Brissett NC, Doherty AJ. 2007. Nonhomologous end-joining in bacteria: a microbial perspective. *Annu Rev Microbiol* 61:259–282. <https://doi.org/10.1146/annurev.micro.61.080706.093354>.
- Bartlett EJ, Brissett NC, Doherty AJ. 2013. Ribonucleolytic resection is required for repair of strand displaced nonhomologous end-joining intermediates. *Proc Natl Acad Sci U S A* 110:E1984–E1991. <https://doi.org/10.1073/pnas.1302616110>.
- Malkova A, Haber JE. 2012. Mutations arising during repair of chromosome breaks. *Annu Rev Genet* 46:455–473. <https://doi.org/10.1146/annurev-genet-110711-155547>.
- Bétermier M, Bertrand P, Lopez BS. 2014. Is non-homologous end-joining really an inherently error-prone process? *PLoS Genet* 10:e1004086. <https://doi.org/10.1371/journal.pgen.1004086>.
- Seol J-H, Shim EY, Lee SE. 2018. Microhomology-mediated end joining: good, bad and ugly. *Mutat Res* 809:81–87. <https://doi.org/10.1016/j.mrfmmm.2017.07.002>.
- Koole W, van Schendel R, Karambelas AE, van Heteren JT, Okihara KL, Tijsterman M. 2014. A polymerase theta-dependent repair pathway suppresses extensive genomic instability at endogenous G4 DNA sites. *Nat Commun* 5:3216. <https://doi.org/10.1038/ncomms4216>.
- Wood RD, Doublie S. 2016. DNA polymerase θ (POLQ), double-strand break repair, and cancer. *DNA Repair (Amst)* 44:22–32. <https://doi.org/10.1016/j.dnarep.2016.05.003>.
- Robert I, Dantzer F, Reina-San-Martin B. 2009. Parp1 facilitates alternative NHEJ, whereas Parp2 suppresses IgH/c-myc translocations during immunoglobulin class switch recombination. *J Exp Med* 206:1047–1056. <https://doi.org/10.1084/jem.20082468>.
- Rass E, Grabarz A, Plo I, Gautier J, Bertrand P, Lopez BS. 2009. Role of Mre11 in chromosomal nonhomologous end joining in mammalian cells. *Nat Struct Mol Biol* 16:819–824. <https://doi.org/10.1038/nsmb.1641>.
- Lee-Theilen M, Matthews AJ, Kelly D, Zheng S, Chaudhuri J. 2011. CtIP promotes microhomology-mediated alternative end joining during class-switch recombination. *Nat Struct Mol Biol* 18:75–80. <https://doi.org/10.1038/nsmb.1942>.
- Ma J-L, Kim EM, Haber JE, Lee SE. 2003. Yeast Mre11 and Rad1 proteins

- define a Ku-independent mechanism to repair double-strand breaks lacking overlapping end sequences. *Mol Cell Biol* 23:8820–8828. <https://doi.org/10.1128/mcb.23.23.8820-8828.2003>.
27. Paul K, Wang M, Mladenov E, Bencsik-Theilen A, Bednar T, Wu W, Arakawa H, Iliakis G. 2013. DNA ligases I and III cooperate in alternative non-homologous end-joining in vertebrates. *PLoS One* 8:e59505. <https://doi.org/10.1371/journal.pone.0059505>.
 28. He P, Yang W. 2018. Template and primer requirements for DNA Pol θ -mediated end joining. *Proc Natl Acad Sci U S A* 115:7747–7752. <https://doi.org/10.1073/pnas.1807329115>.
 29. Chan SH, Yu AM, McVey M. 2010. Dual roles for DNA polymerase theta in alternative end-joining repair of double-strand breaks in *Drosophila*. *PLoS Genet* 6:e1001005. <https://doi.org/10.1371/journal.pgen.1001005>.
 30. Yu AM, McVey M. 2010. Synthesis-dependent microhomology-mediated end joining accounts for multiple types of repair junctions. *Nucleic Acids Res* 38:5706–5717. <https://doi.org/10.1093/nar/gkq379>.
 31. Roerink SF, van Schendel R, Tijsterman M, Schendel R, Tijsterman M. 2014. Polymerase theta-mediated end joining of replication-associated DNA breaks in *C. elegans*. *Genome Res* 24:954–962. <https://doi.org/10.1101/gr.170431.113>.
 32. Sfeir A, Symington LS. 2015. Microhomology-mediated end joining: a back-up survival mechanism or dedicated pathway? *Trends Biochem Sci* 40:701–714. <https://doi.org/10.1016/j.tibs.2015.08.006>.
 33. Pannunzio NR, Watanabe G, Lieber MR. 2018. Nonhomologous DNA end-joining for repair of DNA double-strand breaks. *J Biol Chem* 293:10512–10523. <https://doi.org/10.1074/jbc.TM117.000374>.
 34. Beucher A, Birraux J, Tchouandong L, Barton O, Shibata A, Conrad S, Goodarzi AA, Krempler A, Jeggo PA, Löbrich M. 2009. ATM and Artemis promote homologous recombination of radiation-induced DNA double-strand breaks in G2. *EMBO J* 28:3413–3427. <https://doi.org/10.1038/emboj.2009.276>.
 35. Deriano L, Roth DB. 2013. Modernizing the nonhomologous end-joining repertoire: alternative and classical NHEJ share the stage. *Annu Rev Genet* 47:433–455. <https://doi.org/10.1146/annurev-genet-110711-155540>.
 36. Sallmyr A, Tomkinson AE. 2018. Repair of DNA double-strand breaks by mammalian alternative end-joining pathways. *J Biol Chem* 293:10536–10549. <https://doi.org/10.1074/jbc.TM117.000375>.
 37. Burton P, McBride DJ, Wilkes JM, Barry JD, McCulloch R. 2007. Ku heterodimer-independent end joining in *Trypanosoma brucei* cell extracts relies upon sequence microhomology. *Eukaryot Cell* 6:1773–1781. <https://doi.org/10.1128/EC.00212-07>.
 38. Lee AH, Symington LS, Fidock DA. 2014. DNA repair mechanisms and their biological roles in the malaria parasite *Plasmodium falciparum*. *Microbiol Mol Biol Rev* 78:469–486. <https://doi.org/10.1128/MMBR.00059-13>.
 39. Gill EE, Fast NM. 2007. Stripped-down DNA repair in a highly reduced parasite. *BMC Mol Biol* 8:24. <https://doi.org/10.1186/1471-2199-8-24>.
 40. Conway C, Proudfoot C, Burton P, Barry JD, McCulloch R. 2002. Two pathways of homologous recombination in *Trypanosoma brucei*. *Mol Microbiol* 45:1687–1700. <https://doi.org/10.1046/j.1365-2958.2002.03122.x>.
 41. Glover L, McCulloch R, Horn D. 2008. Sequence homology and microhomology dominate chromosomal double-strand break repair in African trypanosomes. *Nucleic Acids Res* 36:2608–2618. <https://doi.org/10.1093/nar/gkn104>.
 42. Kirkman LA, Lawrence EA, Deitsch KW. 2014. Malaria parasites utilize both homologous recombination and alternative end joining pathways to maintain genome integrity. *Nucleic Acids Res* 42:370–379. <https://doi.org/10.1093/nar/gkt881>.
 43. Peng D, Kurup SP, Yao PY, Minning TA, Tarleton RL. 2014. CRISPR-Cas9-mediated single-gene and gene family disruption in *Trypanosoma cruzi*. *mBio* 6:e02097-14. <https://doi.org/10.1128/mBio.02097-14>.
 44. Zhang W-W, Matlashewski G. 2015. CRISPR-Cas9-mediated genome editing in *Leishmania donovani*. *mBio* 6:e00861-15. <https://doi.org/10.1128/mBio.00861-15>.
 45. Zhang W-W, Lypczewski P, Matlashewski G. 2017. Optimized CRISPR-Cas9 genome editing for *Leishmania* and its use to target a multigene family, induce chromosomal translocation, and study DNA break repair mechanisms. *mSphere* 2:e00340-16. <https://doi.org/10.1128/mSphere.00340-16>.
 46. Galindo LJ, Torruella G, Moreira D, Timpano H, Paskerova G, Smirnov A, Nassonova E, López-García P. 2018. Evolutionary genomics of *Metchnikovella incurvata* (Metchnikovellidae): an early branching microsporidium. *Genome Biol Evol* 10:2736–2748. <https://doi.org/10.1093/gbe/evy205>.
 47. Bowater R, Doherty AJ. 2006. Making ends meet: repairing breaks in bacterial DNA by non-homologous end-joining. *PLoS Genet* 2:e8. <https://doi.org/10.1371/journal.pgen.0020008>.
 48. Weller GR, Kysela B, Roy R, Tonkin LM, Scanlan E, Della M, Devine SK, Day JP, Wilkinson A, d'Adda di Fagagna F, Devine KM, Bowater RP, Jeggo PA, Jackson SP, Doherty AJ. 2002. Identification of a DNA non-homologous end-joining complex in bacteria. *Science* 297:1686–1689. <https://doi.org/10.1126/science.1074584>.
 49. Heidenreich E, Novotny R, Kneidinger B, Holzmann V, Wintersberger U. 2003. Non-homologous end joining as an important mutagenic process in cell cycle-arrested cells. *EMBO J* 22:2274–2283. <https://doi.org/10.1093/emboj/cdg203>.
 50. Karathanasis E, Wilson TE. 2002. Enhancement of *Saccharomyces cerevisiae* end-joining efficiency by cell growth stage but not by impairment of recombination. *Genetics* 161:1015–1027.
 51. Ivics Z, Izsvák Z. 2015. Sleeping Beauty transposition. *Microbiol Spectr* 3:853–874.
 52. Chaudhuri J, Alt FW. 2004. Class-switch recombination: interplay of transcription, DNA deamination and DNA repair. *Nat Rev Immunol* 4:541–552. <https://doi.org/10.1038/nri1395>.
 53. Matic I, Taddei F, Radman M. 2004. Survival versus maintenance of genetic stability: a conflict of priorities during stress. *Res Microbiol* 155:337–341. <https://doi.org/10.1016/j.resmic.2004.01.010>.
 54. Deng W, Henriot S, Chourrout D. 2018. Prevalence of mutation-prone microhomology-mediated end joining in a chordate lacking the c-NHEJ DNA repair pathway. *Curr Biol* 28:3337–3341.e4. <https://doi.org/10.1016/j.cub.2018.08.048>.
 55. Chayot R, Montagne B, Mazel D, Ricchetti M. 2010. An end-joining repair mechanism in *Escherichia coli*. *Proc Natl Acad Sci U S A* 107:2141–2146. <https://doi.org/10.1073/pnas.0906355107>.
 56. El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, Aggarwal G, Caler E, Renauld H, Worthey EA, Hertz-Fowler C, Ghedin E, Peacock C, Bartholomeu DC, Haas BJ, Tran A-N, Wortman JR, Alsmark UCM, Angiuoli S, Anupama A, Badger J, Bringaud F, Cadag E, Carlton JM, Cerqueira GC, Creasy T, Delcher AL, Djikeng A, Embley TM, Hauser C, Ivens AC, Kummerfeld SK, Pereira-Leal JB, Nilsson D, Peterson J, Salzberg SL, Shallom J, Silva JC, Sundaram J, Westenberger S, White O, Melville SE, Donelson JE, Andersson B, Stuart KD, Hall N. 2005. Comparative genomics of trypanosomatid parasitic protozoa. *Science* 309:404–409. <https://doi.org/10.1126/science.1112181>.
 57. Weatherly DB, Peng D, Tarleton RL. 2016. Recombination-driven generation of the largest pathogen repository of antigen variants in the protozoan *Trypanosoma cruzi*. *BMC Genomics* 17:729. <https://doi.org/10.1186/s12864-016-3037-z>.
 58. Jackson AP, Berry A, Aslett M, Allison HC, Burton P, Vavrova-Anderson J, Brown R, Browne H, Corton N, Hauser H, Gamble J, Gildertthorp R, Marcello L, McQuillan J, Otto TD, Quail MA, Sanders MJ, van Tonder A, Ginger ML, Field MC, Barry JD, Hertz-Fowler C, Berriman M. 2012. Antigenic diversity is generated by distinct evolutionary mechanisms in African trypanosome species. *Proc Natl Acad Sci U S A* 109:3416–3421. <https://doi.org/10.1073/pnas.1117313109>.
 59. Flegontov P, Butenko A, Firsov S, Kraeva N, Eliáš M, Field MC, Filatov D, Flegontova O, Gerasimov ES, Hlaváčková J, Ishemgulova A, Jackson AP, Kelly S, Kostygov AY, Logacheva MD, Maslov DA, Opperdoes FR, O'Reilly A, Sádlová J, Ševčíková T, Venkatesh D, Vlček Č, Volf P, Votýpka J, Záhonová K, Yurchenko V, Lukeš J. 2016. Genome of *Leptomonas pyrrhocoris*: a high-quality reference for monoxenous trypanosomatids and new insights into evolution of *Leishmania*. *Sci Rep* 6:23704. <https://doi.org/10.1038/srep23704>.
 60. Matsuzaki M, Misumi O, Shin-I T, Maruyama S, Takahara M, Miyagishima SY, Mori T, Nishida K, Yagisawa F, Nishida K, Yoshida Y, Nishimura Y, Nakao S, Kobayashi T, Momoyama Y, Higashiyama T, Minoda A, Sano M, Nomoto H, Oishi K, Hayashi H, Ohta F, Nishizaka S, Haga S, Miura S, Morishita T, Kabeya Y, Terasawa K, Suzuki Y, Ishii Y, Asakawa S, Takano H, Ohta N, Kuroiwa H, Tanaka K, Shimizu N, Sugano S, Sato N, Nozaki H, Ogasawara N, Kohara Y, Kuroiwa T. 2004. Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. *Nature* 428:653–657. <https://doi.org/10.1038/nature02398>.
 61. Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, Wheeler PR, Honoré N, Garnier T, Churcher C, Harris D, Mungall K, Basham D, Brown D, Chillingworth T, Connor R, Davies RM, Devlin K, Duthoy S, Feltwell T, Fraser A, Hamlin N, Holroyd S, Hornsby T, Jagels K, Lacroix C, Maclean

- J, Moule S, Murphy L, Oliver K, Quail MA, Rajandream MA, Rutherford KM, Rutter S, Seeger K, Simon S, Simmonds M, Skelton J, Squares R, Squares S, Stevens K, Taylor K, Whitehead S, Woodward JR, Barrell BG. 2001. Massive gene decay in the leprosy bacillus. *Nature* 409:1007–1011. <https://doi.org/10.1038/35059006>.
62. Gardner MJ, Bishop R, Shah T, De Villiers EP, Carlton JM, Hall N, Ren Q, Paulsen IT, Pain A, Berriman M, Wilson RJM, Sato S, Ralph SA, Mann DJ, Xiong Z, Shallom SJ, Weidman J, Jiang L, Lynn J, Weaver B, Shoabi A, Domingo AR, Wasawo D, Crabtree J, Wortman JR, Haas B, Angiuoli SV, Creasy TH, Lu C, Suh B, Silva JC, Utterback TR, Feldblyum TV, Pertea M, Allen J, Nierman WC, Taracha ELN, Salzberg SL, White OR, Fitzhugh HA, Morzaria S, Venter JC, Fraser CM, Nene V. 2005. Genome sequence of *Theileria parva*, a bovine pathogen that transforms lymphocytes. *Science* 309:134–137. <https://doi.org/10.1126/science.1110439>.
63. 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. <https://doi.org/10.1038/nature02977>.
64. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, Paulsen IT, James K, Eisen JA, Rutherford K, Salzberg SL, Craig A, Kyes S, Chan M-S, Nene V, Shallom SJ, Suh B, Peterson J, Angiuoli S, Pertea M, Allen J, Selengut J, Haft D, Mather MW, Vaidya AB, Martin D, Fairlamb AH, Fraunholz MJ, Roos DS, Ralph SA, McFadden GI, Cummings LM, Subramanian GM, Mungall C, Venter JC, Carucci DJ, Hoffman SL, Newbold C, Davis RW, Fraser CM, Barrell B. 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419:498–511. <https://doi.org/10.1038/nature01097>.
65. Kissinger JC, Gajria B, Li L, Paulsen IT, Roos DS. 2003. ToxoDB: accessing the *Toxoplasma gondii* genome. *Nucleic Acids Res* 31:234–236. <https://doi.org/10.1093/nar/gkg072>.
66. Katinka MD, Duprat S, Cornillot E, Méténier 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, Vivarès CP. 2001. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature* 414:450–453. <https://doi.org/10.1038/35106579>.
67. Carlton JM, Hirt RP, Silva JC, Delcher AL, Schatz M, Zhao QQ, Wortman JR, Bidwell SL, Alsmark UCM, Besteiro S, Sicheritz-Ponten T, Noel CJ, Dacks JB, Foster PG, Simillion C, Van de Peer Y, Miranda-Saavedra D, Barton GJ, Westrop GD, Müller S, Dessi D, Fiori PL, Ren Q, Paulsen I, Zhang H, Bastida-Corcuera FD, Simoes-Barbosa A, Brown MT, Hayes RD, Mukherjee M, Okumura CY, Schneider R, Smith AJ, Vanacova S, Villalvazo M, Haas BJ, Pertea M, Feldblyum TV, Utterback TR, Shu CLC-L, Osoegawa K, de Jong PJ, Hrdy I, Horvathova L, Zubacova Z, Dolezal P, Malik S-BSB, Logsdon JM, Henze K, Gupta A, Wang CC, Dunne RL, Upcroft JA, Upcroft P, White O, Salzberg SL, Tang P, Chiu CHC-H, Lee YSY-S, Embley TM, Coombs GH, Mottram JC, Tachezy J, Fraser-Liggett CM, Johnson PJ. 2007. Draft genome sequence of the sexually transmitted pathogen *Trichomonas vaginalis*. *Science* 315:207–212. <https://doi.org/10.1126/science.1132894>.
68. Benabdelkader S, Andreani J, Gillet A, Terrer E, Pignoly M, Chaudet H, Aboudharam G, La Scola B. 2019. Specific clones of *Trichomonas tenax* are associated with periodontitis. *PLoS One* 14:e0213338. <https://doi.org/10.1371/journal.pone.0213338>.
69. Cavalier-Smith T. 2005. Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion. *Ann Bot* 95:147–175. <https://doi.org/10.1093/aob/mci010>.
70. Poulin R, Randhawa H. 2015. Evolution of parasitism along convergent lines: from ecology to genomics. *Parasitology* 142:S6–S15. <https://doi.org/10.1017/S0031182013001674>.
71. Mari P-O, Florea BI, Persengiev SP, Verkaik NS, Bruggenwirth HT, Modesti M, Giglia-Mari G, Bezstarosti K, Demmers JAA, Luider TM, Houtsmuller AB, van Gent DC. 2006. Dynamic assembly of end-joining complexes requires interaction between Ku70/80 and XRCC4. *Proc Natl Acad Sci U S A* 103:18597–18602. <https://doi.org/10.1073/pnas.0609061103>.
72. Mimitou EP, Symington LS. 2010. Ku prevents Exo1 and Sgs1-dependent resection of DNA ends in the absence of a functional MRX complex or Sae2. *EMBO J* 29:3358–3369. <https://doi.org/10.1038/emboj.2010.193>.
73. Bhargava R, Sandhu M, Muk S, Lee G, Vaidehi N, Stark JM. 2018. C-NHEJ without indels is robust and requires synergistic function of distinct XLF domains. *Nat Commun* 9:2484. <https://doi.org/10.1038/s41467-018-04867-5>.
74. Malkova A, Ira G. 2013. Break-induced replication: functions and molecular mechanism. *Curr Opin Genet Dev* 23:271–279. <https://doi.org/10.1016/j.gde.2013.05.007>.
75. Boulton SJ, Jackson SP. 1996. *Saccharomyces cerevisiae* Ku70 potentiates illegitimate DNA double-strand break repair and serves as a barrier to error-prone DNA repair pathways. *EMBO J* 15:5093–5103. <https://doi.org/10.1002/j.1460-2075.1996.tb00890.x>.
76. Difilippantonio MJ, Zhu J, Chen HT, Meffre E, Nussenzweig MC, Max EE, Ried T, Nussenzweig A. 2000. DNA repair protein Ku80 suppresses chromosomal aberrations and malignant transformation. *Nature* 404:510–514. <https://doi.org/10.1038/35006670>.
77. Ferguson DO, Sekiguchi JM, Chang S, Frank KM, Gao Y, DePinto RA, Alt FW. 2000. The nonhomologous end-joining pathway of DNA repair is required for genomic stability and the suppression of translocations. *Proc Natl Acad Sci U S A* 97:6630–6633. <https://doi.org/10.1073/pnas.110152897>.
78. Keeling PJ, Slamovits CH. 2005. Causes and effects of nuclear genome reduction. *Curr Opin Genet Dev* 15:601–608. <https://doi.org/10.1016/j.gde.2005.09.003>.
79. Vinogradov AE. 2004. Evolution of genome size: multilevel selection, mutation bias or dynamical chaos? *Curr Opin Genet Dev* 14:620–626. <https://doi.org/10.1016/j.gde.2004.09.007>.
80. Garcia-Diaz M, Kunkel TA. 2006. Mechanism of a genetic glissando: structural biology of indel mutations. *Trends Biochem Sci* 31:206–214. <https://doi.org/10.1016/j.tibs.2006.02.004>.
81. de Jong WW, Rydén L. 1981. Causes of more frequent deletions than insertions in mutations and protein evolution. *Nature* 290:157–159. <https://doi.org/10.1038/290157a0>.
82. Graur D, Shuali Y, Li WH. 1989. Deletions in processed pseudogenes accumulate faster in rodents than in humans. *J Mol Evol* 28:279–285. <https://doi.org/10.1007/BF02103423>.
83. Guirouilh-Barbat J, Lambert S, Bertrand P, Lopez BS. 2014. Is homologous recombination really an error-free process? *Front Genet* 5:175. <https://doi.org/10.3389/fgene.2014.00175>.
84. Bahmed K, Nitiss KC, Nitiss JL. 2010. Yeast Tdp1 regulates the fidelity of nonhomologous end joining. *Proc Natl Acad Sci U S A* 107:4057–4062. <https://doi.org/10.1073/pnas.0909917107>.
85. Daley JM, Wilson TE. 2005. Rejoining of DNA double-strand breaks as a function of overhang length. *Mol Cell Biol* 25:896–906. <https://doi.org/10.1128/MCB.25.3.896-906.2005>.
86. Moore JK, Haber JE. 1996. Cell cycle and genetic requirements of two pathways of nonhomologous end-joining repair of double-strand breaks in *Saccharomyces cerevisiae*. *Mol Cell Biol* 16:2164–2173. <https://doi.org/10.1128/mcb.16.5.2164>.
87. Boulton SJ, Jackson SP. 1996. Identification of a *Saccharomyces cerevisiae* Ku80 homologue: roles in DNA double strand break rejoining and in telomeric maintenance. *Nucleic Acids Res* 24:4639–4648. <https://doi.org/10.1093/nar/24.23.4639>.
88. Chico L, Ciudad T, Hsu M, Lue NF, Larriba G. 2011. The *Candida albicans* Ku70 modulates telomere length and structure by regulating both telomerase and recombination. *PLoS One* 6:e23732. <https://doi.org/10.1371/journal.pone.0023732>.
89. d'Adda di Fagagna F, Hande MP, Tong WM, Roth D, Lansdorp PM, Wang ZQ, Jackson SP. 2001. Effects of DNA nonhomologous end-joining factors on telomere length and chromosomal stability in mammalian cells. *Curr Biol* 11:1192–1196. [https://doi.org/10.1016/S0960-9822\(01\)00328-1](https://doi.org/10.1016/S0960-9822(01)00328-1).
90. Barry JD, Ginger ML, Burton P, McCulloch R. 2003. Why are parasite contingency genes often associated with telomeres? *Int J Parasitol* 33:29–45. [https://doi.org/10.1016/S0020-7519\(02\)00247-3](https://doi.org/10.1016/S0020-7519(02)00247-3).
91. Merrick CJ, Duraisingh MT. 2006. Heterochromatin-mediated control of virulence gene expression. *Mol Microbiol* 62:612–620. <https://doi.org/10.1111/j.1365-2958.2006.05397.x>.
92. Jackson AP, Otto TD, Darby A, Ramaprasad A, Xia D, Echaide IE, Farber M, Gahlot S, Gamble J, Gupta D, Gupta Y, Jackson L, Malandrin L, Malas TB, Moussa E, Nair M, Reid AJ, Sanders M, Sharma J, Tracey A, Quail MA, Weir W, Wastling JM, Hall N, Willadsen P, Lingelbach K, Shiels B, Tait A, Berriman M, Allred DR, Pain A. 2014. The evolutionary dynamics of variant antigen genes in *Babesia* reveal a history of genomic innovation underlying host-parasite interaction. *Nucleic Acids Res* 42:7113–7131. <https://doi.org/10.1093/nar/gku322>.
93. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renaud H, Bar-

- tholomeu DC, Lennard NJ, Caler E, Hamlin NE, Haas B, Böhme U, Hannick L, Aslett MA, Shallom J, Marcello L, Hou L, Wickstead B, Alsmark UCM, Arrowsmith C, Atkin RJ, Barron AJ, Bringaud F, Brooks K, Carrington M, Cherevach I, Chillingworth T-J, Churcher C, Clark LN, Corton CH, Cronin A, Davies RM, Doggett J, Djikeng A, Feldblyum T, Field MC, Fraser A, Goodhead I, Hance Z, Harper D, Harris BR, Hauser H, Hostetler J, Ivens A, Jagels K, Johnson D, Johnson J, Jones K, Kerhornou AX, Koo H, Larke N, Landfear S, Larkin C, Leech V, Line A, Lord A, Macleod A, Mooney PJ, Moule S, Martin DMA, Morgan GW, Mungall K, Norbertczak H, Ormond D, Pai G, Peacock CS, Peterson J, Quail MA, Rabinowitsch E, Rajandream M-A, Reitter C, Salzberg SL, Sanders M, Schobel S, Sharp S, Simmonds M, Simpson AJ, Tallon L, Turner CMR, Tait A, Tivey AR, Van Aken S, Walker D, Wanless D, Wang S, White B, White O, Whitehead S, Woodward J, Wortman J, Adams MD, Embley TM, Gull K, Ullu E, Barry JD, Fairlamb AH, Opperdoes F, Barrell BG, Donelson JE, Hall N, Fraser CM, Melville SE, El-Sayed NM. 2005. The genome of the African trypanosome *Trypanosoma brucei*. *Science* 309:416–422. <https://doi.org/10.1126/science.1112642>.
94. Navarro M, Cross GA, Wirtz E. 1999. *Trypanosoma brucei* variant surface glycoprotein regulation involves coupled activation/inactivation and chromatin remodeling of expression sites. *EMBO J* 18:2265–2272. <https://doi.org/10.1093/emboj/18.8.2265>.
95. Stringer JR, Keely SP. 2001. Genetics of surface antigen expression in *Pneumocystis carinii*. *Infect Immun* 69:627–639. <https://doi.org/10.1128/IAI.69.2.627-639.2001>.
96. Meyer TF, Mlawer N, So M. 1982. Pilus expression in *Neisseria gonorrhoeae* involves chromosomal rearrangement. *Cell* 30:45–52. [https://doi.org/10.1016/0092-8674\(82\)90010-1](https://doi.org/10.1016/0092-8674(82)90010-1).
97. Maskell DJ, Szabo MJ, Butler PD, Williams AE, Moxon ER. 1992. Molecular biology of phase-variable lipopolysaccharide biosynthesis by *Haemophilus influenzae*. *J Infect Dis* 165:S90–S92. https://doi.org/10.1093/infdis/165-Supplement_1-S90.
98. Plasterk RHA, Simon MI, Barbour AG. 1985. Transposition of structural genes to an expression sequence on a linear plasmid causes antigenic variation in the bacterium *Borrelia hermsii*. *Nature* 318:257–263. <https://doi.org/10.1038/318257a0>.
99. Li B. 2015. DNA double-strand breaks and telomeres play important roles in *Trypanosoma brucei* antigenic variation. *Eukaryot Cell* 14:196–205. <https://doi.org/10.1128/EC.00207-14>.
100. Freitas-Junior LH, Bottius E, Pirrit LA, Deitsch KW, Scheidig C, Guinet F, Nehrbass U, Wellems TE, Scherf A. 2000. Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of *P. falciparum*. *Nature* 407:1018–1022. <https://doi.org/10.1038/35039531>.
101. Taylor HM, Kyes SA, Newbold CI. 2000. Var gene diversity in *Plasmodium falciparum* is generated by frequent recombination events. *Mol Biochem Parasitol* 110:391–397. [https://doi.org/10.1016/S0166-6851\(00\)00286-3](https://doi.org/10.1016/S0166-6851(00)00286-3).
102. Calhoun SF, Reed J, Alexander N, Mason CE, Deitsch KW, Kirkman LA. 2017. Chromosome end repair and genome stability in *Plasmodium falciparum*. *mBio* 8:e00547-17. <https://doi.org/10.1128/mBio.00547-17>.
103. Celli GB, Denchi EL, de Lange T. 2006. Ku70 stimulates fusion of dysfunctional telomeres yet protects chromosome ends from homologous recombination. *Nat Cell Biol* 8:885–890. <https://doi.org/10.1038/ncb1444>.
104. Conway C, McCulloch R, Ginger ML, Robinson NP, Browitt A, Barry JD. 2002. Ku is important for telomere maintenance, but not for differential expression of telomeric VSG genes, in African trypanosomes. *J Biol Chem* 277:21269–21277. <https://doi.org/10.1074/jbc.M200550200>.
105. Janzen CJ, Lander F, Dreesen O, Cross G. 2004. Telomere length regulation and transcriptional silencing in KU80-deficient *Trypanosoma brucei*. *Nucleic Acids Res* 32:6575–6584. <https://doi.org/10.1093/nar/gkh991>.
106. Maslov DA, Votýpka J, Yurchenko V, Lukeš J. 2013. Diversity and phylogeny of insect trypanosomatids: all that is hidden shall be revealed. *Trends Parasitol* 29:43–52. <https://doi.org/10.1016/j.pt.2012.11.001>.
107. Záhonová K, Kostygov AY, Ševčíková T, Yurchenko V, Eliáš M. 2016. An unprecedented non-canonical nuclear genetic code with all three termination codons reassigned as sense codons. *Curr Biol* 26:2364–2369. <https://doi.org/10.1016/j.cub.2016.06.064>.
108. Genois M-M, Paquet ER, Laffitte M-C, Maity R, Rodrigue A, Ouellette M, Masson J-Y. 2014. DNA repair pathways in trypanosomatids: from DNA repair to drug resistance. *Microbiol Mol Biol Rev* 78:40–73. <https://doi.org/10.1128/MMBR.00045-13>.
109. Laffitte M-C, Leprohon P, Papadopoulou B, Ouellette M. 2016. Plasticity of the *Leishmania* genome leading to gene copy number variations and drug resistance. *F1000Res* 5:2350. <https://doi.org/10.12688/f1000research.9218.1>.
110. da Silva MS, Hovel-Miner GA, Briggs EM, Elias MC, McCulloch R. 2018. Evaluation of mechanisms that may generate DNA lesions triggering antigenic variation in African trypanosomes. *PLoS Pathog* 14:e1007321. <https://doi.org/10.1371/journal.ppat.1007321>.
111. Li H, Marple T, Hasty P. 2013. Ku80-deleted cells are defective at base excision repair. *Mutat Res* 745–746:16–25. <https://doi.org/10.1016/j.mrfmmm.2013.03.010>.
112. Reis H, Schwebs M, Dietz S, Janzen CJ, Butter F. 2018. TelAP1 links telomere complexes with developmental expression site silencing in African trypanosomes. *Nucleic Acids Res* 46:2820–2833. <https://doi.org/10.1093/nar/gky028>.
113. van Schendel R, van Heteren J, Welten R, Tijsterman M. 2016. Genomic scars generated by polymerase theta reveal the versatile mechanism of alternative end-joining. *PLoS Genet* 12:e1006368. <https://doi.org/10.1371/journal.pgen.1006368>.
114. Kent T, Mateos-Gomez PA, Sfeir A, Pomerantz RT. 2016. Polymerase θ is a robust terminal transferase that oscillates between three different mechanisms during end-joining. *Elife* 5:e13740. <https://doi.org/10.7554/eLife.13740>.
115. Ajawatanawong P, Baldauf SL. 2013. Evolution of protein indels in plants, animals and fungi. *BMC Evol Biol* 13:140. <https://doi.org/10.1186/1471-2148-13-140>.
116. Hamilton WL, Claessens A, Otto TD, Kekre M, Fairhurst RM, Rayner JC, Kwiatkowski D. 2017. Extreme mutation bias and high AT content in *Plasmodium falciparum*. *Nucleic Acids Res* 45:1889–1901. <https://doi.org/10.1093/nar/gkw1259>.
117. Chen JQ, Wu Y, Yang H, Bergelson J, Kreitman M, Tian D. 2009. Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria. *Mol Biol Evol* 26:1523–1531. <https://doi.org/10.1093/molbev/msp063>.
118. Vanichanankul J, Taweetchai S, Yuwaniyama J, Vilaivan T, Chitnumsub P, Kamchonwongpaisan S, Yuthavong Y. 2011. Trypanosomal dihydrofolate reductase reveals natural antifolate resistance. *ACS Chem Biol* 6:905–911. <https://doi.org/10.1021/cb200124r>.
119. Timm J, Valente M, García-Caballero D, Wilson KS, González-Pacanowska D. 2017. Structural characterization of acidic M17 leucine aminopeptidases from the TriTryps and evaluation of their role in nutrient starvation in *Trypanosoma brucei*. *mSphere* 2:e00226-17. <https://doi.org/10.1128/mSphere.00226-17>.
120. Mercaldi GF, Pereira HM, Cordeiro AT, Michels PAM, Thiemann OH. 2012. Structural role of the active-site metal in the conformation of *Trypanosoma brucei* phosphoglycerate mutase. *FEBS J* 279:2012–2021. <https://doi.org/10.1111/j.1742-4658.2012.08586.x>.
121. Timm J, González-Pacanowska D, Wilson KS. 2014. Structures of adenosine kinase from *Trypanosoma brucei brucei*. *Acta Crystallogr F Struct Biol Commun* 70:34–39. <https://doi.org/10.1107/S2053230X13033621>.

OPEN

Selection of suitable reference genes for gene expression studies in myxosporean (Myxozoa, Cnidaria) parasites

Anush Kosakyan^{1*}, Gema Alama-Bermejo^{1,2,3}, Pavla Bartošová-Sojková¹, Ana Born-Torrijos¹, Radek Šíma¹, Anna Nenarokova^{1,4}, Edit Eszterbauer⁵, Jerri Bartholomew² & Astrid S. Holzer¹

Myxozoans (Cnidaria: Myxozoa) are an extremely diversified group of endoparasites some of which are causative agents of serious diseases in fish. New methods involving gene expression studies have emerged over the last years to better understand and control myxozoan diseases. Quantitative RT-PCR is the most extensively used approach for gene expression studies. However, the accuracy of the results depends on the normalization of the data to reference genes. We studied the expression of eight commonly used reference genes, adenosylhomocysteinase (AHC1), beta actin (ACTB), eukaryotic translation elongation factor 2 (EF2), glyceraldehyde-3-phosphate dehydrogenase (GAPDH), hypoxanthine-guanine phosphoribosyltransferase 1 (HPRT1), DNA-directed RNA polymerase II (RPB2), 18S ribosomal RNA (18S), 28S ribosomal RNA (28S) across different developmental stages of three myxozoan species, *Sphaerospora molnari*, *Myxobolus cerebralis* and *Ceratonova shasta*, representing the three major myxozoan lineages from the largest class Myxosporea. The stable reference genes were identified using four algorithms: geNorm, NormFinder, Bestkeeper and ΔCq method. Additionally, we analyzed transcriptomic data from *S. molnari* proliferative and spore-forming stages to compare the relative amount of expressed transcripts with the most stable reference genes suggested by RT-qPCR. Our results revealed that GAPDH and EF2 are the most uniformly expressed genes across the different developmental stages of the studied myxozoan species.

Myxozoans are a cnidarian group of obligate parasites documented mainly from fish in marine and freshwater habitats. These microscopic endoparasites have a two-host life cycle that involves an invertebrate (annelids and bryozoans) and a vertebrate host (mostly fish, few are known from other vertebrates) where infectious actinospores and myxospores are formed, respectively, serving as transmission stages in aquatic habitats^{1,2}. The current classification of the Myxozoa into classes mainly reflects spore morphology and invertebrate host types. Taxa are ranked in the class Myxosporea Bütschli 1881 according to their hardened shell valves and annelid definitive hosts while the ones with soft spore valves and bryozoan definitive hosts are representatives of the second class Malacosporea Canning, Curry, Feist, Longshaw et Okamura 2000. Myxosporea represents the largest class comprising 19 families and 67 genera while Malacosporea have only one family and two genera³.

Myxosporeans have received considerable attention since some of them are reported to cause severe fish diseases. These parasites can have a strong impact on wild and cultured fish worldwide by reducing fillet marketability and causing important mortalities in fish populations⁴⁻⁸. Given the fact that aquaculture is one of the fastest growing food sectors^{9,10} comprising an ample part of global food production, economic losses caused by parasites such as myxosporeans are of major concern^{11,12}. Furthermore, disease severity has been linked to increasing water

¹Institute of Parasitology, Biology Centre, Czech Academy of Sciences, 37005, Ceske Budejovice, Czech Republic.

²Department of Microbiology, Oregon State University, Corvallis, Oregon, USA. ³Centro de Investigación Aplicada y Transferencia Tecnológica en Recursos Marinos Almirante Storni (CIMAS), CCT CONICET – CENPAT, San Antonio Oeste, Argentina. ⁴Faculty of Science, University of South Bohemia, Ceske Budejovice, Czech Republic. ⁵Institute for Veterinary Medical Research, Centre for Agricultural Research, Hungarian Academy of Sciences, Budapest, Hungary.

*email: anna.kosakyan@gmail.com

temperatures (i.e. in *Ceratonova shasta*^{13,14}), predicting emerging numbers of these organisms in the future as a result of climate change.

In this study we have focused on three myxosporean species, *Sphaerospora molnari*, *Myxobolus cerebralis* and *Ceratonova shasta* that belong to three (sphaerosporids, oligochaete-infecting and polychaete-infecting lineages¹⁵) out of four main myxozoan phylogenetic lineages and transcriptomic data are available (¹⁶, Alama-Bermejo *et al.* submitted, Hartigan *et al.* submitted). These species are considered serious pathogens for highly commercialized fishes such as cyprinids and salmonids which represent a significant proportion of worldwide aquaculture production. *Sphaerospora molnari* causes respiratory and osmoregulatory failure in host's gill epithelia¹⁷, while proliferative blood stages induce a massive systemic inflammatory response¹⁸. It was also shown that *S. molnari* may be an important co-factor for swim bladder inflammation in carp, the disease responsible for up to 100% mortalities of carp fingerling stocks in Central Europe¹⁹. The invertebrate definitive host of *S. molnari* is unknown. *Myxobolus cerebralis* causes serious damage to farmed and wild salmonid fish populations worldwide. It is responsible for whirling disease, a condition caused by infection of the hosts central nervous system and cartilage resulting high mortalities⁶. The life cycle of *M. cerebralis* involves the oligochaete *Tubifex tubifex* and salmonids as vertebrate hosts. *Ceratonova shasta* is a serious pathogen of wild and cultured salmonids in the Pacific Northwest of North America, including endangered Coho and Chinook salmon. *C. shasta* causes intestinal enteritis with up to 100% mortality in certain populations. The definitive invertebrate host of *C. shasta* is the freshwater polychaete *Manayunkia speciosa*²⁰.

There are currently no disease control methods for myxozoans in general, as no vaccines or commercial treatments for fish destined for human consumption are available. In order to design efficient methods for prediction and control of myxozoan diseases it is important to explore genes that these parasites use for invasion of and survival within their host, e.g. genes involved in immune evasion, since they are potential candidates for targeted antiparasitic treatments and vaccine development. As a result of this need, the number of gene expression studies are presently expanding considerably in order to predict and control such functional genes.

RT-qPCR (reverse transcriptase quantitative PCR) is one of the most rapidly incorporated techniques in scientific studies. Its application in mRNA quantification has grown from ~8% to ~73–88% in the last decade²¹.

Being considered highly sensitive, RT-qPCR is one of the most extensively used approaches for gene expression studies in all organisms^{22–24}. To achieve accurate gene expression results, it is critical that RT-qPCR results are normalized to an internal control, since gene expression can be influenced by different factors, i.e. variation in the amount of starting material, differences in RNA contents between cells or developmental stages, technical variability, and transcription efficiency. Traditionally, housekeeping genes (hereafter HKGs) are used as internal control. HKGs are present in all cell types because they are necessary for basic cell survival. HKGs commonly used as internal controls include beta actin, glyceraldehyde-3-phosphate dehydrogenase, several ribosomal genes such as 18S rRNA, 28S rRNA and eukaryotic elongation factors. Due to their key roles in metabolism, cytoskeleton and ribosome structure, the mRNA/rRNA synthesis of these genes was considered to be stable or uniformly expressed in various tissues, during ontogeny and development, even under different treatments^{19–21} and thus these genes were considered good reference genes (hereafter RGs).

However, it was shown that HKGs independent of organism do not always perform as good RG, and their expression may be differentially regulated and vary under certain experimental conditions. That is why it is highly recommended to validate the HKGs for each organism and study before performing gene expression studies²³.

For cnidarians in general, data on RGs are scarce, although few differential expression studies were performed^{25–30}. For myxozoans, comprehensive gene expression studies are inexistent and only three reports study parasite gene expression^{31–33} and mainly rely on RGs that were “used in previous publications” (e.g.³¹), or the validation of RGs was focused on a limited part of parasite development (e.g. early intrapiscine development studied by Eszterbauer *et al.*³²). Myxozoans are some of the oldest metazoan parasites with an extremely accelerated evolutionary rate and high heterogeneity across genes¹⁵, and likely functional derivation of genes. Therefore, it can be expected that genes that serve as RGs in other organisms are not constantly expressed during the complex life cycle of myxozoans and across the different developmental stages, within the vertebrate and invertebrate hosts. 18S rDNA is presently the most commonly used gene region for phylogenetic studies¹⁵ and especially for PCR and qPCR based detection and quantification assays (e.g.^{34–36}), since rDNA occurs in tandem repeats and multiple copies in the genome, however it has not previously been tested as a RG. Considering the increasing need to understand and evaluate gene expression in myxozoans, our aim was to investigate the suitability of candidate reference genes in different developmental stages of three myxozoan species. Furthermore, we want to propose “optimal” reference genes that can be used in future myxozoan gene expression studies aimed at the discovery of functional target proteins to control emerging myxozoan diseases.

Material and Methods

Parasite collection. For each species different developmental stages (*S. molnari*) and different life cycle stages (*M. cerebralis* and *C. shasta*) were isolated from fish and definitive worm hosts (Table 1).

S. molnari proliferative, presporogonic blood stages were collected from a laboratory line that has been cycled (2+ years) from fish to fish by intraperitoneal injection of parasites into specific parasite-free (SPF) common carp (*Cyprinus carpio*) (methodology detailed in¹⁸). *S. molnari* blood stages (n = 5 fish) were concentrated and co-isolated with host white blood cells from whole blood of carp, by centrifugation for 5 minutes at 3500 rpm in heparinized hematocrit tubes¹⁸. Spore-forming stages (infected gills, n = 5) were obtained from carp held at the recirculation system of the Research Institute for Fisheries and Aquaculture (Szarvas, Hungary).

M. cerebralis actinospores used for exposure trials originated from *Tubifex tubifex* cultures maintained in the laboratory of the Institute for Veterinary Medical Research (IVMR), Budapest, Hungary, over several years. SPF rainbow trout, *Onchorhynchus mykiss* (Kamloops strain) was obtained from the Lillafüred Trout Hatchery, Hungary (yolk sac stage) and reared at the IVMR. Fish were infected individually with 5000 freshly filtered

Parasite species	Host species	Origine of samples	Parasite developmental stage	Host tissue	Number of samples
<i>Sphaerospora molnari</i>	<i>Cyprinus carpio</i>	Malá Outrata pond, CZ	presporogonic blood stage	blood	5
	<i>Cyprinus carpio</i>	Malá Outrata pond, CZ	non infected	blood	2
<i>Sphaerospora molnari</i>	<i>Cyprinus carpio</i>	Szarvas, HU	sporogonic stage	gills	5
	<i>Cyprinus carpio</i>	Szarvas, HU	non infected	gills	2
<i>Myxobolus cerebralis</i>	<i>Oncorhynchus mykiss</i>	Inst for Veterinary Med Res, Budapest, HU	sporogonic stage	cartilage (head)	4
	<i>Oncorhynchus mykiss</i>	Inst for Veterinary Med Res, Budapest, HU	non infected	cartilage (head)	2
<i>Myxobolus cerebralis</i>	<i>Tubifex tubifex</i>	Inst for Veterinary Med Res, Budapest, HU	triacinomyxon stage	whole worm	4
	<i>Tubifex tubifex</i>	Inst for Veterinary Med Res, Budapest, HU	non infected	whole worm	2
<i>Ceratonova shasta</i>	<i>Oncorhynchus mykiss</i>	Roaring River Hatchery, Scio, OR, USA	mix of presporogonic and sporogonic stages	intestine	3
	<i>Oncorhynchus mykiss</i>	Roaring River Hatchery, Scio, OR, USA	non infected	intestine	2
<i>Ceratonova shasta</i>	<i>Oncorhynchus mykiss</i>	Roaring River Hatchery, Scio, OR, USA	mix of presporogonic and sporogonic stages	ascites	3
	<i>Oncorhynchus mykiss</i>	Roaring River Hatchery, Scio, OR, USA	non infected	ascites	2
<i>Ceratonova shasta</i>	<i>Manayunkia</i> sp.	Fryer Aquatic Animal Health Lab (OSU), USA	tetractinomyxon stage	whole worm	3
	<i>Manayunkia</i> sp.	Fryer Aquatic Animal Health Lab (OSU), USA	tetractinomyxon stage	whole worm	2

Table 1. Sampling details of selected parasites across different developmental stages and non- infected host used as control.

actinospores according to³⁷. From infected fish, pieces of skulls containing myxospores and sporogonic plasmodia (spore-forming stages) were collected 90 days' post exposure (n = 4). Laboratory *T. tubifex* cultures were exposed with spores isolated from the head cartilage as per³⁸. Worms (n = 4) infected with triactinomyxon spore-forming stages were collected 100 days' post exposure.

The species composition of naive worm cultures was regularly checked by DNA sequencing and microscopy, and worm specimens with long hair chaetae (which all belong to *Tubifex tubifex* s.l. in the culture) were selected for individual exposure.

Ceratonova shasta was collected from ascitic fluid of the abdominal cavity and from infected intestines of rainbow trout infected with genotype IIR (n = 3). Naive rainbow trout were from Roaring River Hatchery strain (Scio, OR, Oregon Department of Fish and Wildlife) and they were infected by an intraperitoneal injection of ascites collected from an infected rainbow trout that was previously exposed in the Williamson River, Oregon, USA. Fish were held at 18 °C in 100 L tanks at the Aquatic Animal Health Laboratory at Oregon State University (AAHL, OSU). Fish were sampled when developing typical clinical signs of enteronecrosis⁶. A wet mount of ascites was examined using a Zeiss 47 30 28 light microscope and the presence of different developmental stages (plasmodia and spores) was confirmed. Genotype was confirmed using the ITS rDNA region³⁹. Fish were euthanized by an overdose of buffered MS-222 (tricaine methanesulfonate; Argent Laboratories). Ascites was collected with a sterile syringe. Intestine was removed by dissection. Fluid and tissues were flash frozen in liquid nitrogen and kept at -80 °C. The infection had been achieved by transmission of ascites stages from fish to fish by intraperitoneal injection⁴⁰.

Manayunkia sp. worms (n = 3) infected with actinospores of the same genotype were obtained from laboratory cultures (methodology of⁴¹ at the John L. Fryer Aquatic Animal Health Lab (OSU). Worms originated from the Upper Klamath River and were regularly seeded with myxospores from IIR transfected rainbow trout⁴². RNA in blood, gills, skull pieces and worms was stabilized in 100 µl of RNAProtect Cell Reagent (Qiagen) and stored at -80 °C prior to RNA extraction. Intestine and ascites infected with *C. shasta* were flash-frozen in liquid nitrogen and stored at -80 °C.

Ethics statement. Fish manipulation and sampling techniques were performed in accordance with Czech legislation (Protection of Animals Against Cruelty Act No. 246/1992) and approved by the Czech Ministry of Agriculture. Rainbow trout sampling at Oregon State University (OSU) was carried out in accordance with the recommendations of OSU - Institutional Animal Care and Use Committee (IACUC). The protocol was approved by ACUP #4666. For rainbow trout exposure to *M. cerebralis*, the Hungarian Scientific Ethical Committee on Animal Experimentation provided approval (PEI/001/4087-4/2015).

RNA extraction and reverse transcription. Total host + parasite RNA for all samples with exception of *Manayunkia* worms, was isolated using the Nucleospin RNA Kit (Macherey-Nagel) following manufacturer's instructions. RNA from *Manayunkia* worms was isolated using guanidine/thiocyanate/phenol/chloroform extraction method⁴³ to ensure higher concentrations of RNA compared to the column-based RNA extraction methods. A DNase digestion step ensuring elimination of genomic DNA was included into the protocol of the Nucleospin RNA Kit (manufacturer's instructions). For *Manayunkia* samples, DNA was removed using the DNAFree Kit (Invitrogen). RNA concentration and purity were checked using a Nano Drop - 1000 Spectrophotometer (Thermo Fisher Scientific Inc.). All RNA samples with 260/280 ratio in range of 1.9–2.0, and 260/230 ratio in range of 2.0–2.4 were chosen for cDNA synthesis. Approximately 500 ng RNA was used as an input for synthesis of 20 µl of cDNA using the Transcriptor High Fidelity cDNA synthesis Kit (Roche) following the manufacturer's protocol.

Gene Ids	Protein encoded	Generally accepted function	Accepted reference gene for	Used as RG for myxozoans
ACTB	Beta actin	Cytoskeletal structural protein	Commonly accepted reference gene ^{25,32,86}	<i>Myxobolus cerebralis</i> ^{32,33}
AHC1	Adenosylhomocysteinase	Homocystein synthesis protein	Corals ²⁵	
EF2	Eukaryotic Translation Elongation Factor 2	Nascent protein synthesis protein	Commonly accepted reference gene ^{67,68}	
GAPDH	Glyceraldehyde-3-Phosphate Dehydrogenase	Metabolic protein (glycolytic enzyme)	Commonly accepted reference gene ²³	
HPRT 1	Hypoxanthine-Guanine Phosphoribosyltransferase 1	Purine nucleotide synthesis protein	Commonly used for humans ²²	
RPB2	DNA-directed RNA polymerase II	RNA polymerase II transcription machinery protein	Humans ⁸⁷	
18S rRNA	18S ribosomal RNA gene SSU	Ribosome structural protein	Commonly used reference gene ^{23,88}	<i>Myxobolus cerebralis</i> ³¹
28S rRNA	28S ribosomal RNA gene LSU	Ribosome structural protein	Commonly used reference gene ⁸⁹	

Table 2. Details of selected candidate reference genes.

Candidate reference gene selection and data mining. A list of eight commonly used cnidarian and other metazoan candidate reference genes were selected for this study: adenosylhomocysteinase (AHC1), beta actin (ACTB), eukaryotic translation elongation factor 2 (EF2), glyceraldehyde-3-phosphate dehydrogenase (GAPDH), hypoxanthine-guanine phosphoribosyltransferase 1 (HPRT1), DNA-directed RNA polymerase II (RPB2), 18S ribosomal RNA (18S), 28S ribosomal RNA (28S) (Table 2). Initially, this list included six more genes used as reference genes for cnidarians and other metazoans, such as eukaryotic translation factor 1 (EF1), NADH dehydrogenase iron-sulfur protein 2 ubiquinone (NADH), heat shock protein 70 (HSP70), ribosomal protein L11 (RPL11), TATA-Box Binding Protein Associated Factor 6 (TAF6), PHD finger protein 8 (PHF8). However, these genes were later excluded from the study/analysis, because either we were not able to find suitable homologues of these genes in our transcriptome/s, or primer design/ PCR was not successful. The eight candidate reference genes were mined from their respective parasite transcriptome data (RNA sequences) or from DNA sequences available in GenBank or at private databases. All available homologue amino acid sequences of these genes (GenBank) from common representatives of cnidarians such as *Acropora tenuis*, *Aurelia aurita*, *Hydra vulgaris*, *Hydra magnipapillata*, *Nematostella vectensis*, *Polypodium hydriforme*, and different myxozoan species (*C. shasta*, *Kudoa iwatai*, *M. cerebralis*, *Sphaerospora dicentrarchi*, *Thelohanellus kitauei*, *Buddenbrockia sp.*, etc.) were combined for queries. The search was performed using the tBLASTn algorithm with the e-value cutoff set to 10^{-10} . The top hits (highest e-value) were analyzed using the NCBI conserved domains platform (<https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>, to confirm their identity). To re-confirm the myxozoan origin of the mined sequences phylogenetic trees including other metazoan taxa (cnidarians, fish, etc.) were reconstructed using maximum likelihood methods in RAxML web-server (<https://raxml-ng.vital-it.ch/#/>). Details on chosen sequences are included in Suppl. Mat. 1.

Primer design and specificity of PCR. Gene-specific primers were designed to amplify short 70–150 bp regions suitable for RT-qPCR assays (Table 3). Primer pairs were designed with optimal T_m at 58–60 °C and GC content between 45–50%, using the NCBI online primer-design tool (<https://www.ncbi.nlm.nih.gov/tools/primer-blast/>). All primers were tested for specificity using conventional PCR prior to performing RT-qPCR. Details on PCR conditions are described in Suppl. Mat. 5. Primer specificity was determined by obtaining single amplicons of the expected size from infected samples and no amplification in uninfected fish and worm samples (indicating that primers are not annealing with fish cDNA). Controls without reverse transcriptase (–RT) were tested for genomic DNA contamination. The presence of infection was confirmed, and parasites identified microscopically and by specific published PCR assays^{31,44,45}. The identity of PCR products was confirmed by sequence comparison. RT-qPCR primer specificity was also checked by running melting curve analysis (see Suppl. Mat. 3).

Quantitative real-time PCR. RT-qPCR was performed using the LightCycler[®] 480 Real-Time PCR System (Roche). Reactions of 25 µl were comprised of 12.5 µl of FastStart Universal SYBR Green PCR Master Mix (Roche, Germany, 2X conc.), 1 µl of each forward and reverse primer (10 µM conc., 0.4 µM final conc.), 5.5 µl of PCR grade water, and 5 µl of cDNA (generally at ~150–170 ng/µl dilutions). The cycling conditions were denaturation at 95 °C for 5 min, followed by 50 cycles of 95 °C for 10 s, 58 °C for 10 s and 72 °C for 10 s. Melting curve analysis was performed after each cycle to ensure primer specificity. All samples were amplified in technical triplicates and a mean value was calculated. Four (*M. cerebralis*) to five (*S. molnari*) biological replicates were used for each sample, with the exception of *C. shasta* (only 3 replicates available). qPCR efficiency was predicted for each gene based on the slope of a linear regression model⁴⁶ using a series of 5-fold dilutions (1:5, 1:25, 1:125, 1:625). Standard curves were built using Roche LightCycler 480 Software version 1.5.0 SP4. Generally, for best amplification results efficiency ranges of 90–110% and standard curve slopes of –3.58 to –3.10 were considered optimal⁴⁷.

Ranking and quantitative analysis of reference genes. Differential expression levels and abundance of candidate reference genes within the sample was analyzed by a direct comparison of C_q (quantification cycle) values (Fig. 1, Suppl. Mat. 2). The stability of the candidate reference genes was analyzed using four algorithms: ΔC_q , NormFinder, geNorm, and BestKeeper. The comparative ΔC_q method manually compares relative expression of ‘pairs of genes’ within each sample. If the ΔC_q value between the two genes remains constant when analyzed in different samples it means that either both genes are expressed at relatively constant rates among those samples, or they are co-regulated (here we assume the stability of both genes)⁴⁸.

Organism	Gene	Primer sequence (5'-3')	Amplicon length (bp)	Melting T°(C)	GenBank access. numbers*
<i>Sphaerospora molnari</i>	ACTB	F: AATCCACGAGACCACCTTCG	149	59.75	see Suppl. Mat. 1
		R: CAGCAGCCAAACCGGTGATA		60.68	
	AHC1	F: TTCCCATGGTGTGCGAGAAA	138	58.94	see Suppl. Mat. 1
		R: TCAATGACACCTCGAACACAGT		59.9	
	EF2	F: TCCGGCAGGCAAGAAGGTTT	140	62	see Suppl. Mat. 1
		R: CCAAGTTGGATACGGATTACGAGT		60.44	
	GAPDH	F: TATCGACCTGGCCGTACTG	118	59.63	see Suppl. Mat. 1
		R: GTTGCTGCTGTCAATGACCC		59.9	
	HPRT 1	F: TCTCATCTGTGACCGTGCTC	84	59.47	see Suppl. Mat. 1
		R: ACGCACAAAACTCGGATCTG		59.47	
	RPB2	F: ATTAGTTACGGTGCCGAGG	143	59.54	see Suppl. Mat. 1
		R: GCTGTGACATGGAAGATGCG		59.62	
18S rRNA	F: ATCCCAGGTCGTATCCGCTA	73	59.89	see Suppl. Mat. 1	
	R: ACTGCCCTGTTGATGCGATT		60.32		
28S rRNA	F: ATCTGCTCGCACCTCATACG	143	59.97	see Suppl. Mat. 1	
	R: CCGAGTTTGCTTGCGTTACC		60.11		
<i>Myxobolus cerebralis</i>	ACTB	F: TTGCCTGATGGTCAGGTGAT	110	59.01	AY156508.2
		R: AGTGTCTCGTGAAGTCCACTG		59.66	
	AHC1	F: GTTCAGCGTCGCTAAGAGGA	124	59.83	GBKL01003454.1
		R: GCCCGAGAGACACAGTCATC		60.18	
	EF2	F: ATGGATCCGGGCCTAACCTT	149	60.7	GBKL01021688.1
		R: CAAGTCCAGACGAACACCCC		60.6	
	GAPDH	F: GTGGCAAACCCGCAACTAA	95	59.61	GBKL01017634.1
		R: TGTGCGTCGACAACTGGAT		60.25	
	HPRT 1	F: TGGTCTCCTGGTGAAGAAA	119	59.16	GBKL01050483.1
		R: GAGGTCGTCCATCCAGTTT		59.39	
	RPB2	F: AATGGAGGGCTGGCTAAACG	127	60.39	GBKL01027608.1
		R: TAATCCGATGTCAGGGCACC		59.53	
18S rRNA	F: TAGAGTGTGCCGAACGAGTC	85	59.48	EF370479.1	
	R: GGTCCCAAGGCATCATGACA		60.03		
28S rRNA	F: AGTCGAAGTAGAGCAGCGTG	141	59.83	AY302740.1	
	R: CATCCTCAGGGATGCACTGT		59.45		
<i>Ceratonova shasta</i>	ACTB	F: GTCGGCAATTCTGGGTACA	149	60.04	see Suppl. Mat. 1
		R: TCCAACCGGCATTTTAGGA		57.41	
	AHC1	F: TTCGGTTACCACGACTCGGC	82	62.47	see Suppl. Mat. 1
		R: TGTAGTGGGTGGCTATGGTGA		60.55	
	EF2	F: CTGGATTCCAATGGGCAACT	147	58.14	KM392431.1
		R: AAATAACTCTTCGAGCAGTAGGT		57.34	
	GAPDH	F: TGGGGCTAAACAGTTGGTGG	152	60.18	see Suppl. Mat. 1
		R: GTGGACATTTGAAAGGAGGCG		59.8	
	RPB2	F: TGGAGGTGAAGGTACGTGT	156	58.58	see Suppl. Mat. 1
		R: TCTGCCCTTTATAGGACGA		57.54	
	18S rRNA	F: CCAAGTTGGTCTCTCCGTGA	121	59.32	AF001579.1
		R: CAAATTAAGCCGAGGCTCC		59.9	
28S rRNA	F: ACGTGAAACCGTTAACATGGA	132	58.16	FJ981818.1	
	R: CCACTGGCCTTGAAGATTGT		58.08		

Table 3. Genes and their primer sequences used in this study. *Accession numbers are provided for the gene sequences that are available in GenBank/either mined from transcriptomic data under review, but see sequences in Suppl. Mat. 1.

NormFinder⁴⁹ was performed using original Microsoft Excel-based software. It determines the stability of the candidate genes based on an estimate of inter- and intragroup variation. It calculates the most stably expressed candidate genes and suggests two of them as references.

geNorm was performed using the qbase + package software⁵⁰. This program is based on the assumption that if the ratios between samples are uniformly expressed, non-normalized target genes should remain regular. The genes with the most irregular expression are excluded from further analysis while the last two remaining genes are selected as the most stable⁵¹. We used two values to interpret geNorm results: (1) geNorm M (geNorm expression

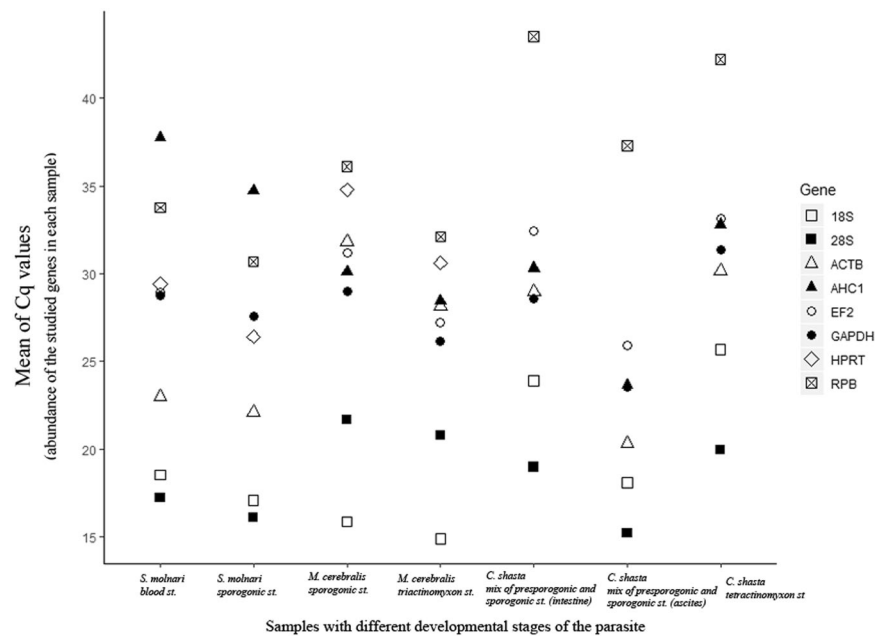


Figure 1. RNA transcription levels of candidate reference genes (in absolute Cq values) representing the abundance of the studied genes in each sample.

stability value of reference genes, lowest M value indicates higher stability); and (2) geNorm V (pairwise variation). geNorm V further determines the optimal number of reference genes to be used in subsequent analyses. A $V_{n/n+1}$ value is shown for every comparison between two consecutive numbers (n and $n+1$) of candidate reference genes. As a general guideline (www.qbaseplus.com, qbase + manual, rev2017.04.27) it is stated that the benefit of using an extra ($n+1$) reference gene is limited as soon as the $V_{n/n+1}$ value drops below the 0.15 threshold, indicated with a horizontal line (Fig. 2).

BestKeeper was performed using the original Microsoft Excel-based formulas⁵². It calculates the standard deviation of the Cq value between the whole data set, and the gene with the lowest standard deviation (SD) is proposed as most suitable.

Finally, we used RefFinder (<https://www.heartcure.com.au/reffinder/?type=reference> accessed at 25 June 2019), a comprehensive software platform which integrates all four algorithms providing an overall ranking of the used genes.

Transcriptomic data analyses. *De novo* transcriptome assemblies of *S. molnari* (unpublished) were used to observe expression of candidate reference genes in blood and sporogonic stages of parasite. We used *S. molnari* 11 samples (5 from blood stages and 6 from sporogonic stages) based transcriptomic data to estimate transcript expression values (TPM: Transcripts Per Million) using the Salmon software⁵³. These TPM expression values were scaled and served to generate a cross-sample normalized TMM gene expression matrix (TMM: trimmed mean of M-values: scaling normalization that aims to account for differences in total cellular RNA across all samples), using the Trinity package^{54,55}. We extrapolated TMM values for eight candidate genes expression values from the gene expression matrix and compared it across the 11 samples manually. Average values for each developmental stage were calculated. The most stable gene was considered the one for which the ratio between the average values of both developmental stages was closest to 1 (Table 4).

Results

PCR specificity and primer efficiency. Primer specificity was confirmed by obtaining single amplicons of the expected size, together with negative results in uninfected fish and worm samples. Primer specificity was also confirmed based on the occurrence of a single peak in the melting curve (Suppl. Mat. 3). Absence of genomic DNA contamination was confirmed by no amplification in $-RT$ samples. The efficiency of our candidate RG primers in the present study ranged from 88 to 129%, which slightly surpasses the acceptable optimum range (90–110%). However, we obtained similar efficiencies for the given genes in two different developmental stages of parasite.

Cq data. Transcript abundance of each gene within each biological replicate was roughly estimated from the raw Cq values. The most abundantly expressed genes were 28S and 18S for all three studied species with Cq ranging about 14–25. The least expressed genes were AHC1 for *S. molnari* (Cq > 34) and RPB for *M. cerebralis* (Cq > 32) and *C. shasta* (Cq > 37.3). The rest of the genes fell in the range of Cq = 22–34 (Fig. 1 and Suppl. Mat. 2).

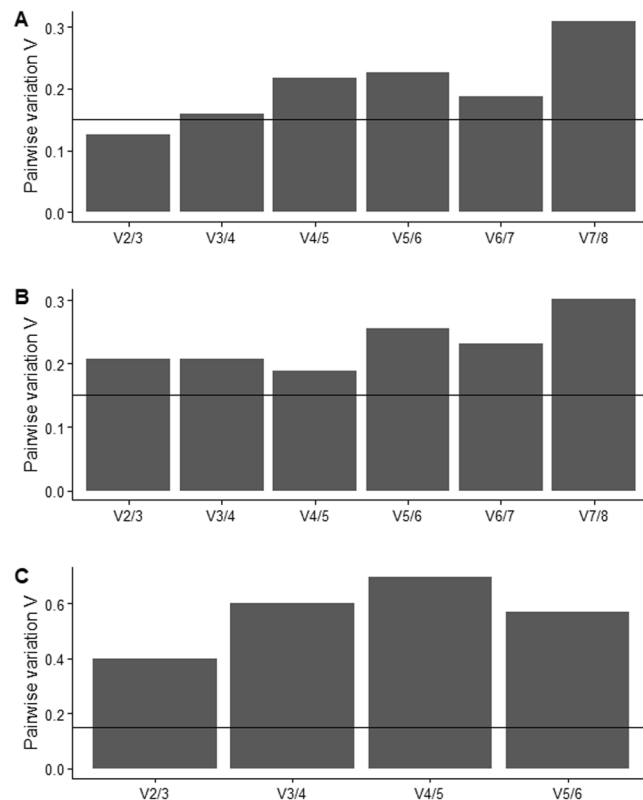


Figure 2. GeNorm pairwise variation (with threshold value = 0.15) suggesting optimal number of reference genes for normalization for A. *Sphaerospora molnari*, B. *Myxobolus cerebralis*, C. *Ceratonova shasta*.

Stability of candidate genes. For *S. molnari*, the expression of EF2 and GAPDH was shown to be the most stable among the eight studied genes according to ΔCq method (lowest average SD values 1.08 for EF2 and 1.12 for GAPDH), NormFinder (with lowest stability value = 0.18 for EF2 and 0.38 for GAPDH) and geNorm (lowest M = 0.37 for both genes). We obtained pairwise variation (geNorm V-value) $V_{2/3} < 0.15$, indicating that in this case 2 genes are sufficient for normalization, and that the additional inclusion of more reference genes will not provide a significant improvement for the normalization of target genes (Fig. 2A). While GAPDH was ranked in the second most stable place ($SD \pm Cq = 2.15$), EF2 was ranked in the third most stable place ($SD \pm Cq = 2$) by BestKeeper (Table 5, Suppl. Mat. 4). The same genes occur in the first 4 positions across all the algorithms: geNorm: EF2 > GAPDH > ACTB > 28S, ΔCq method: EF2 > GAPDH > ACTB > 28S, NormFinder: EF2 > GAPDH > ACTB > 28S, BestKeeper: ACTB > GAPDH > 28S > EF2. To obtain a comprehensive ranking and summary of all the algorithms used, we loaded our raw data to the RefFinder web-based platform, which includes four above mentioned algorithms (Fig. 3). Overall EF2, GAPDH, and ACTB were observed as the most stable genes by RefFinder (Fig. 3A).

We observed a similar pattern for *M. cerebralis*. geNorm suggested ACTB, EF2 and GAPDH with the lowest M value ($M = 0.41, 0.41$ and 0.57 , respectively). Ge Norm pairwise variation showed middle stability, suggesting to consider 4 genes for optimal normalization (Fig. 2B). ΔCq method (with lowest average SD values 1.11 for GAPDH, 1.12 for ACTB and 1.17 for EF2) and NormFinder (with lowest stability value = 0.27 for GAPDH, 0.56 for ACTB and 0.65 for EF2) also suggested the same genes. However, Bestkeeper's ranking was different as GAPDH occurred in the fourth place ($SD \pm Cq = 1.69$), ACTB in the fifth place ($SD \pm Cq = 1.95$) and EF2 in the sixth place ($SD \pm Cq = 2$). Overall, out of the eight *M. cerebralis* genes studied, the following genes ranked in the first 4 positions: geNorm: ACTB > EF2 > GAPDH > RPB2, ΔCq method: GAPDH > ACTB > EF2 > RPB2, NormFinder: GAPDH > ACTB > EF2 > RPB2, BestKeeper: 28S > 18S > AHC1 > GAPDH. RefFinder suggested GAPDH, EF2 and ACTB as the most stable reference genes in comprehensive ranking (Fig. 3B).

For *C. shasta* the combination of the genes used in the analysis was slightly different. We obtained no expression or very low expression for HPRT1 gene in all samples, and very low expression ($\Delta Cq > 40$) of RPB gene in worm samples and thus these two genes were excluded from the final gene stability analysis. geNorm showed that AHC1 and ACTB and GAPDH had the lowest M value ($M = 1.27, 1.27$ and 1.54 respectively). However, pairwise variation (geNorm $V > 0.15$) could not determine the optimal number of genes to be used for normalization. (Fig. 2C). ΔCq method (with lowest average SD values 2.63 for GAPDH, 2.79 for EF2 and 2.86 for AHC1) and NormFinder (with lowest stability value = 0.94 for EF2, 1.15 for GAPDH and 1.92 AHC1) suggested EF2 and GAPDH, however Bestkeeper's ranking was different as 28S occurred in the second place ($SD \pm Cq = 1.97$), while GAPDH and EF2 were in fourth place ($SD \pm Cq = 2.87$), and fifth place (3.06). Overall, out of the studied *C. shasta* genes the following genes ranked in the first 4

Biological replicate	Sporogonic stage					Sporogonic average	Blood stage						Blood average	Ratio between two stages (Sp. average/Bl. average)
	biol repl1 (fish 1)		biol repl2 (fish 2)		biol rep 3 (fish 3)		biol repl1 (fish 1)		biol repl2 (fish 2)		biol rep 3 (blood stage mix of several fishes)			
Technical Replicate	tech repl 1	tech repl 2	tech repl 1	tech repl 2	no tech repl		tech repl 1	tech repl 2	tech repl 1	tech repl 2	tech repl 1	tech repl 2		
Sample name	3A	3B	4A	4B	5A		RNA1_L001	RNA1_L002	RNA2_L001	RNA2_L002	RNA3_L001	RNA3_L002		
GAPDH	81.29	73.27	62.23	77.49	47.42	68.34	29.83	29.25	83.18	84.50	106.20	118.88	75.31	0.91
EF2	269.91	230.57	193.22	282.79	143.82	224.06	74.14	73.91	331.30	331.03	313.90	298.76	237.17	0.94
18S rRNA	225.78	408.70	489.16	372.05	851.90	469.52	1301.10	1320.74	1154.50	1148.18	4751.65	4818.87	2415.84	0.19
HPRT1	1038.46	1134.00	1121.06	1068.91	1201.93	1112.87	73.16	70.86	371.42	371.66	438.28	470.23	299.27	3.72
RPB2	8.59	6.75	6.41	7.78	4.55	6.82	5.97	4.47	5.25	4.02	5.18	4.87	4.96	1.37
ACTB	3785.16	2868.36	2975.15	2916.40	3780.47	3265.11	4167.18	4235.62	10470.21	10430.10	10841.77	10873.28	8503.03	0.38
28S rRNA	273.61	504.40	392.36	390.14	1058.62	523.83	92.39	93.65	193.23	187.74	364.82	365.67	216.25	2.42
AHC1	22.11	23.63	18.93	13.89	40.43	23.80	1.93	1.84	5.80	3.98	0.29	1.39	2.54	9.37

Table 4. TMM expression values of candidate HKGs for sporogonic and blood stages of *Sphaerospora molnari*. The genes for which ratio between sporogonic and blood stages are closer to 1 are considered the most uniformly expressed between two stages.

Genes	GeNorm	NormFinder	BestKeeper	Δ Ct	Comprehensive ranking	TMM
<i>Sphaerospora molnari</i>						
ACTB	2	4	2	4	3	4
AHC1	5	7	7	6	6	8
EF2	1	1	5	1	1	1
GAPDH	1	2	3	2	2	2
HPRT1	4	5	6	5	7	7
RPB2	7	8	8	8	8	3
18S rRNA	6	6	4	7	5	5
28S rRNA	3	3	1	3	4	6
<i>Myxobolus cerebralis</i>						
ACTB	1	2	5	2	2	
AHC1	5	5	3	6	6	
EF2	1	3	6	3	3	
GAPDH	2	1	4	1	1	
HPRT1	4	6	8	5	8	
RPB2	3	4	7	4	5	
18S rRNA	7	8	2	8	7	
28S rRNA	6	7	1	7	4	
<i>Ceratonova shasta</i>						
ACTB	1	4	5	4	6	
AHC1	2	3	4	3	3	
EF2	4	1	3	2	2	
GAPDH	3	2	2	1	1	
HPRT1	Excluded from analysis					
RPB2	Excluded from analysis					
18S rRNA	8	6	6	6	5	
28S rRNA	7	4	1	5	4	

Table 5. Comprehensive ranking of studied genes using a combination of four algorithms. TMM ranking is based on normalized transcript expression values from NGS data. Lower ranking values indicate higher gene stability. Additionally, gene stability values generated by each algorithm are given in Suppl. Mat. 4.

positions: geNorm: AHC1 > ACTB > GAPDH > EF2, Δ Cq: GAPDH > EF2 > AHC1 > ACTB, NormFinder: EF2 > GAPDH > AHC1 > ACTB, BestKeeper: 28S > GAPDH, EF2 > AHC1 (Table 5). RefFinder suggested EF2, GAPDH and AHC1 as the most stable reference genes in comprehensive ranking (Fig. 3C).

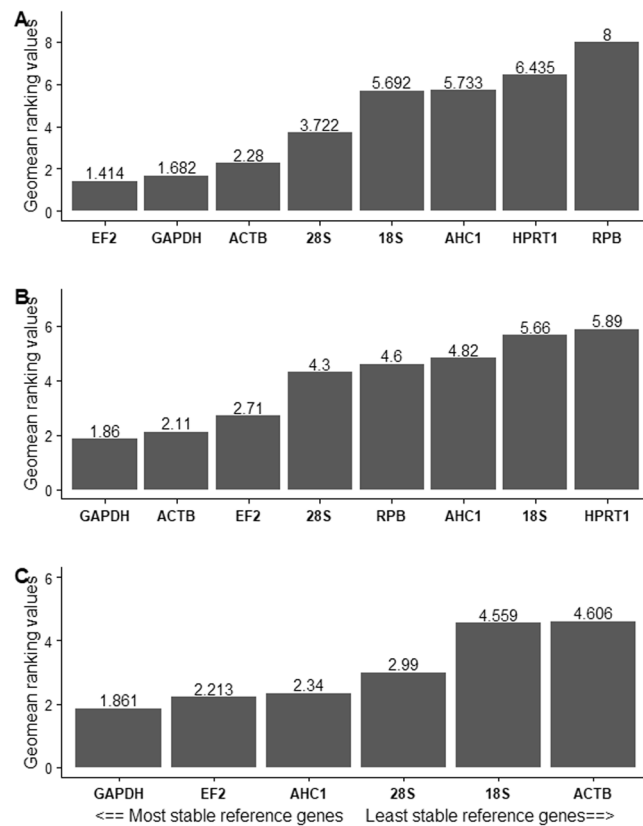


Figure 3. Stability of candidate reference genes expression for studied myxozoan species based on RefFinder comprehensive ranking: (A). *Sphaerospora molnari*, (B). *Myxobolus cerebralis*, (C). *Ceratovona shasta*. Y axis represents Genes Geomean of ranking values (lower value indicates higher stability). In X axis genes are ordered from high to low expression stability.

Differential gene expression from transcriptome data. We have obtained TMM values for candidate reference genes in two *S. molnari* developmental stages (pre-sporogonic blood stage and sporogonic stage). We have calculated the closest to 1 value for EF2 (ratio between two stages = 0.94) and GAPDH (ratio between two stages = 0.91) indicating these two genes as the most stable of the eight studied genes (see Tables 4 and 5).

Discussion

Expectations for real-time RT-qPCR are high as it serves as a first step of generating a data which will be a reference for the next steps of research applications (i.e. studies involving gene expression data).

While, numerous studies are based on qPCR data, in the past, only few reference gene validation studies were conducted. Many papers would use only a single gene as a reference, without verification of its utility under the used experimental conditions. This was especially common for majority of the articles concerning the analysis of RNA transcripts published in high impact journals in late 1990s and early 2000s in different organisms including myxozoans^{23,24,31–33}.

In order to obtain reliable results, reference genes normalization and its rational interpretation are essential. It is complicated to determine if a two-fold variation in gene expression is of biological importance because this genetic variation can be triggered by intrinsic noise of biochemical reactions. Discrepancies with regard to organism strains, experimental design, and algorithms calculating differential expression further add to this noise.

While not everything can be controlled for, the first step for producing meaningful (true) data is careful evaluation of reference genes.

Since previous data evaluating the stability of reference genes in myxozoans are missing, in this study, we evaluated eight candidate genes for their suitability as a reference for future RT-qPCR assays in gene expression studies of myxozoans. We designed a comprehensive setup for testing these genes in a comparative approach by using RNA extractions from different developmental stages of the parasites' dixenous life cycle, using three species from different phylogenetic lineages, covering the fields of biological and technical replicates and different calculation algorithms and methods by using RT-qPCR and transcriptomic data. Here, we discuss the parameters we used to ensure the best choice of reference genes, possible pitfalls that should be taken into consideration before final conclusions, and we provide recommendations for future RT-qPCR studies in this unique group of highly derived cnidarian parasites.

Stability of the candidate reference gene and choice of algorithm. We used four well-accepted algorithms, geNorm, ΔCq , NormFinder, and BestKeeper in combination with TMM expression values mined in transcriptome expression data to evaluate the stability of the examined genes. Since these algorithms have

different calculating approach, it might expect that the rankings of candidate genes could be different depending on the software applied. In previous studies we observed that the results produced by BestKeeper can oppose those of geNorm and NormFinder^{56,57}.

Each approach has its strengths and weaknesses and there is no commonly accepted opinion on which one is the best. A consensus ranking of RGs is useful as it combines the data obtained from different algorithms and creates a meaningful outcome reflecting an overall agreement^{58,59}. We used RefFinder for consensus ranking. We additionally checked the expression levels of examined genes in 6 transcriptomic datasets of highly proliferative and motile feeding stages vs 5 localized, predominantly intracellular spore-forming stages, for further confirmation of qPCR data obtained. In principle, RNAseq data can be used to identify good reference genes^{60,61} without previous selection according to published data. This is offering attractive perspectives regarding new RG discoveries, since the proposed workflows can be used for already generated transcriptomic datasets, regardless of sequencing technology, library size or organism⁶⁰. With regard to myxozoans, there are not enough transcriptomic datasets of different developmental stages available for a single species and we hence compared preselected genes in new transcriptomic data of two different developmental stages of *S. molnari*. While other studies used TPM (Transcripts Per Million) or FPKM (Fragments Per Kilobase of transcript per Million mapped reads) values from RNA-Seq studies to analyse gene expression stability via CV (coefficient of variation) and fold change calculation methods^{62–64} we used a relatively straightforward approach, simply comparing already cross-sample normalized TMM values of the gene between studied conditions.

Stability of the candidate reference genes and experimental conditions. Numerous studies showed that the stability of proposed HKGs vary across organisms and most importantly experimental conditions (i.e. developmental stages, drug/dietary treatment, temperature such as heat shock, cold stress, drought stress, etc.)^{23,64,65}. Only genes that have stable expression under a condition to be analysed can be used as an RG for the given study. In the present case, we were mainly focused on providing RGs that are stable during myxosporean development, rather than e.g. under different temperature or dietary regimes. This is of particular importance to be able to investigate the stage-specific expression profile of parasite genes.

In our study, comprehensive ranking together with transcriptomic TMM calculations suggested that EF2 and GAPDH are the most stable genes across all the studied myxozoan species (Table 5). EF2 promotes the GTP-dependent translocation of the ribosome⁶⁶. It is an essential factor for protein synthesis and thus, like other ribosomal genes, is assumed to have a constant expression throughout different tissues, different treatments or developmental stages of the organism. It has been shown to be the most stable gene for mouse DNBS disease treated and non-treated colon tissue⁶⁷, or for plant tissues exposed to biotic and abiotic stress⁶⁸.

GAPDH is one of the most used RGs in different organisms, including corals, fish, human, etc.^{23,29,69}. Although several studies showed that GAPDH can be regulated under some experimental conditions, i.e. gene expression in thermal and light studies^{70,71} it is reported to be a suitable reference for normalizing gene expression in various life stages, for instance in red algae⁷², plants⁷³ or during the metamorphosis of free living cnidarian representatives such as corals (*Porites astreoides*²⁹), which coincides well with the general scheme of our study. Our results suggested GAPDH as the first most stable gene for *M. cerebralis* and *C. shasta*, and the second most stable gene for *S. molnari*.

Actin is one of the most conserved proteins in eukaryotes, whose structure has been conserved despite the numerous actin isoforms reported with different biological functions^{64,65}. Beta actin is one of the most common reference genes used in gene expression studies as it is known to be a key component involved in the development of cytoskeletal filaments⁶⁶. It was listed as one of the best-performing reference genes in cnidarian/dinoflagellate studies²⁶. Similarly, it was the best performing reference gene, along with 28S rRNA, in the parasitic apicomplexan *Theileria parva* from different host tissues⁶⁷. Beta actin was used as a reference gene for *M. cerebralis*² and a number of gene expression studies in free-living cnidarians^{26–28}, however, it has been shown to be regulated under various experimental conditions and was redefined as an unsuitable reference gene in some cases²³. Multiple actin isoforms are known from myxozoans^{74,75} and in *S. molnari*, the expression level of two highly divergent isoforms differs about 15-fold, since likely only isoform 1 is responsible for the unique parasite motility during proliferation in the blood⁷⁵. These highly variable expression levels of different isoforms suggest different functions and even though we used actin isoform 2, which was expressed at a low level, for the design of our qPCR assay, our results did not support beta actin to be an optimal reference gene for *S. molnari* or the other myxozoan species studied. It was ranked only in third place for *S. molnari* and *M. cerebralis*, and in sixth for *C. shasta*. TMM data also placed it in the fourth place for *S. molnari*. Actin isoforms are very similar to beta actin and gamma actin, differing only by four biochemically similar residues and being conserved from birds to mammals⁷⁶, however, actins show highly divergent DNA sequences in myxozoans⁷⁵. Hence, the possibility of misidentification of the same actin isoform in different species of the highly derived myxozoans gives an additional reason for excluding beta actin from qPCR analyses in myxozoans.

Alongside with beta actin, ribosomal genes especially 18S and 28S rRNA are traditionally used as reference genes. Being structural components of small and large eukaryotic ribosomal subunits (40S and 60S), they are one of the most basic components of eukaryotic cells. However, the suitability of 18S and 28S rRNA as a reference gene varies in the literature. In myxozoan studies 18S was used for phylogenetic studies and for detection and quantification of parasites in environment or host tissue^{34,35}. Thus, we have included it in our study to test its utility as reference gene, since it is shown to have high stability. However, despite its high stability in a number of organisms^{77–79} many studies indicated that these are highly expressed genes and are often unsuitable for comparison. It can be challenging to compare them with target genes expressed at a low level which can lead to erroneous results^{23,80}. Indeed, 18S and 28S rRNA were the most highly expressed genes (Cq < 21) in our study, except for *C. shasta* (Cq for 18S = 23.9–25.6 in intestine and worm samples, respectively). In our study 28S was the fourth most stable gene for *S. molnari* and *C. shasta*.

Stability of the candidate reference genes and the influence of used methodology. Despite RT-qPCR being one of the most reliable techniques to accurately measure the expression level of a gene, there are number of factors that may affect the consistency of expression data. For instance, different dyes can influence PCR inhibition in a concentration-dependent manner and can have effects on DNA melting temperature or can preferentially bind to certain DNA sequences⁸¹. PCR efficiency can be influenced by PCR inhibitors present in the sample or by non-optimal primer design. This information is critical, since these factors can produce different results even if the experimental design or study organism is similar (see also notes in Suppl. Mat. 5). Overall, we obtained consistent results for the suitability of RGs in three myxosporean species from different phylogenetic clades: EF2, GAPDH together with ACTB were the most stable genes for *S. molnari* and *M. cerebralis*, and EF2, GAPDH and AHC1 for *C. shasta*. Additionally, comparable TMM values of *S. molnari* demonstrate the robustness of our predictions. Thus, it can be hypothesized that these genes will also have stable expression in other related myxozoan species. These results are useful in particular for studies involving developmental stages of these parasites, however furthermore it would be of great interest to check the stability of these genes under different experimental conditions such as temperature, drug treatment, different dietary treatment, water quality, etc.

Possible pitfalls of detecting less abundant transcripts and further recommendations.

Likewise, highly expressed genes, such as 18S and 28S rRNA, reference gene may not be suitable to use in gene expression studies, if the gene of interest (i.e. target gene) has low expression in comparison to the reference gene. For instance, in our study, suggested reference genes (EF2, GAPDH) showed a suitable expression range ($C_q = 23\text{--}33$) in all samples (see Fig. 1, and Suppl. Mat. 2 for the abundance of each gene in each sample). However, we noticed that all the investigated genes in these samples showed low expression levels, which could be related to the low amount of parasite concentration in the sample or PCR inhibition, which may occur in guts and soil samples^{82–84}.

PCR inhibition is something to be aware of, especially in invertebrate host extractions procedures, and need to be carefully evaluated to ensure sufficient representation of quantifiable transcripts in subsequent myxozoan studies.

Several attempts (i.e. using inhibitor removal column and reagents, see details in Suppl. Mat. 5) were undertaken to reduce inhibitory effect in this study, which did not improve our overall results. Another useful way to evaluate the inhibitory effect is to use serial dilutions, since inhibitory effect can be lost in high dilutions. However, for the samples where the parasite concentration is already extremely low, dilution may not always be an optimal solution.

While we cannot exclude PCR inhibition as the reason for low detection of some of our genes, it is possible that the high C_q values observed are simply related to low parasite concentrations in the sample or these transcripts are expressed in low levels.

In either way, detecting less abundant transcripts remain an open question, especially in invertebrate samples, and more analyses (i.e. using new inhibitory effects removal kits, testing different concentration of RNA, carefully evaluating the primer design) may help to clarify this problem.

Finally, using a single reference gene for gene normalization is generally less reliable than the use of a set of genes⁸⁵, and based on our data, we propose using a combination of at least 2 to 3 genes for myxozoans. To our knowledge, this is the first study to validate RGs for myxozoan species, and we are convinced that the results presented here serve as an essential aid for subsequent gene expression studies of this group of extremely derived parasites.

Data availability

The data that support the findings of this study are available in Supplementary Material.

Received: 13 January 2019; Accepted: 2 October 2019;

Published online: 21 October 2019

References

1. Wolf, K. & Markiw, M. E. Biology contravenes taxonomy in the Myxozoa: New discoveries show alternation of invertebrate and vertebrate hosts. *Science* **225**, 1449–1452 (1984).
2. Eszterbauer, E. *et al.* Myxozoan life cycles: Practical approaches and insights. In *Myxozoan Evolution, Ecology and Development* 175–198, https://doi.org/10.1007/978-3-319-14753-6_10 (Springer International Publishing, 2015).
3. Kent, M. L. *et al.* Recent advances in our knowledge of the Myxozoa. *J. Eukaryot. Microbiol.* **48**, 395–413 (2001).
4. Sterud, E. *et al.* Severe mortality in wild atlantic salmon *Salmo salar* due to proliferative kidney disease (PKD) caused by *Tetracapsuloides bryosalmonae* (Myxozoa). *Dis. Aquat. Organ.* **77**, 191–198 (2007).
5. Peeler, E. J. & Taylor, N. G. The application of epidemiology in aquatic animal health -opportunities and challenges. *Vet. Res.* **42**, 94 (2011).
6. Hallett, S. L. & Bartholomew, J. L. *Myxobolus cerebralis* and *Ceratomyxa shasta*. In *Fish parasites: pathobiology and protection* (eds Woo, P. T. K. & Buchmann, K.) 141–172 ((CABI), 2012).
7. True, K., Voss, A. & Foott, J. S. Myxosporean parasite (*Ceratonova shasta* and *Parvicapsula minibicornis*) prevalence of infection in Klamath river basin juvenile Chinook salmon, March– August 2017. *Calif. -Nevada Fish Heal. Cent* (2013).
8. Fontes, I., Hallett, S. L. & Mo, T. A. Comparative epidemiology of myxozoan diseases. In *Myxozoan Evolution, Ecology and Development*. (eds Okamura, B. *et al.*) 85–110 (Springer Int. Pub, 2015).
9. FAO. The State of world fisheries and aquaculture 2016. Contributing to food security and nutrition for all, Rome, 200 pp (2016).
10. FAO. Aquaculture newsletters. No 56 Rome, 51pp (2017).
11. Clifton-Hadley, R. S., Bucke, D. & Richards, R. H. Economic importance of proliferative kidney disease in salmonid fish in England and Wales. *Vet. Res.* **119**, 305–306 (1986).
12. Granath, W. O., Gilbert, M. A., Wyatt-Pescador, E. J. & Vincent, E. R. Epizootiology of *Myxobolus cerebralis*, the causative agent of salmonid whirling disease in the rock creek of west-central Montana. *J. Parasitol.* **93**, 104–119 (2007).
13. Ray, R. A., Holt, R. A. & Bartholomew, J. L. Relationship between temperature and *Ceratomyxa shasta*-induced mortality in Klamath river salmonids. *J. Parasitol.* **98**, 520–526 (2012).

14. Ray, R. A., Alexander, J. D., De Leenheer, P. & Bartholomew, J. L. Modeling the effects of climate change on disease severity: A case study of *Ceratonova* (syn *Ceratomyxa*) *shasta* in the Klamath river. In *Myxozoan Evolution, Ecology and Development* 363–378, https://doi.org/10.1007/978-3-319-14753-6_19 (Springer International Publishing, 2015).
15. Holzer, A. S. *et al.* The joint evolution of the Myxozoa and their alternate hosts: A cnidarian recipe for success and vast biodiversity. *Mol. Ecol.* **27**, 1651–1666 (2018).
16. Chang, E. S. *et al.* Genomic insights into the evolutionary origin of Myxozoa within Cnidaria. *Proc. Natl. Acad. Sci.* **112**, 14912–14917 (2015).
17. Molnár, K. Gill sphaerosporosis in the common carp and grass carp. *Acta. Vet. Acad. Sci. Hungary* **27**, 99–113 (1979).
18. Korytář, T. *et al.* The kinetics of cellular and humoral immune responses of common carp to presporogonic development of the myxozoan *Sphaerospora molnari*. *Parasites & Vectors* **12**, 208 (2019).
19. Holzer, A. S. *et al.* Molecular fingerprinting of the myxozoan community in common carp suffering Swim Bladder Inflammation (SBI) identifies multiple etiological agents. *Parasites & Vectors* **7**, 398 (2014).
20. Bartholomew, J. L., Whipple, M. J., Stevens, D. G. & Fryer, J. L. The life cycle of *Ceratomyxa shasta*, a myxosporean parasite of salmonids, requires a freshwater polychaete as an alternate host. *J. Parasitol.* **83**, 859–868 (1997).
21. Thellin, O., El-Moualij, B., Heinen, E. & Zorzi, W. A decade of improvements in quantification of gene expression and internal standard selection. *Biotechnol. Adv.* **27**, 323–33 (2009).
22. Valente, V. *et al.* Selection of suitable housekeeping genes for expression analysis in glioblastoma using quantitative RT-PCR. *BMC Mol. Biol.* **10**, 17 (2009).
23. Kozera, B. & Rapacz, M. Reference genes in real-time PCR. *J. Appl. Genet.* **54**, 391–406 (2013).
24. San Segundo-Val, I. & Sanz-Lozano, C. S. Introduction to the gene expression analysis. *Methods in molecular biology (Clifton, N.J.)* **1434**, 29–43 (2016).
25. Lehnert, E. M. *et al.* Extensive differences in gene expression between symbiotic and aposymbiotic cnidarians. *G3(Bethesda); Genes|Genomes|Genetics* **4**, 277–295 (2014).
26. Levy, O. *et al.* Light-responsive cryptochromes from a simple multicellular animal, the coral *Acropora millepora*. *Science (80-)* **318**, 467–470 (2007).
27. Moya, A. *et al.* Cloning and use of a coral 36B4 gene to study the differential expression of coral genes between light and dark conditions. *Mar. Biotechnol.* **10**, 653–663 (2008).
28. Rodriguez-Lanetty, M., Phillips, W. S., Dove, S., Hoegh-Guldberg, O. & Weis, V. M. Analytical approach for selecting normalizing genes from a cDNA microarray platform to be used in q-RT-PCR assays: A cnidarian case study. *J. Biochem. Biophys. Methods* **70**, 985–991 (2008).
29. Kenkel, C. D. *et al.* Development of gene expression markers of acute heat-light stress in reef-building corals of the genus *Porites*. *PLoS One* **6**, e26914 (2011).
30. Sorek, M. *et al.* Setting the pace: host rhythmic behaviour and gene expression patterns in the facultatively symbiotic cnidarian *Aiptasia* are determined largely by *Symbiodinium*. *Microbiome* **6**, 83 (2018).
31. Kelley, G. O., Adkison, M. A., Leutenegger, C. M. & Hedrick, R. P. *Myxobolus cerebralis*: identification of a cathepsin Z-like protease gene (MyxCP-1) expressed during parasite development in rainbow trout, *Oncorhynchus mykiss*. *Exp. Parasitol.* **105**, 201–210 (2003).
32. Eszterbauer, E., Kallert, D. M., Grabner, D. & El-Matbouli, M. Differentially expressed parasite genes involved in host recognition and invasion of the triactinomyxon stage of *Myxobolus cerebralis* (Myxozoa). *Parasitology* **136**, 367 (2009).
33. Sarker, S., Menanteau-Ledouble, S., Kotob, M. H. & El-Matbouli, M. A RNAi-based therapeutic proof of concept targets salmonid whirling disease *in vivo*. *PLoS One* **12**, e0178687 (2017).
34. Hallett, S. L. & Bartholomew, J. L. Application of a real-time PCR assay to detect and quantify the myxozoan parasite *Ceratomyxa shasta* in river water samples. *Dis Aquat Organ.* **71**, 109–118 (2006).
35. Jorgensen, A. *et al.* Real-time PCR detection of *Parvicapsula pseudobranchicola* (Myxozoa: Myxosporae) in wild salmonids in Norway. *J. Fish Dis.* **34**, 365–371 (2011).
36. Alama-Bermejo, G., Sima, R., Raga, J. A. & Holzer, A. S. Understanding myxozoan infection dynamics in the sea: Seasonality and transmission of *Ceratomyxa puntazzi*. *Int. J. Parasitol.* **9**, 771–780 (2013).
37. Sipos, D. *et al.* Susceptibility-related differences in the quantity of developmental stages of *Myxobolus spp.* (Myxozoa) in fish blood. *PLoS One* **13**, e0204437 (2018).
38. Marton, S. & Eszterbauer, E. The susceptibility of diverse species of cultured oligochaetes to the fish parasite *Myxobolus pseudodispar* Gorbunova (Myxozoa). *J. Fish Dis.* **35**, 303–314 (2012).
39. Atkinson, S. D. & Bartholomew, J. L. Disparate infection patterns of *Ceratomyxa shasta* (Myxozoa) in rainbow trout *Oncorhynchus mykiss* and Chinook salmon *Oncorhynchus tshawytscha* correlate with ITS-1 sequence variation in the parasite. *Int. J. Parasitol.* **40**, 599–604 (2010).
40. Ibarra, A. M., Hedrick, R. P. & Gall, G. A. E. Inheritance of susceptibility to *Ceratomyxa Shasta* (Myxozoa) in rainbow trout and the effect of length of exposure on the liability to develop ceratomyxosis. *Aquaculture* **104**, 217–229 (1992).
41. Willson, S. J., Wilzbach, M. A., Malakauskas, D. M. & Cummins, K. Lab rearing of a freshwater polychaete (*Manayunkia speciosa*, Sabellidae) host for Salmon pathogens. *Northwest Science* **84**, 183–191 (2010).
42. Bjork, S. J. & Bartholomew, J. L. Effects of *Ceratomyxa shasta* dose on a susceptible strain of rainbow trout and comparatively resistant Chinook and Coho salmon. *Dis. Aquat. Organ.* **86**, 29–37 (2009).
43. Chomczynski, P. & Sacchi, N. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal. Biochem.* **162**, 156–159 (1987).
44. Eszterbauer, E. *et al.* Molecular characterization of *Sphaerospora molnari* (Myxozoa), the agent of gill sphaerosporosis in common carp *Cyprinus carpio carpio*. *Dis Aquat Organ.* **104**(1), 59–67 (2013).
45. Atkinson, S. D., Hallett, S. L. & Bartholomew, J. L. Genotyping of individual *Ceratonova shasta* (Cnidaria: Myxosporae) myxospores reveals intra-spore ITS-1 variation and invalidates the distinction of genotypes II and III. *Parasitology* **145**, 1588–1593 (2018).
46. Pfaffl, M. W. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res.* **29**, e45 (2001).
47. Chong, G., Kuo, F.-W., Tsai, S. & Lin, C. Validation of reference genes for cryopreservation studies with the gorgonian coral endosymbiont *Symbiodinium*. *Sci. Rep.* **7**, 39396 (2017).
48. Silver, N., Best, S., Jiang, J. & Thein, S. L. Selection of housekeeping genes for gene expression studies in human reticulocytes using real-time PCR. *BMC Mol. Biol.* **7**, 33 (2006).
49. Andersen, C. L., Jensen, J. L. & Ørntoft, T. F. Normalization of Real-Time Quantitative Reverse Transcription-PCR Data: A Model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. *Cancer Res.* **64**, 5245–5250 (2004).
50. Hellemans, J., Mortier, G., De Paep, A., Speleman, F. & Vandesompele, J. qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. *Genome Biol.* **8**, R19 (2007).
51. Vandesompele, J. *et al.* Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* **3**, research0034.1 (2002).
52. Pfaffl, M. W., Tichopad, A., Prgomet, C. & Neuvians, T. P. Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper–Excel-based tool using pair-wise correlations. *Biotechnol. Lett.* **26**, 509–515 (2004).

53. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
54. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).
55. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
56. Hibbeleer, S., Scharnsack, J. P. & Becker, S. Housekeeping genes for quantitative expression studies in the three-spined stickleback *Gasterosteus aculeatus*. *BMC Mol. Biol.* **9**, 18 (2008).
57. Dzaki, N., Ramli, K. N., Azlan, A., Ishak, I. H. & Azzam, G. Evaluation of reference genes at different developmental stages for quantitative real-time PCR in *Aedes aegypti*. *Sci. Rep.* **7**, 43618 (2017).
58. Pihur, V., Datta, S. & Datta, S. RankAggreg, an R package for weighted rank aggregation. *BMC Bioinformatics* **10**, 62 (2009).
59. Vanhauwaert, S. *et al.* Expressed repeat elements improve RT-qPCR normalization across a wide range of zebrafish gene expression studies. *PLoS One* **9**, e109091 (2014).
60. Carmona, R. *et al.* Automated identification of reference genes based on RNA-seq data. *Biomed. Eng. Online* **16**, 65 (2017).
61. Gao, D., Kong, F., Sun, P., Bi, G. & Mao, Y. Transcriptome-wide identification of optimal reference genes for expression analysis of *Pyropia yezoensis* responses to abiotic stress. *BMC Genomics* **19**, 251 (2018).
62. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
63. Stanton, K. A. *et al.* A Whole-transcriptome approach to evaluating reference genes for quantitative gene expression studies: A case study in *Mimulus. G3 (Bethesda)* **7**, 1085–1095 (2017).
64. Brunner, A. M., Yakovlev, I. A. & Strauss, S. H. Validating internal controls for quantitative plant gene expression studies. *BMC Plant Biol.* **4**, 14 (2004).
65. de Jonge, H. J. M. *et al.* Evidence based selection of housekeeping genes. *PLoS One* **2**, e898 (2007).
66. Susorov, D. *et al.* Eukaryotic translation elongation factor 2 (eEF2) catalyzes reverse translocation of the eukaryotic ribosome. *J. Biol. Chem.* **293**, 5220–5229 (2018).
67. Eissa, N. *et al.* Stability of reference genes for messenger RNA quantification by Real-Time PCR in mouse dextran sodium sulfate experimental colitis. *PLoS One* **11**, e0156289 (2016).
68. Boava, L. P. *et al.* Selection of endogenous genes for gene expression studies in Eucalyptus under biotic (*Puccinia psidii*) and abiotic (acibenzolar-S-methyl) stresses using RT-qPCR. *BMC Res. Notes* **3**, 43 (2010).
69. Zainuddin, A., Chua, K. H., Abdul Rahim, N. & Makpol, S. Effect of experimental treatment on GAPDH mRNA expression as a housekeeping gene in human diploid fibroblasts. *BMC Mol. Biol.* **11**, 59 (2010).
70. Dombrowski, J. E. & Martin, R. C. Evaluation of reference genes for quantitative RT-PCR in *Lolium temulentum* under abiotic stress. *Plant Sci.* **176**, 390–396 (2009).
71. Lovdal, T. & Lillo, C. Reference gene selection for quantitative real-time PCR normalization in tomato subjected to nitrogen, cold, and light stress. *Anal. Biochem.* **387**, 238–242 (2009).
72. Wu, X. *et al.* Variation of expression levels of seven housekeeping genes at different life-history stages in *Porphyra yezoensis*. *PLoS One* **8**, e60740 (2013).
73. Li, M.-Y. *et al.* Validation and comparison of reference genes for qPCR normalization of Celery (*Apium graveolens*) at different development stages. *Front. Plant Sci.* **7**, 313 (2016).
74. Kelley, G. O., Beauchamp, K. A. & Hedric, R. Phylogenetic comparison of the myxosporea based on an actin cDNA isolated from *Myxobolus cerebralis*. *J. Eukaryot. Microbiol.* **51**, 660–663 (2004).
75. Hartigan, A. *et al.* New cell motility model observed in parasitic cnidarian *Sphaerospora molnari* (Myxozoa: Myxosporea) blood stages in fish. *Sci. Rep.* **6**, 39093 (2016).
76. Perrin, B. J. & Ervasti, J. M. The actin gene family: function follows isoform. *Cytoskeleton* **67**, 630–634 (2010).
77. Kim, B.-R., Nam, H.-Y., Kim, S.-U., Kim, S.-I. & Chang, Y.-J. Normalization of reverse transcription quantitative-PCR with housekeeping genes in rice. *Biotechnol. Lett.* **25**, 1869–72 (2003).
78. Huggett, J., Dheda, K., Bustin, S. & Zumla, A. Real-time RT-PCR normalisation; strategies and considerations. *Genes Immun.* **6**, 279–284 (2005).
79. Kuchipudi, S. V. *et al.* 18S rRNA is a reliable normalisation gene for real time PCR based on influenza virus infected cells. *Virology* **9**, 230 (2012).
80. Bär, M., Bär, D. & Lehmann, B. Selection and validation of candidate housekeeping genes for studies of human Keratinocytes—Review and recommendations. *J. Invest. Dermatol.* **129**, 535–537 (2009).
81. Gudnason, H., Dufva, M., Bang, D. D. & Wolff, A. Comparison of multiple DNA dyes for real-time PCR: effects of dye concentration and sequence composition on DNA amplification and melting temperature. *Nucleic Acids Res.* **35**, e127 (2007).
82. Penn, H. J., Chapman, E. G. & Harwood, J. D. Overcoming PCR inhibition during DNA-based gut content analysis of ants. *Environ. Entomol.* **45**, 1255–1261 (2016).
83. Schrader, C., Schielke, A., Ellerbroek, L. & Johne, R. PCR inhibitors—Occurrence, properties and removal. *J. Appl. Microbiol.* **113**, 1014–1026 (2012).
84. Matheson, C. D., Gurney, C., Esau, N. & Lehto, R. Assessing PCR Inhibition from humic substances. *Open Enzym. Inhib. J.* **3**, 38–45 (2014).
85. Bustin, S. A. *et al.* The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments. *Clin. Chem.* **55**, 611–622 (2009).
86. Najafpanah, M. J., Sadeghi, M. & Bakhtiarizadeh, M. R. Reference genes selection for Quantitative Real-Time PCR using RankAggreg method in different tissues of *Capra hircus*. *PLoS One* **8**, e83041 (2013).
87. Radonić, A. *et al.* Guideline to reference gene selection for quantitative real-time PCR. *Biochem. Biophys. Res. Commun.* **313**, 856–62 (2004).
88. Stürzenbaum, S. R. & Kille, P. Control genes in quantitative molecular biological techniques: the variability of invariance. *Comp. Biochem. Physiol. Part B* **130**, 281–289 (2001).
89. Tsotetsi, T. N., Collins, N. E., Oosthuizen, M. C. & Sibeko-Matjila, K. P. Selection and evaluation of housekeeping genes as endogenous controls for quantification of mRNA transcripts in *Theileria parva* using quantitative real-time polymerase chain reaction (qPCR). *PLoS One* **13**, e0196715 (2018).

Acknowledgements

We acknowledge financial support by the Czech Science Foundation (grant numbers 19-28399X (to AK) and 14-28784P (to GAB)), the European Commission (project reference 634429, ParaFishControl (to ASH)), Ministry of Education, Youth and Sports of the Czech Republic (project no. LTAUSA17201 (to PBS)), the Hungarian National Research, Development and Innovation Office (project no. NN124220 (to EE)) and Consellería de Educación, Investigación, Cultura y Deporte, Valencia, Spain (APOSTD/2013/087 (to GAB)). We thank Stephen D. Atkinson and Julie Alexander (OSU, USA) for the help in Manayunkia worm sampling.

Author contributions

A.K. and A.S.H. conceived and designed the study. A.S.H., G.A.B., P.B.S., A.B.T., E.E. and J.B. collected/provided the scientific material. A.K., A.S.H., R.S. and A.N. design the methodology. A.K. G.A.B. and P.B.S. generate the data. A.K. analyzed the data. All authors contributed for writing and approving the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41598-019-51479-0>.

Correspondence and requests for materials should be addressed to A.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019

A uniquely complex mitochondrial proteome from *Euglena gracilis*

Michael J. Hammond¹, Anna Nenarokova^{1,2}, Anzhelika Butenko^{1,3}, Martin Zoltner^{4,5}, Eva Lacová Dobáková¹, Mark C. Field^{1,4} & Julius Lukeš^{1,2,*}

¹Biology Centre, Institute of Parasitology, Czech Academy of Sciences, České Budějovice (Budweis), Czech Republic

² Faculty of Sciences, University of South Bohemia, České Budějovice (Budweis), Czech Republic

³ Faculty of Science, University of Ostrava, Ostrava, Czech Republic

⁴ School of Life Sciences, University of Dundee, Dundee, United Kingdom

⁵ Faculty of Science, Charles University, Biocev, Vestec, Czech Republic

*Corresponding author: jula@paru.cas.cz

Abstract

Euglena gracilis is a metabolically flexible, photosynthetic and adaptable free-living protist of considerable environmental importance and biotechnological value. By label-free LC-MSMS, a total of 1,787 proteins were identified from the *E. gracilis* purified mitochondria, representing one of the largest mitochondrial proteomes so far described. Despite this apparent complexity, protein machinery responsible for the extensive RNA editing, splicing and processing in the sister clades diplomonads and kinetoplastids, is absent. This strongly suggests that the complex mechanisms of mitochondrial gene expression in diplomonads and kinetoplastids occurred late in euglenozoan evolution, arising independently. By contrast, the alternative oxidase pathway and numerous ribosomal subunits presumed to be specific for parasitic trypanosomes are present in *E. gracilis*. We investigated the evolution of unexplored protein families, including import complexes, cristae formation proteins and translation termination factors, as well as canonical and unique metabolic pathways. We additionally compare this mitoproteome with the transcriptome of *Eutreptiella gymnastica*, illuminating conserved features of Euglenidae mitochondria as well as those exclusive to *E. gracilis*. This

is the first mitochondrial proteome of a free-living protist from the Excavata, and one of few available for protists as a whole. This study alters our views of the evolution of the mitochondrion, and indicates early emergence of complexity within euglenozoan mitochondria, independent of parasitism.

Introduction

The mitochondrion is an important and versatile organelle with core activities in energy production, iron-sulphur cluster biosynthesis, regulation of apoptosis and metabolism of lipids and amino acids. Arising from the endosymbiosis of an alpha-proteobacterium, before evolving into a truly integrated organelle, mitochondria played a leading role in eukaryogenesis (Lane and Martin 2010). With the remarkable exception of a single known truly amitochondriate protist group (Karnkowska et al. 2016), all extant eukaryotes retain mitochondria or mitochondrion-related organelles which claim descent from the mitochondria of the last eukaryotic common ancestor (Roger et al. 2017).

While the alpha-proteobacterial ancestor is reconstructed as possessing a genome of ~5000 protein-coding genes (Boussau et al. 2004), all but a few genes in modern mitochondria have been either transferred to the nuclear genome or lost. Therefore, most mitochondrial proteins are imported from the cytosol *via* a specialised translocation apparatus (Mokranjac and Neupert 2009). Accordingly, the mitochondrial genome is of limited use for predicting the total mitochondrial proteome (mitoproteome) and its functions. Indeed, the proteins of modern mitochondria are derived from multiple eukaryotic and prokaryotic sources, with only a small fraction estimated to be contributed by the original endosymbiont (Gray 2015). Although many *in silico* methods predict localisation to mitochondria (Fukasawa et al. 2015; Guda et al. 2004), these predictions are hampered by variability in mitochondrial targeting signals and mechanisms (Santos et al. 2018), as well as divergence across the eukaryotic domain (Emanuelsson et al. 2007; Fukasawa et al. 2017). Thus, direct proteomic approaches remain essential for determining a mitoproteome.

Mitoproteomes have been established for a few unicellular (Casaletti et al. 2017; Sickmann et al. 2003) and multicellular animals, fungi (Calvo et al. 2016; Li et al. 2009; Taylor et al. 2003) and plants (Heazlewood et al. 2003; Hochholdinger et al. 2004; Lee et al. 2013; Mueller et al. 2014). Considering their abundance and diversity, protists remain under-investigated in this regard, and only mitoproteomes for parasitic *Trypanosoma brucei*

(Panigrahi et al. 2009), free-living *Tetrahymena thermophila* (Smith et al. 2007), *Chlamydomonas reinhardtii* (Atteia et al. 2009), and *Acanthamoeba castellanii* (Gawryluk et al. 2014), as well as the mitochondria-related organelles of parasitic *Trichomonas vaginalis* (Schneider et al. 2011), *Giardia intestinalis* (Jedelský et al. 2011) and *Entamoeba histolytica* have been studied in some detail (Mi-ichi et al. 2009).

The protist *Euglena gracilis* arguably represents one of the most comprehensively studied organisms within Euglenida, a group of diverse flagellates distinguished by a striated cell surface or pellicle (Adl et al. 2019). Euglenida belong to the phylum Euglenozoa along with Diplonemea, a recently discovered group of highly abundant and diverse marine flagellates (Flegontova et al. 2016), and Kinetoplastea, a group notable for numerous parasite members of public health importance (Gibson 2017). From an evolutionary perspective, the extensively studied kinetoplastids are more distant to euglenids than diplomonids (Vesteg et al. 2019). Therefore, any common traits shared between euglenids and kinetoplastids likely represent features possessed by the euglenozoan common ancestor and are expected to be distributed throughout extant members of the phylum. Moreover, their basal phylogenetic position makes euglenids important from an evolutionary perspective, especially since close relatives evolved extremely complex systems for mitochondrial RNA editing and/or *trans*-splicing (Faktorová et al. 2018; Read et al. 2016).

Euglenids are versatile organisms that possess a variety of nutritional strategies, including eukaryotrophy, bacteriotrophy and osmotrophy (Leander et al. 2017). *E. gracilis* can additionally employ photosynthesis due to the presence of a triple membrane-bound plastid acquired through a secondary endosymbiotic event (Zakryś et al. 2017) and thanks to an anaerobically capable mitochondrion, which generates energy *via* fatty acid fermentation, can grow in anoxic environments (Zimorski et al. 2017). The anaerobically produced wax esters that result from this fermentation are of biotechnological interest as a source of biofuel (Inui et al. 2017), along with a number of other compounds, such as the storage polysaccharide paramylon and essential amino acids which also serve as food supplements in parts of Asia (Krajčovič et al. 2015).

Recent advances in elucidating the molecular biology of *E. gracilis* include the sequencing of a draft genome, transcriptome and determining a whole cell proteome, identifying a large genome in excess of 500 Mb with a comparatively small coding region (<1%) that nonetheless is estimated to include over 36,000 protein-coding genes (Ebenezer et al. 2019).

Comparative transcriptomes for cells grown in light *versus* dark (Ebenezer et al. 2019), rich *versus* minimal media (O'Neill et al. 2015) and anoxic conditions have been described (Yoshida et al. 2016). Subcellular analyses are also being pursued, with the completion of a chloroplast proteome, revealing a total of 1,345 proteins of multiple origins and a seemingly divergent protein translocation apparatus (Novák Vanclová et al. 2019). The *E. gracilis* mitochondrial genome revealed only seven protein-coding genes and no evidence for post-transcriptional editing and/or splicing that are extensive in sister lineages (Dobáková et al. 2015). *In silico* prediction of the mitoproteome was estimated at approximately 1,100 proteins (Ebenezer et al. 2019), representing a cohort similar to the related *T. brucei* (Peikert et al. 2017), albeit with the caveats discussed above.

Here we report a mitoproteome for *E. gracilis*, obtained from purified organelles and analysed by liquid chromatography-mass spectrometry, which contains 1,787 proteins, and is complemented by *in silico* analysis. Notably, we report the identification of five of seven mitochondrially encoded proteins. We find no evidence for complex RNA editing and processing machineries orthologous with the mitochondrion of kinetoplastids and diplomonids, yet still encountered metabolic and structural complexity that challenges the assumption that protist mitochondria have compositional simplicity.

Materials and Methods

Cell culture and isolation of mitochondria

E. gracilis strain Z1 cells were axenically cultured in total volume of 600 ml in Hutner's medium at 27 °C in aerobic conditions under permanent light ($10 \mu\text{m}/\text{m}^{-2}\text{s}^{-1}$) with constant shaking at 150 rpm until they reached exponential growth ($1.5\text{--}2 \times 10^6$ cells/ml). Cells were collected by centrifugation at $800 \times g$ for 10 min and resuspended in SHE buffer (250 mM sucrose, 10 mM HEPES, 1 mM EDTA, pH 7.3) supplemented with 0.4% fatty acid-free bovine serum albumin (BSA). All the following steps were performed on ice. *Euglena* cells were disrupted by sonication at 80% power using a thick 19.5-mm probe (Ultrasonic homogenizer 3,000; Biologics, Inc.). Sonication was performed in six 10 s pulses cycles with 2-min breaks. The sonicate was centrifuged for 15 min at $800 \times g$ and 4 °C and the resulting supernatant centrifuged for 15 min at $8,500 \times g$ at 4 °C. The pellet was resuspended in 3 ml of STM buffer (250 mM sucrose, 20 mM Tris-HCl, 2 mM MgCl₂, pH 8.0) with 40 U of DNase I (Thermo Scientific) and incubated on ice for 30–60 min. DNase-treated lysate (5

ml) was loaded on top of a discontinuous sucrose density gradient. The gradient was prepared by layering decreasingly dense sucrose solutions upon one another from 2.0, 1.75, 1.5, 1.25, 1, to 0.5 M (5 ml each) and centrifuged in a SW-28 rotor at 87,000 x g at 4 °C for 4.5 h (L8-M, Beckman).

The mitochondrial fraction was separated from chloroplast and peroxisome fractions, located at the interface of 1.7 and 1.5 M sucrose layers and collected with a syringe. To remove excess sucrose mitochondria were washed twice in SHE buffer. The final pellet was gently resuspended with a cut-off pipette tip in 500–1,000 µl of SHE buffer supplemented with 0.4% Bovine Serum Albumin (BSA), and then spun for 30 min at 16,000 × g at 4 °C and stored at -80 °C.

Mass spectrometry-based protein identification and quantification

Samples were sonicated in NuPAGE LDS sample buffer (Thermo Sci.) containing 2 mM dithiothreitol and separated on a NuPAGE Bis-Tris 4–12% gradient polyacrylamide gel (Thermo) under reducing conditions. The sample lane was divided into eight slices that were excised from the Coomassie-stained gel, destained and subjected to tryptic digest and reductive alkylation. The treated fractions were subjected to liquid chromatography tandem mass spectrometry (LC-MSMS) on an UltiMate 3000 RSLCnano System coupled to a Q exactive mass spectrometer (Thermo Sci.). Mass spectra were analysed using MaxQuant V. 1.56, (Cox and Mann 2008) and by searching the predicted translated transcriptome of Ebenezer et al (2019). Minimum peptide length was set to six amino acids, isoleucine and leucine were considered indistinguishable and false discovery rates (FDR) of 0.01 were calculated at the levels of peptides, proteins, and modification sites based on the number of hits against the reversed sequence database. Ratios were calculated from label-free quantification intensities using only peptides that could be uniquely mapped to a given protein. If the identified peptide sequence set of one protein contained the peptide set of another protein, these two proteins were assigned to the same protein group. P values were calculated applying t-test based statistics using Perseus (Tyanova et al. 2016). 2,704 protein sequences were initially identified in the mitochondrial fraction which satisfied these criteria (Suppl. Table 1). Proteomics data has been deposited to the ProteomeXchange Consortium by the PRIDE (Vizcaíno et al. 2016) partner repository with the dataset identifier PXD014767.

Identification of mitochondrially–encoded proteins

Proteins encoded in the *E. gracilis* mitochondrial genome, determined by Miranda-Astudillo et al. (2018) and Dobáková et al. (2015) (Suppl. Table 2), were missing from the translated transcriptome, except for *cox3*. The MaxQuant LFQ analysis was repeated using this search database composed of seven protein sequences. The resulting quantifications of four additional mitochondrial-encoded proteins were included in the candidate dataset.

Reference mitoproteome comparison

Proteins of the mitochondrial fraction of *E. gracilis* were compared to datasets containing the mitoproteomes of *Arabidopsis thaliana*, *Mus musculus*, *Saccharomyces cerevisiae* and *T. brucei* (Suppl. Table 3). For *S. cerevisiae*, 1,010 mitochondrial proteins were selected using a combination of the following criteria: i) GO term ‘mitochondrion’ assigned to the protein; ii) annotation as mitochondrial assigned manually based on experimental studies, or identified as mitochondrial in ≥ 2 high-throughput analyses, or one high-throughput analysis plus computational evidence. The dataset for *A. thaliana* is composed of 722 mitochondrial proteins based on Lee et al. (2013) (Suppl. Table 3). Mitochondrial proteins of *M. musculus* were downloaded from the Mitocarta database v.2.0, while for *T. brucei*, mitochondrial importome comprising 1,120 proteins identified using ImportOmics method was used. For each species within the dataset, only one representative protein isoform for each gene was included. *E. gracilis* sequences were included in accepted mitoproteome if an orthologous sequence was detected in at least one of the mitoproteomes (995 sequences) (Suppl. Table 4).

Predicted proteins from three available transcriptomes (Ebenezer et al. 2019; O'Neill et al. 2015; Yoshida et al. 2016) were analysed via Target-P 1.1 for predicted organellar localisation (Emanuelsson et al. 2000). Transcript ORFs starting with methionine and predicted to be imported into the mitochondrion in the majority of transcriptomes were considered as mitoproteome components (77 sequences). Mitochondrial fraction proteins were additionally compared against *in silico* predicted *E. gracilis* mitoproteome (Ebenezer et al. 2019), including any predicted sequences that were identified by LC-MSMS from the mitochondrial fraction (552) (Suppl. Table 4). Proteins with mitochondria to chloroplast and mitochondria to whole cell ratios greater than one, indicating mitochondrial enrichment

compared to the chloroplast and whole cell, were considered as mitochondrial (1,544) (Suppl. Table 5). Mitochondrial-to-chloroplast enriched sequences were assigned confidence values of the logarithmic enrichment ratio of the mitochondrial versus chloroplast preparation ($\log_{10}(\text{Mt/Cp})$), and sorted into groups of 0.0 -1.0, >1.0 (Suppl. Table 5).

Functional characterisation and mitochondrial protein families

Proteins considered as candidate mitoproteome components were annotated automatically using BLAST against the NCBI non-redundant protein database (August 2016 version). These annotations were checked manually consulting UniProt database and putative homologues inferred using OrthoFinder 2.2.7 as necessary (Emms and Kelly 2015). Proteins annotated as having roles in specific metabolic pathways or molecular complexes were sorted manually into 16 custom-defined categories roughly based on the KEGG pathway classification. 24 proteins were removed from the mitoproteome based on functional annotation showing clear non-mitochondrial function as well as a lack of orthologues in any reference mitoproteomes. Annotations of the predicted mitoproteome of *E. gracilis* (Ebenezer et al. 2019) were first compared to the 2,704 sequences of mitochondrial fraction (Suppl. Table 1) to determine functional proteins that were present (552). Functional protein families determined were then complemented via manual search for functionally annotated members from the experimentally verified mitoproteome. Missing key sequences from established mitochondrial protein families were then investigated with a combination of BLAST and Hidden Markov Model searches.

Hidden Markov Models

Sequences underwent ClustalW alignment via Mega10.0.5 (Kumar et al. 2018) and were searched using HMMER software against proteins generated primarily from the *E. gracilis* transcriptome (Ebenezer et al. 2019; O'Neill et al. 2015; Yoshida et al. 2016), which were accepted with an E value of 10^{-3} or lower and a bias score lower than 1. Sorting and assembly machinery SAM35, SAM37, atypical translocase of outer membrane (ATOM) subunits 19, 14, 12 and 11, peripheral ATOM36, translocase of inner membrane (TIM) 8, 11, 13, 40, 47, 54, 62, tiny TIMs, mitochondrial contact site and cristae organizing system (MICOS) subunits 10-1, 10-2, 16, 17, 20, 32, 34, 40 and 60, Hsp10, Nfu's, and Coq7, were queried for

with HMM profiles constructed using all available functionally annotated kinetoplast sequences from TriTryp website (Aslett et al. 2010). Inner membrane peptidases (IMP)1 and IMP2, were also determined using a spread of sequences from Amoebozoa, Archaeplastida and Alveolata. Pre-sequence translocated-associated motor subunit 17 and 18 were searched for using a selection of metazoan sequences. Metaxin sequences were searched for using a combination of metazoan, plants and protist sequences. TIM18 and TIM54 (both non-trypanosome), mammalian MICOS19/25, Coq6 and TOM22, were searched for using a selection of opisthokont sequences. Proteins identified as MICOS hits were analysed with TMHMM (Krogh and Rapacki 2013) and Coiled Coils (Combet et al. 2000) to identify structural features.

BLAST searches

Mitochondrial ribosomes, subunits of respiratory complexes, and protein import machinery (TIM11, TIM13 RHOM 1, RHOM 2, TINY TIM, Oxa2, TIM42) were searched using BLAST against the assembled *E. gracilis* transcriptome (Ebenezer et al. 2019) with *T. brucei* sequences as the query. Sequences with an E value of 10^{-3} or less were accepted.

Mitochondrial ribosomal subunits from databases of *T. brucei* (Desmond et al. 2011; Ramrath et al. 2018; Zíková et al. 2008), and *Diplonema papillatum* (our unpubl. data) were employed for BLAST searches. For mitochondrial ribosomal subunits not present in excavates, sequences from *H. sapiens* and *S. cerevisiae* were used as queries (Desmond et al. 2011).

Mitochondrial termination factor-like proteins (mTERF) of *A. thaliana* and *H. sapiens* (Kleine and Leister 2015) were used for BLAST searches. Protein sequences of *E. gracilis* respiratory complex subunits from previous studies (Miranda-Astudillo et al. 2018; Perez et al. 2014) were used as queries.

Survey of aminoacyl-tRNA synthetases and amino acid metabolism

The search for aminoacyl-tRNA synthetases (aaRSs) and related enzymes (tRNA-dependent amidotransferases, or AdTs) was performed using annotated sequences of *H. sapiens*, *Plasmodium falciparum* and *A. thaliana* as queries and all transcripts and predicted proteins of *E. gracilis* as database for BLAST searches with an E-value cut-off of 10^{-20} . For AdTs (Pfam models PF01425, PF02686 and PF02934) additional HMM-based searches were

performed using HMMER package v.3.1 (Eddy 2009) and predicted proteins from this study, as well as publicly available *E. gracilis* transcriptomes (O'Neill et al. 2015; Yoshida et al. 2016). InterProScan, Pfam and BlastKoala were used to facilitate functional annotation of the hits (El-Gebali et al. 2019; Kanehisa 2017; Kanehisa et al. 2016; Mitchell et al. 2019). NLStradamus, SeqNLS and NLS-Mapper were employed for nuclear localization signals prediction (Ba et al. 2009; Kosugi et al. 2009; Lin and Hu 2013). Reconstruction of mitochondrial amino acid metabolic pathways was performed with KEGG Mapper v.2.8 following KEGG IDs assignment to putative mitochondrial proteins using BlastKoala (Kanehisa et al. 2016).

Comparison to *Eutreptiella gymnastica*

The experimentally determined *E. gracilis* mitoproteome was compared to the translated transcriptome of *Eutreptiella gymnastica* strain NIES-381 with Orthofinder 2.2.7 (Emms and Kelly 2015). The *E. gymnastica* transcriptome additionally underwent Hidden Markov Model searches (with previously described parameters) against all available kinetoplastid sequences from TriTrypDB for the oxoglutarate decarboxylase complex subunits E1 and E2.

Results and discussion

Construction of a *Euglena* mitoproteome

For experimental determination of the mitoproteome, pelleted cells were disrupted by sonication, treated with DNaseI and subjected to discontinuous sucrose density gradient centrifugation, which clearly separated several distinct cellular fractions. The mitochondrial fraction formed a sharp and rather narrow band at the 1.5 M and 1.75 M sucrose interface (Suppl. Fig. 1). This mitochondrial fraction initially yielded 2,704 candidate proteins, of which 994 were orthologous to proteins in reference mitoproteomes, 1,544 were mitochondrially-enriched in comparison to both the chloroplast fraction and the whole cell, and 77 had a majority consensus of mitochondrial targeting based on signal peptide predictions (Fig. 1).

Together, the verified (i.e. experimentally determined and *in silico* validated) *E. gracilis* mitoproteome consists of 1,787 majority identified proteins, which were distributed into 1,756 protein groups (Suppl. Table 2; Suppl. Table 4). Of these, four sequences have been

shown to be encoded by the organellar genome (Suppl. Table 2) and the remaining 1,783 are of nuclear origin. This total cohort excludes 24 proteins which were identified through functional annotation as likely contaminants for their clear non-mitochondrial functions but includes another 35 proteins, which were initially excluded based on low enrichment values, but reintegrated based on predicted mitoproteome data and their functional annotation, indicating clear mitochondrial function (Suppl. Table 4).

Excluding razor and redundant peptides, 1,667 protein groups were identified with more than one unique peptide while 85 were identified from a single unique peptide only (Suppl. Table 2). The proteome showed wide variation in protein molecular weight, ranging from 416.09 to 10.33 kDa, with 1,307 proteins exhibiting a predicted molecular weight less than 50 kDa, representing approximately 74% of all experimentally verified proteins. When compared to the predicted mitoproteome, 552 of the 1,092 previously predicted proteins were experimentally verified (Ebenezer et al. 2019). Other organisms also show similar differences when confirming the expressed proteome, with two thirds of the *A. thaliana* predicted mitoproteome still to be recovered as expressed proteins (Lee et al. 2013). As mentioned previously, 35 of these predicted proteins were initially rejected due to insufficient mitochondrial enrichment; 12 had greater enrichment in the chloroplast and 23 in whole cells, but these were ultimately included due to their functional annotation and previous transcriptome-based prediction (Ebenezer et al. 2019) (Suppl. Table 4). Together with unconfirmed *in silico* sequences (Ebenezer et al. 2019) and the additional sequences identified in this study, we currently expect the complete mitoproteome of *E. gracilis* to consist of ~2,500 proteins (Suppl. Table 5). While mitochondrial enrichment represents an obvious criterion for verifying presence within the organelle, there are clearly exceptions to be accounted for, particularly proteins with multiple localisations.

Indeed, of the verified mitoproteome, 211 proteins showed greater enrichment in the chloroplast fraction (Suppl. Fig. 2), which were retained based on the presence of orthologues in reference mitoproteomes and/or high likelihood of mitochondrial targeting sequence, indicating a very low level of contamination of the chloroplast fraction with mitochondrial components. Importantly, the vast majority (1,567) shows mitochondrial fractional enrichment, with volcano plots revealing minimal contamination with other organelle components (Suppl. Fig. 3; Suppl. Table 6), supporting the validity of the mitoproteome predictions. Of the experimentally verified fraction, less than half of proteins (716) were found to possess a mitochondrial import signal in at least one of the available *E. gracilis*

transcriptomes as predicted via TargetP, likely arising from variability of import signals and 5' truncation of some cDNA sequences, highlighting the weakness of relying solely on target prediction software. Bioinformatic predictions of dual-targeted proteins is notoriously difficult, but certain protein families found in the mitochondrial matrix have shown a tendency for plastid presence as well, including those involved in nucleotide metabolism, DNA replication, tRNA biogenesis and translation (Carrie and Small 2013), which we consider the most likely candidates within our mitoproteome for dual-localisation.

Functional annotation could be assigned to 788 proteins, leaving an unexpected total of 999 (56%) proteins with unknown function. Of the annotated proteins, “core metabolic pathways”, “ribosome, aminoacyl-tRNA biosynthesis and translation” and “protein transport, folding, processing and degradation” were most represented (Fig. 2). When compared with the categorisation of the predicted mitoproteome, we see notable increase in the proportion of proteins with functions attributed to the “ribosome” and “protein transport” categories as well as a reduction in “oxidative phosphorylation and electron transport proteins” (Suppl. Fig. 4).

The *E. gracilis* mitoproteome is complex

The *E. gracilis* mitoproteome is larger than that predicted for all other eukaryotes, for which well-curated mitoproteomes based on both predictions and experimental data are available, and specifically *A. thaliana* (843) (Lee et al. 2013), *T. brucei* (1,120) (Peikert et al. 2017), *S. cerevisiae* (1,187) (Gonczarowska-Jorge et al. 2017) and *M. musculus* (1,158) (Calvo et al. 2016). While varied criteria are often used in different studies to determine genuine mitochondrial components making direct comparisons and estimates of reliability difficult, we consider the number of *E. gracilis* mitoproteins to be notable, especially as the estimated proteome sizes in these other model organisms were initially reported as lower (Chang et al. 2003; Panigrahi et al. 2009; Sickmann et al. 2003; Werhahn and Braun 2002), and have steadily increased over the past 20 years through improved protein identification, increased mass spectrometry sensitivity and experimental mitochondrial analysis (Calvo et al. 2016; Gonczarowska-Jorge et al. 2017; Peikert et al. 2017).

Moreover, when compared to the high confidence protist mitoproteomes of *Acanthamoeba castellanii* (709) (Gawryluk et al. 2014), *Tetrahymena thermophila* (573) (Smith et al. 2007) and *Chlamydomonas reinhardtii* (347) (Atteia et al. 2009), the complexity of the analysed mitoproteome is even more impressive. *E. gracilis* displays the largest number of unknown

sequences (57%) for any mitoproteome surveyed thus far, with corresponding frequencies in other mitoproteomes ranging from 17% in *C. reinhardtii* to 53% in *T. brucei*.

With ~2,500 sequences, the total predicted mitoproteome of *E. gracilis* (Suppl. Table 5) represents a size more typical of plant mitochondria, as several software prediction analyses of *A. thaliana* have consistently predicted 2,000 to 3,000 proteins residing within its mitochondria (Heazlewood et al. 2004; Lee et al. 2013; Millar et al. 2005). We suggest that a common factor contributing to potentially accelerated mitochondrial complexity is the presence of the plastid. *E. gracilis* has numerous proteins dually localised to both mitochondria and plastids, and additionally carries some plastid-derived pathways within the mitochondria, as well as what appears to be metabolic cycles which complement plastid function, indicating that plastid interactions with the mitochondrion can foster an overall greater complexity. In contrast to *E. gracilis*, experimental verification of many predicted plant mitoproteomes has remained challenging (Hochholdinger et al. 2004; Huang et al. 2009; Mueller et al. 2014). Indeed, experimentally verified mitoproteomes greater than 1,000 sequences remain a rarity (Salvato et al. 2014), a situation thought to reflect difficulty in detecting hydrophobic and low-abundance proteins (Millar et al. 2005). Recent investigations into the subcellular compartmentalisation of metabolic pathways in *E. gracilis* have been lacking extensive proteomic information on the mitochondria (Inwongwan et al. 2019), which we hope to rectify here.

mTERF proteins

Mitochondrial termination factor-like proteins (mTERF) are involved in transcription termination/activation and ribosome biogenesis (Roberti et al. 2009). While initially discovered in mitochondria, mTERFs are also present in chloroplasts, sometimes being dually localised (Kleine and Leister 2015). Whereas four groups of mTERFs have been described in vertebrates, plants have undergone a major expansion with 31 and 35 mTERFs in *Zea mays* and *A. thaliana*, respectively (Kleine and Leister 2015), of which 10 *A. thaliana* mTERFs have been confirmed by proteomics as mitochondrial (Lee et al. 2013). Of an unprecedented 192 mTERFs identified in the *E. gracilis* transcriptome, at least eight are identified in the mitoproteome (Suppl. Table 7), of which four show exclusive localisation to the mitochondrion, raising the possibility that the remaining four have dual localisations. We observed six additional mTERFs with considerable enrichment in the chloroplast (Suppl.

Table 7), suggesting that *E. gracilis* also partitions its mTERF family between these two endosymbiont-derived organelles. This raises questions of why *E. gracilis* contains such a large suite of mTERF proteins. The large mitochondrial genomes of plants, which exhibit considerable gene rearrangement and numerous introns, may require greater transcriptional regulation facilitated by the additional mTERFs (Kleine and Leister 2015). However, such an explanation does not seemingly apply to *E. gracilis* which, while containing fragmented mitoprotein-coding genes, nonetheless has a reduced set, importantly lacking the complex post-transcriptional modifications seen in other euglenozoan lineages (Dobáková et al. 2015).

Experimental studies suggest that in plants the expanded mTERF family enables functional diversification, fostering resistance to various abiotic stresses such as increased light, heat and salt concentrations (Quesada 2016; Robles et al. 2018), while mTERF18 confers heat tolerance to *A. thaliana* through regulating redox-related gene expression (Kim et al. 2012). mTERF-like gene MOC1-deficient mutants of *C. reinhardtii* display reduced growth under high light intensity (Schönfeld et al. 2004). Adaptions to abiotic stress may underly the dramatic expansion of mTERFs in *E. gracilis*. Moreover, since in metazoans mTERFs have also been observed mediating ribosome biogenesis (Wredenberg et al. 2013), we suggest that the expanded set of ribosomal subunits in *E. gracilis* (see below) may require additional mTERFs for proper ribosome assembly and/or function.

RNA editing and processing

Kinetoplastids and diplomonids developed highly complex, yet distinct systems of RNA editing and processing of their mitochondrial transcripts. While in kinetoplastids uridines are extensively inserted into and deleted from mRNAs (Read et al. 2016), in diplomonids mRNAs of systematically fragmented genes undergo *trans*-splicing and C-to-U, A-to-I and G-to-A, as well as U- and A-appendage editing (Valach et al. 2017) Kaur et al., accepted). Although *E. gracilis* does not have RNA editing (Dobáková et al. 2015), we raise the question of whether the common euglenozoan ancestor possessed pre-adaptations for these baroque mechanisms.

From the ~90 proteins involved in mitochondrial RNA editing and processing in *T. brucei*, only homologues to 12 proteins were present in the *E. gracilis* mitoproteome: MRB3010, RBP16, RGG2, PAMC, KPAF1, MHEL61, KRET2, KREX2, poly(A) polymerase, p22 precursor, a PAMC (polyadenylation mediator complex) component Tb927.6.3350, and

pentatricopeptide (PPR) protein Tb927.10.10160. An additional 19 proteins are predicted to be involved in mitochondrial RNA processing in *E. gracilis*. These include six DEAD-box RNA helicases, one PPR proteins, eight and two proteins similar to MRB3010 and RGG2, respectively, and two splicing factors belonging to the arginine/serine-rich 4/5/6-like family. Remarkably, these splicing factors have homologues only in diplomonids (our unpubl. data). Our data indicates that while the last common euglenozoan ancestor did not perform mitochondrial RNA editing, it already had a basic set of RNA processing proteins in its mitochondrion that were repurposed in the evolution of kinetoplastids and potentially diplomonids as key RNA editing enzymes. Those include RNA binding proteins, RNA helicases, terminal uridyl-transferase (TUTase) and the exonuclease listed above. This prompts the question of which pre-adaptions within the common euglenozoan ancestor triggered emergence of highly complex RNA editing and/or processing systems. One driving force might have been an increasingly scrambled mitochondrial genome (Flegontov et al. 2011). The emergence and evolution of excessive molecular complexity are best explained via purely neutral drift, which becomes irreversible due to ratchet mechanisms preventing subsequent simplification (Lukeš et al. 2011; McShea and Hordijk 2013; Stoltzfus 1999).

Mitochondrial-encoded proteins

The mitochondrial genome of *E. gracilis* encodes only seven proteins (nad1, nad4, nad5, cob, cox1, cox2 and cox3), all of which are respiratory complex subunits (Dobáková et al. 2015). Our data revealed nad1 of complex I, cob of complex III, and cox1, 2 and 3 of complex IV, providing proteomic evidence for all complexes consisting of mitochondrial-encoded subunits (Suppl. Table 2). In trypanosomes, recovery of these complexes has required specific extraction procedures (Horváth et al. 2000; Škodová-Sveráková et al. 2015), and hence detection of five mitochondrially-encoded proteins is notable and further demonstrates the quality of our dataset. Since the transcriptome was assembled from polyadenylated RNAs (Ebenezer et al. 2019), the presence of cox3 was surprising (Suppl. Table 4). The status of RNA polyadenylation in the *E. gracilis* mitochondrion is currently unknown (Dobáková et al. 2015), but it is assumed that in most eukaryotes the polyA tails in mitochondrial transcripts are generally shorter than their nuclear-encoded counterparts (Chang and Tong 2012). While the C-terminal end of cox1 is nuclear-encoded, arising from an ancient gene fission event (Gawryluk and Gray 2010), no sequences showing homology to this region were found in the transcriptome. Cox3-like sequences are also present in the transcriptomes of Yoshida and

O'Neill, both of which show a sequence of extended length, whose N-terminal region contains a predicted mitochondrial target peptide. Thus, we suggest *cox3* is also dual-encoded, and that the recovered transcript in fact represents the nuclear encoded region of *cox3*.

Respiratory complexes

So far, 25 conventional respiratory complex subunits have been detected in *E. gracilis* from complex I along with 14 euglenozoan-specific subunits and 20 unknown proteins (Miranda-Astudillo et al. 2018; Perez et al. 2014). Through our analysis, we recovered 19 of these conventional subunits, along with subunit NDUFA8, which has not been identified previously, seven euglenozoan-specific subunits and ten unknown proteins. Additionally, we found two potentially novel subunits of complex I (Suppl. Table 7). It is plausible that some complex I subunits may represent highly diverged paralogues of conventional subunits so far not identified in euglenids, such as Nad2,3,4L and 6, but presumably with potential to increase functional flexibility.

Of complex III, eight conventional and two euglenozoan-specific subunits have been identified, along with three novel proteins (Miranda-Astudillo et al. 2018; Perez et al. 2014). Within this mitoproteome, we recovered five conventional and one euglenozoan-specific subunit, along with all three novel proteins. Cytochrome *c* was detected as three paralogues, indicating duplication events (Suppl. Table 7). The possibility that *E. gracilis* possesses functional variants of cytochrome *c* is intriguing and recommends itself for further study.

Seven conventional subunits, nine euglenozoan-specific subunits and eight novel proteins have been identified for complex IV (Miranda-Astudillo et al. 2018; Perez et al. 2014). Nine conventional subunits were also detected in the mitoproteome, including five previously identified ones and four subunits (*cox15*, *cox4*, *cox11* and *cox10*) that are new additions. Moreover, we found five euglenozoan-specific subunits and eight previously assigned unknown proteins (Suppl. Table 7). Finally, while seven conventional and eight euglenozoan-specific subunits have previously been found to encompass the ATP synthase (Miranda-Astudillo et al. 2018; Perez et al. 2014), our mitoproteome contains five subunits of both these groups, as well as two OSCP homologues (Suppl. Table 7), which again suggests increased functional flexibility.

MICOS complex

The mitochondrial contact site and cristae-organising system (MICOS) mediates the formation of cristae and increases membrane area available for respiratory complexes (Kozjak-Pavlovic 2017). MICOS supports this function across eukaryotes and may also be involved in protein import into the inner membrane (Kaurov et al. 2018). While only Mic10 was identified in the *E. gracilis* genome (Ebenezer et al. 2017), we report its proteomic detection along with Mic20, Mic40, Mic60 and a putative Mic34 (Suppl. Fig. 5). Mic10 typically carries two transmembrane domains, yet these were not predicted for *E. gracilis* with high confidence. The Mic20 of *E. gracilis* contains, same as its orthologue in trypanosomes, a thioredoxin-like domain (Suppl. Fig. 5), and thus likely also represents a functional analogue to Mia40, which is seemingly absent from excavates (Kaurov et al. 2018). Since *T. brucei* Mic60 lacks a mitofilin domain (Suppl. Fig. 5), which is present in opisthokonts and responsible for interaction with the TOB complex (Kozjak-Pavlovic 2017), it may be unable to fulfil this function (Kaurov et al. 2018). *E. gracilis* Mic60 not only lacks a C-terminal mitofilin domain, but is also 100 amino acids shorter than its trypanosome orthologue, possessing a transmembrane domain near to the N-terminus. Mic34 of both *E. gracilis* and *T. brucei* carries two coiled-coil domains (Suppl. Fig. 5), which in the latter has been suggested to support mitofilin-based interactions with the TOB complex (Kaurov et al. 2018). While Mic34 of *E. gracilis* was initially dismissed because of a high bias HMM score, the presence of two coiled-coil domains (Suppl. Fig. 5) and detection in the mitoproteome fraction, along with its seemingly essential organellar role has led us to conclude this as a genuine component. In comparison to the nine MICOS subunits in *T. brucei*, *E. gracilis* presents itself as an evolutionary intermediate between opisthokonts and the expanded and diverged MICOS apparatus of trypanosomes.

Alternative respiratory pathway

In the bloodstream stage of *T. brucei*, an alternative respiratory pathway, composed of glyceraldehyde 3-phosphate dehydrogenase (G3PD), alternative oxidase (AOX) and a type II alternative NADH dehydrogenase (Verner et al. 2015) is essential. The pathway is present in other life stages in conjunction with the oxidative phosphorylation machinery, where it fosters greater metabolic flexibility in response to nutrient and oxygen availability

(Chaudhuri et al. 2006), and also likely plays a role in regulating the level of reactive oxygen species (Fang and Beattie 2003). Both trypanosome-like AOX and G3PD have been identified in *E. gracilis* previously (Miranda-Astudillo et al. 2018; Perez et al. 2014), as well as here (Suppl. Table 7). For the first time we also report the detection of type II NADH dehydrogenase. The presence of a full alternative respiratory pathway and multiple paralogues in *E. gracilis* further encourages the view of its mitochondrion as highly versatile and adaptable to a variety of environmental conditions.

Protein import - outer mitochondrial membrane

Opisthokonts make use of a translocation of outer membrane (TOM) complex to translocate proteins across the outer mitochondrial membrane (Mani et al. 2016). In kinetoplastids, a diverged atypical translocation of outer membrane (ATOM) complex fulfils the same role (Schneider 2018). Both complexes consist of a central pore-forming subunit, two or more protein receptors and many smaller subunits (Mani et al. 2016).

The *E. gracilis* mitoproteome contains 22 proteins associated with import and subsequent processing in the mitochondria that were recovered at the protein level, with an additional 12 identified *in silico* (Suppl. Table 7). There are two orthologues of 40 kDa β -barrel pore-forming ATOM (Fig. 3), which is essential in related species for proper mitochondrial import through the outer membrane (Eckers et al. 2012). Homologues to ATOM46 and 69 are also present, which in trypanosomes display preference for hydrophobic carrier proteins and pre-sequence containing substrates, respectively (Mani et al. 2016). Another characteristic subunit of the *T. brucei* complex, the peripheral receptor pATOM36, which mediates insertion and assembly of N-terminal anchored outer membrane proteins (Schneider 2018) and was previously assumed to be trypanosome-specific, was also confirmed at protein level (Fig. 3). A homologue of TOM34, which in opisthokonts serves as a co-chaperone with heat shock proteins for mitochondrial translocation of sequences (Faou and Hoogenraad 2012), was identified *in silico* (Fig. 3). While a protein similar to ATOM19 was predicted within transcriptome and also recovered in the verified mitoproteome, only weak hits were found for ATOM11 and 12 (Fig. 3). Given that opisthokonts also contain four small, stabilizing subunits within their TOM complex (TOM5,6,7 and 22, corresponding to ATOM11, 12, 14 and 19 in trypanosomatids) (Schneider 2018), it is likely that they are present in *E. gracilis*, but have diverged beyond recognition. Peculiarly, a homologue to ATOM14 and TOM22,

which represents the single subunit conserved between opisthokonts and kinetoplastids (Schneider 2018), could not be identified (Fig. 3).

TOB55 inserts β -barrel-containing proteins into the outer mitochondrial membrane from the intermembrane space. In *E. gracilis*, two orthologues of TOB55 were detected by proteomics (Suppl. Table 7), reminiscent of the duplicated TOB55 of *Trypanosoma cruzi* (Eckers et al. 2012), while TOB38 was identified only *in silico*. TOB55 and TOB38 are essential in fungi (Sharma et al. 2010), while the third subunit of the TOB55 complex, SAM37, is dispensable (Desmond et al. 2011). Indeed, this subunit appears absent from *E. gracilis*. Associating with the TOB55 complex as a peripheral protein and with an undefined function are the metaxins, which have been identified in both opisthokonts and trypanosomes (Verner et al. 2015). Two orthologues of metaxin were identified in the *E. gracilis* transcriptome, one of which is homologous to SAM37 (Suppl. Table 7).

Protein import - intermembrane space, inner mitochondrial membrane and matrix

Translocation of inner membrane (TIM) protein TIM9 of the hexameric TIM9/10 complex was identified at the protein level, while only *in silico* evidence is available for its binding partner TIM10 (Fig. 3). The two together traditionally serve to chaperone hydrophobic precursors across the intermembrane space (Wiedemann et al. 2004). We found no orthologues for TIM complex 11/13 or TIM8 and tiny TIMs (Fig. 3). The TIM23 complex imports signal-bearing proteins into the inner membrane and the mitochondrial matrix (Wiedemann et al. 2004). We provide mass spectrometric evidence for the existence, in *E. gracilis*, of this complex containing orthologues of TIM16, TIM17, TIM23, TIM44, TIM50, Hsp70 and Mge1 (Fig. 3). TIM17, TIM23 and TIM50 constitute the membrane-anchoring component, with TIM50 traditionally working as a receptor for precursors, while TIM17 and TIM23 serve as the pore-forming units (Schneider 2018). Moreover, TIM14 and TIM15 were identified *in silico* (Fig. 3), and presumably serve as a part of the motor protein complex and prevent protein aggregation (Fraga et al. 2013).

All trypanosomes studied so far prominently lack the TIM23 pore-forming subunit, presumably forming a channel only via TIM17, suggesting that *E. gracilis* contains a less restricted import apparatus. The presence of two orthologues of TIM17, confirmed by mass spectrometry in *E. gracilis*, is similar to *T. cruzi* (Eckers et al. 2012), suggesting that a specific duplication event likely occurred in the common euglenozoan ancestor. Moreover,

the absence of TIM21 and Pam17 in both protist species implies that they represent Euglenozoa-specific losses (Verner et al. 2015). *T. brucei* exhibits an expanded set of membrane-anchored subunits (TIM47, TIM54 and TIM62) with undefined functions, of which only TIM47 was identified in our data with confidence (Suppl. Table 7). Of the integral membrane proteins TIM42, RHOM 1 and RHOM 2, all of which have been shown to be essential to *T. brucei*, only a weakly homologous sequence to TIM42 was recovered (Harsman et al. 2016).

The TIM22 complex, which in opisthokonts inserts proteins into the inner mitochondrial membrane (Mokranjac and Neupert 2009), seems to be absent in trypanosomes (Schneider 2018). Hence the detection by mass spectrometry of an orthologue to the main functional subunit TIM22 (Fig. 3) supports the idea of a relatively expanded import system in euglenids, when compared to its highly diverged and streamlined version in trypanosomes. However, other subunits of the TIM22 complex were not found, so that the functional significance remains unclear (Fig. 3).

The membrane-anchored Oxa1 is an insertase of the inner membrane (Mokranjac and Neupert 2009), and whilst an orthologue was identified in the mitoproteome (Fig. 3), its binding partner Mba1 was not. Proteins translocated via Oxa1 are associated with removal of intermembrane space sorting signal, undertaken by inner membrane peptidases (IMPs) (Gakh et al. 2002). An IMP2 orthologue is also present but as in *T. brucei* (Verner et al. 2015), IMP1 seems to be absent (Fig. 3). In opisthokonts, amoebozoans and plants, the Erv1-Mia40 complex imports and oxidatively folds small cysteine-rich proteins in the intermembrane space (Allen et al. 2008). While Erv1 is present in *E. gracilis*, Mia40 could not be identified and is also absent in parasitic chromalveolates (Ebenezer et al. 2019). As proposed for trypanosomes, Mia40 may be functionally complemented by a dedicated subunit of the MICOS complex (Kaurov et al. 2018). *E. gracilis* is notable in that it represents the first non-parasitic excavate lacking Mia40, which supports the notion that this protein is absent from euglenozoans (Allen et al. 2008), and may possibly import intermembrane proteins by an as yet unknown pathway.

Proteins transported through the TIM23 complex require removal of the import signal, mediated by the mitochondrial processing peptidase (MPP) complex consisting of α and β subunits. MPP- α recognises the pre-sequence while MPP- β performs the cleavage (Gakh et al. 2002). As in *T. brucei* (Desy et al. 2012), both proteins were recovered from the

mitoproteome (Fig. 3). Once cleaved, some proteins will self-assemble, while others require assistance, and are folded by heat shock protein (Hsp) 60 and 10 (Martin 1997). Hsp10 was identified *in silico*, while Hsp60 was detected by mass spectrometry (Fig. 3).

TCA cycle

The tricarboxylic acid (TCA) cycle is a key component of all aerobic mitochondria, allowing efficient generation of energy. *E. gracilis* employs a modified TCA cycle reminiscent of certain alpha-proteobacteria, where the oxoglutarate dehydrogenase complex is replaced with 2-oxoglutarate decarboxylase and succinate semialdehyde dehydrogenase (Green et al. 2000). TCA cycle and associated components show numerous duplication events with multiple copies of key enzymes (Suppl. Table 7). Under aerobic conditions, pyruvate is transported to the mitochondrion and metabolised by the pyruvate dehydrogenase complex (PDH), producing acetyl-coenzyme A (CoA), which enters the TCA cycle (Hoffmeister et al. 2004). All four PDH subunits were recovered *in vivo* (Ebenezer et al. 2019) (Fig. 4A). The first three steps of the TCA cycle, following the generation of acetyl-CoA, follow a conventional path and all enzymes from these steps (citrate synthase, aconitase, isocitrate dehydrogenase) were recovered by mass spectrometry (Fig. 4A). Notably, isocitrate dehydrogenase was among the identified proteins initially rejected based on low mitochondrial enrichment compared to the whole cell, but was accepted based on functional annotation.

Traditionally, α -ketoglutarate is catabolised using the α -ketoglutarate dehydrogenase complex, before employing the succinyl-CoA synthetase complex to generate succinate (Zimorski et al. 2017). Again, canonical subunits of succinyl-CoA synthase are present in the mitoproteome, as well as one of three conventional subunits (E3 dihydrolipoyl dehydrogenase) of the α -ketoglutarate dehydrogenase, the other two being absent from the *E. gracilis* genome (Ebenezer et al. 2019), suggesting the pathway is either non-functional or uses unconventional components (Fig. 4). Retention of both E3 and succinyl-CoA subunits for a seemingly non-functional pathway can be explained through their bifunctionality, since E3 exists as a subunit of the PDH complex, while succinyl-CoA synthase also acts in the methylmalonyl-CoA pathway for fatty acid synthesis (Fig. 4A). *E. gracilis* bypasses the need for the α -ketoglutarate dehydrogenase complex through the use of α -ketoglutarate decarboxylase and succinate semialdehyde dehydrogenase. α -ketoglutarate is first catalysed to succinate semi-aldehyde (SSA) by α -ketoglutarate decarboxylase and then converted to

succinate via SSA dehydrogenase (Müller et al. 2012). We find evidence for both essential enzymes of this pathway (Fig 4A.).

A second variant pathway, the gamma-aminobutyric (GABA) shunt, employs glutamate dehydrogenase, glutamate decarboxylase and GABA transaminase to generate SSA, which in turn is converted to succinate by SSA dehydrogenase (Zhang et al. 2016). In the mitoproteome, glutamate dehydrogenase and GABA transaminase were detected, while glutamate decarboxylase remains identified only *in silico* (Fig. 4A). The GABA shunt allows for optimised photosynthetic conditions in cyanobacteria (Nogales et al. 2012), which may also be applicable within *E. gracilis*.

After generation of succinate, the *E. gracilis* TCA cycle proceeds conventionally with succinate dehydrogenase (respiratory complex II) functioning with fumarate hydratase and malate dehydrogenase to complete the cycle, with both of these proteins detected by mass spectrometry. Succinate dehydrogenase is composed of two conserved subunits, one of which, SDH2, is further split into two components (Gawryluk and Gray 2019). These conserved subunits, along with three of eight euglenozoan-specific subunits (SDHTC 6,7, and 8) were identified (Fig. 4A). Two genes encoding for euglenozoan-specific subunit 7 were present, likely representing a recent duplication (Suppl. Table 7).

Notably, *E. gracilis* is one of few organisms to localise the glyoxylate cycle to the mitochondria, as opposed to glyoxysomes (Zimorski et al. 2017). The glyoxylate cycle represents an anabolic variant of the TCA, which enables synthesis of complex carbohydrates from compounds such as lipids, with both cycles sharing several enzymes. In general, the glyoxylate cycle proceeds with the first two TCA reactions, yielding isocitrate, which is then cleaved by isocitrate lyase to form glyoxylate and succinate (Zimorski et al. 2017).

Subsequently, glyoxylate is converted to malate *via* malate synthase, while succinate is converted to fumarate by fumarate reductase and further metabolised into malate by fumarate hydratase again. By the activity of malate dehydrogenase, malate is converted into oxaloacetate, which can re-enter the TCA cycle, or be processed into phosphoenolpyruvate by phosphoenolpyruvate carboxykinase, initiating gluconeogenesis (Dolan and Welch 2018). To facilitate the glyoxylate cycle, *E. gracilis* uses a bifunctional enzyme, containing domains for both isocitrate lysis and malate synthesis (Nakazawa et al. 2011), which was present within the mitoproteome, while fumarate reductase was identified only *in silico* (Fig. 4B). Phosphoenolpyruvate carboxykinase was recovered at the protein level, initially rejected for

low enrichment, yet eventually recalled (Fig. 4B). The *E. gracilis* mitochondrion can additionally initiate gluconeogenesis by converting pyruvate directly to oxaloacetate *via* pyruvate carboxylase, which was proteomically detected (Suppl. Table 7), and is then processed through the activity of phosphoenolpyruvate carboxykinase. Our analysis of the TCA cycle and associated pathways demonstrates a highly versatile organelle, able to generate and store ATP using a variety of energy sources under a range of conditions; the presence of paralogues for several critical enzymes also reinforces the potential for environmental flexibility.

Mitochondrial ribosomes

108 mitochondrial ribosomal (mitoribosomal) subunits were identified *in silico*, representing 68 additional subunits in comparison to the earlier reported complement (Ebenezer et al. 2019). A total of 83 of these were experimentally verified, consisting of 47 large and 34 small mitoribosomal subunits with two unclassified (Suppl. Table 7). From a phylogenetic perspective, these are further classified into a variety of categories: ‘Core’ subunits are of alpha-proteobacterial origin, likely present in the last common eukaryotic ancestor while ‘accessory’ subunits subsequently emerged in different eukaryotic lineages. 35 core and 15 accessory mitoribosomal proteins were recovered, with an additional six and two *in silico*-predicted but not detected, respectively. Compared with the well-studied mitoribosomes of *T. brucei*, 20 subunits (L5, L14, L29, L30, L33, L35, L41, L2, L48, L53, S6, S14, S21, S26, S30, S33, S35, S37, S38, Fyv4) appear lost in euglenids yet retained in trypanosomes (Zíková et al. 2008), whereas 10 subunits (L1, L6, L18, L38, L56, S2, S3, S7, S13, S28) are present solely in the *E. gracilis* mitoribosome (Suppl. Table 7). These compositional distinctions potentially underpin major differences in protein synthesis between euglenids and kinetoplastids.

Kinetoplastids also contain a uniquely large number of novel mitoribosomal subunits, indicating expansions that occurred later in euglenozoan evolution (Desmond et al. 2011). Based on structural analysis, many have been reclassified as core or accessory subunits (Ramrath et al. 2018), with only one found outside of kinetoplastids, in *Naegleria gruberi* (Desmond et al. 2011). Of the 79 kinetoplastid-specific subunits, 15 were recovered in the *E. gracilis* mitoproteome, with a further 20 predicted within the transcriptome (Suppl. Table 7). This finding rules out the possibility that these subunits developed in response to a parasitic

lifestyle, as previously proposed (Desmond et al. 2011). Of shared mitoribosomal proteins, a greater percentage of predicted small ribosomal subunit proteins (7/12) were recovered *in vivo* when compared to the large subunit (7/23). One ribosomal protein identified in *T. brucei* that localises to both large and small subunits (Zíková et al. 2008) was also recovered here (Suppl. Table 7).

A total of 18 mitoribosomal proteins, which did not match previously defined subunit groups were present either in the mitoproteome or predicted *in silico*, and are therefore potential novel *E. gracilis* subunits. Twelve and four of them were predicted to reside in large and small mitoribosomal subunits respectively, while two were unclassified (Suppl. Table 7). Finally, eight putative homologues of mitoribosomal proteins were identified in *E. gracilis*. Given high divergence generally encountered among this category of proteins (Petrov et al. 2019), and the fact that these eight proteins with E-values above the stated cut-off were detected by mass spectrometry, we classify them as putative mitoribosomal components (Suppl. Table 7). Our results suggest that considerable mitoribosome protein expansion occurred already in the common euglenozoan ancestor (Desmond et al. 2011).

Aminoacyl-tRNA synthetases

Aminoacyl-tRNA synthetases (aaRSs) are responsible for attaching amino acids to their cognate tRNAs (de Duve 1988). Two classes exist, which differ in their domain architecture, ATP binding conformation and modes of aminoacylation (Burbaum and Schimmel 1991). Along with direct aminoacylation carried out by aaRSs, some archaea, bacteria and chloroplasts lacking GlnRS and AsnRS enzymes possess indirect aminoacylation pathways, where a non-discriminating GluRS and tRNA-dependent amidotransferase (AdT) carries out biosynthesis of Gln-tRNA^{Gln} and Asn-tRNA^{Asn} (Schön et al. 1988). We identified 46 transcripts encoding all class I and II aaRSs in *E. gracilis* (Suppl. Table 8). Each of the identified proteins had at least one paralogue (or isoform) with predicted mitochondrial localisation, except ArgRS and TrpRS, likely missing in the mitochondrion, while MetRS, IleRS and GluRS appear to be dually targeted to the mitochondrion and chloroplast (Suppl. Table 8). Moreover, each of these aaRSs has an additional paralogue of low sequence identity (varying from 24 to 44%) targeted exclusively to the mitochondrion. The identification of several aaRSs in the mitochondrial fraction suggests that some of them might have unconventional functions (Francklyn et al. 2002). The presence of GlnRS and AsnRS along

with the absence of clear homologues of AdTs in any of the publicly available *E. gracilis* transcriptomes indicates that only the direct aminoacylation pathway is functional in both organelles and cytoplasm. This is in agreement with recent identification of a plastidial version of GlnRS in the non-photosynthetic *Euglena longa* (Záhonová et al. 2018).

Amino acid metabolism

We identified most of the biosynthesis pathways for 11 amino acids: valine, leucine, isoleucine, aspartate, cysteine, glutamate, glutamine, glycine, serine, proline and tyrosine (Suppl. Fig. 6), suggesting that the repertoire of amino acid biosynthesis enzymes is substantially richer in the mitochondrion of *E. gracilis* than in its plastid (Novák Vanclová et al. 2019). While hydroxymethyltransferase catalyses serine and glycine interconversion in *E. gracilis* mitochondria, glutamate is synthesised from 2-oxoglutarate by the action of aspartate aminotransferase and can be subsequently converted to glutamine or proline (Suppl. Fig. 6), and aspartate can be synthesized from pyruvate using pyruvate carboxylase and aspartate aminotransferase.

The biosynthesis of valine and leucine from pyruvate does not appear possible in the *E. gracilis* mitoproteome, since only one enzyme of the pathway, 2-isopropylmalate synthase, is present. Another branched-chain amino acid, isoleucine, can be obtained in a five-step reaction from threonine, while cysteine is synthesized *de novo* from serine, and tyrosine is formed from phenylalanine by phenylalanine hydroxylase (Suppl. Fig. 6).

The use of amino acids as an energy source in mitochondrion in *E. gracilis* is limited to the ability to degrade valine, leucine, isoleucine, glutamate, glycine, proline, aspartate, alanine and possibly cysteine (Suppl. Table 8). Almost all enzymes of the branched-chain amino acids degradation pathway leading to formation of acetyl-CoA, in the case of leucine and isoleucine, and succinyl-CoA or valine, were identified. Proline is converted to glutamate in a two-step reaction carried out by proline dehydrogenase and δ -1-pyrroline-5-carboxylate dehydrogenase, and glutamate can then be degraded either to α -ketoglutarate and ammonia or to succinate. In contrast to human mitochondria, it appears that cysteine degradation in the organelle of *E. gracilis* leads to pyruvate formation. We identified a putative homologue of thiosulphate/3-mercaptopyruvate sulphurtransferase, catalysing the second step of this process. Cysteine aminotransferase was not detected, however, leaving the possibility that the initial reaction of the process is catalysed by aspartate aminotransferase, similar to some

bacteria (Andreeßen et al. 2017). While glycine is oxidatively cleaved by the glycine cleavage system, the saccharopine pathway of lysine degradation appears to be non-functional, since only three enzymes out of nine, were identified (Suppl. Table 8).

Sulphate assimilation

A remarkable feature of the *E. gracilis* mitochondrion is the ability to assimilate and metabolise sulphate (Saidha et al. 1988). This represents an exceptionally rare feature among eukaryotes, as the vast majority employ the sulphate assimilation pathway in either resident plastids or the cytosol. Some of the enzymes involved in this process have a plastid origin, but over time have been re-targeted to the mitochondria, presumably after their genetic incorporation into the nucleus (Patron et al. 2008). Components of the sulphate assimilation pathway have additionally been found in the mitosomes of *E. histolytica*, though these are thought to have been acquired independently from ϵ -proteobacteria and other sources (Miichi et al. 2009). This pathway is important for the generation of sulphur-containing amino acids such as cysteine, methionine and various other metabolites, which supply necessary sulphur for the Fe-S cluster generation in euglenids (Saidha et al. 1988). Except for sulphate permease and cysteine synthase, all enzymes for sulphate assimilation were identified by proteomics (Suppl. Table 7; Suppl. File 1).

Fe-S cluster biosynthesis

Iron sulphur (Fe-S) cluster (ISC) biosynthesis provides essential and ubiquitous co-factors for many proteins (Lill 2009). In most eukaryotes, the first step, referred to as ISC biosynthesis, is located in the mitochondrion (Braymer and Lill 2017). Amongst euglenozoans, Fe-S cluster assembly is best studied in *T. brucei* while in the euglenids and diplomonids the pathway remains unexplored (Ali and Nozaki 2013; Peña-Díaz and Lukeš 2018). We find conservation of all predicted central components of this pathway, with some notable expansions.

The ISC pathway consists of three central steps, starting with Fe-S cluster assembly on a scaffold protein IscU. Sulphur is donated by cysteine desulphurase NfsI and Isd11, after the cluster is reduced by ferredoxin in coordination with ferredoxin reductase (Ollagnier-de-Choudens et al. 2001). Iron (II) cation is imported into the matrix via mitochondria carrier protein 17 (Braymer and Lill 2017), where it is donated to the scaffold with the assistance of

frataxin (Bridwell-Rabb et al. 2014). In *E. gracilis*, two ferredoxin orthologues were identified (Fig. 5), corresponding to *T. brucei* ferredoxin A and B (Changmai et al. 2013). The second step involves detachment of the Fe-S cluster from the IscU scaffold. Grx5, Mge1, Ssc1 and Hep1 assist as chaperones, transporting the cluster to the relevant apoprotein (Maio et al. 2014). Peptides for Grx5, Mge1 and Ssc1 were identified at the protein level, with Hep1 being identified only *in silico* (Fig. 5). The final step involves alternative scaffold proteins Isa1 and Isa2, with assistance from Iba57, which insert the Fe-S clusters into various apoproteins (Sheftel et al. 2012). All three sequences were confirmed at a protein level (Fig. 5).

Depending on protein specificity, various factors are subsequently required for holoprotein maturation, such as Nfu, involved in maturation of respiratory complexes I and II components (Lukeš and Basu 2015). In *E. gracilis*, four Nfu factors are identifiable from the transcriptome, with two being recovered as peptides (Fig. 5; Suppl. Table 7). One Nfu homologue was initially rejected due to insufficient enrichment, but reintegrated based on functional prediction. Most eukaryotes contain only a single Nfu protein, while expansion into three and four Nfus occurred in *T. brucei* and *Leishmania* spp., respectively, of which most are mitochondrially localised (Benz et al. 2016). *A. thaliana* contains a higher number of Nfu paralogues, but most are localized in the plastid (Léon et al. 2003). By contrast, in *E. gracilis* both Nfu proteins show mitochondrial enrichment when compared to the chloroplast (Suppl. Table 4). The Fe-S clusters bound for export out of the mitochondrion are transported by Atm1 and Erv1 (Braymer and Lill 2017), both being identified *in silico*, and Atm1 also at the protein level (Fig. 5).

Fatty acid metabolism

Under anaerobic conditions, the mitochondrion of *E. gracilis* can employ acetyl-CoA as a terminal electron acceptor and synthesise fatty acids in a malonyl-independent manner, enabling net ATP production (Hoffmeister et al. 2005). Acetyl-CoA is first produced from the catalysis of pyruvate, mediated by an oxygen-sensitive pyruvate:NADP⁺ oxidoreductase complex (PNO) (Zimorski et al. 2017). Both α and β subunits of the PNO complex were recovered at a protein level (Suppl. Table 7). Although the β subunit showed low mitochondrial enrichment, likely due to the oxygen-sensitive nature of PNO (Rotte et al. 2001) and the aerobic conditions of this study, it was included based on strong prediction

criteria. The synthesis of fatty acids involves reversal of the β -oxidation of acetyl-CoA pathway. Several enzymes involved in this process can work in both directions (Zimorski et al. 2017). In *E. gracilis* acetyl-CoA is first condensed by acetyl-CoA-C-acetyltransferase, then reduced via β -hydroxyacyl-CoA and dehydrated by enoyl-CoA hydratase. The final step is mediated by unique *trans*-2-enoyl CoA reductase, as the enzyme responsible for the reverse reaction (acyl-CoA dehydrogenase) operates catalytically in one direction (Zimorski et al. 2017). The activity of *trans*-2-enoyl CoA reductase produces elongated fatty acyl-CoA as an end product, which is reduced to alcohol, esterified and deposited in the cytosol as wax esters (Hoffmeister et al. 2005). Upon returning to aerobic conditions, in the mitochondrion, these wax esters can be converted back to acetyl-CoA (Zimorski et al. 2017).

We have detected peptides for all four enzymes required for fatty acid synthesis, while oxidative acyl-CoA dehydrogenase was identified only by a homology search. Although *E. gracilis* was cultivated in aerobic conditions, *trans*-2-enoyl CoA reductase, which is exclusively used for anaerobic fatty acid synthesis, was detected (Suppl. Table 7). Its expression under aerobic conditions is in agreement with previous studies and reflects the adaptability of *E. gracilis* for environments deprived of oxygen (Hoffmeister et al. 2004). All proteins involved in long-chain acyl-CoA import into the mitochondrion, as well as components for odd-numbered fatty acid synthesis (excepting fumarate reductase and propionyl-CoA subunits), were also proteomically recovered (Suppl. File 1).

In-silico mitoproteome of *Eutreptiella gymnastica*

To gain a better understanding of mitoproteome evolution within euglenidae and to resolve the seemingly unique features of *E. gracilis*, we compared the verified mitoproteome to the publicly available transcriptome of fellow photosynthetic euglenid *Eutreptiella gymnastica*, the outcome of which yielded 1,162 transcripts corresponding to 900 orthologue groups (Suppl. Table 9). 620 of these *E. gymnastica* transcripts were orthologous to functionally uncategorised sequences of *E. gracilis*, which suggests shared heritage to a large expansion of these unknown genes within the ancestor of photosynthetic euglenids.

While analysis of the of the *E. gracilis* transcriptome identified an extensive 192 mTERF factors identified in *E. gracilis*, no orthologues could be identified from *E. gymnastica*. Comparison with the transcriptome of *E. gymnastica* indicated conservation of all components of the alpha-proteobacterial shunt, while only orthologues to GABA

transaminase of the GABA shunt could be identified (Fig. 6; Suppl. Table 9). Components necessary for the glyoxylate cycle were not identified, suggesting this pathway may be unique to *E. gracilis*. Interestingly, Hidden Markov model searches identified a sequence above the detection threshold to subunit E2 of oxoglutarate dehydrogenase (Suppl. Table 9), suggesting that the conventional TCA pathway could in fact be functional in *E. gymnastica*.

The recovery of inner membrane translocation pores TIM22 and TIM23 in *E. gracilis* represents a first for Euglenozoan mitochondria, and their identification within the transcriptome of *E. gymnastica* suggests they are distributed throughout euglenida as well (Fig. 6; Suppl. Table 9). While four protein maturation factors were identified for *E. gracilis*, no Nfu factors could be identified from the *E. gymnastica* transcriptome, suggesting that this high Nfu number may be confined to *E. gracilis*.

Orthologues to the majority of sulphate assimilation enzymes (seven of nine) were additionally identified in the *E. gymnastica* transcriptome (Suppl. Table 9), suggesting that the mitochondrial re-targeting of sulphate occurred in the common ancestor of these two species. Additionally, orthologues for the majority of components for fatty acid synthesis and import (nine of seventeen) were also identified, including crucial enzyme *trans*-2-enoyl CoA reductase (Suppl. Table 9), suggesting that malonyl-independent fatty acid synthesis is a conserved feature in euglenids (Fig. 6).

Conclusions

The exceptionally rich mitoproteome of *E. gracilis* can be partially attributed to extensive duplications, including mitoribosomal subunits and the alternative oxidase pathway as well as core metabolic pathways where multiple paralogues may indicate considerable flexibility. However, the main contributing factor to this unique complexity is the exceptionally large fraction of proteins with unknown function. Additionally, over 700 predicted proteins such as the extensive mTERF family still require validation through additional functional studies. While we report the identification of various anaerobic enzymes such as *trans*-enoyl-reductase and oxygen-sensitive PNO, it is likely that many strictly anaerobic proteins have eluded detection due to the aerobic nature of the extracted mitochondria. Subsequently, a study of the organelle from the anaerobically grown *E. gracilis* is a logical successive step.

Characterisation of the *E. gracilis* mitoproteome represents an important step in our understanding of the uniquely rich evolution of this organelle in different euglenozoan lineages, and enables the prediction of mitochondrial traits present in the euglenozoan common ancestor. Such traits include newly identified mitoribosomal and MICOS subunits, protein import machinery, respiratory proteins, components of the Fe-S cluster biosynthesis as well as the prominent absence of RNA editing and processing machinery (Fig. 6). Combined, the mitoproteome of *E. gracilis* demonstrates an unparalleled protein count, which may reflect the influence of the co-occurring plastid, and reveals a remarkable metabolic flexibility and adaptability.

Acknowledgements

We thank Zdeněk Verner, Laurie Read and Michael Gray for helpful comments. This work was supported by an ERC CZ grant (LL1601), the Czech Grant Agency No. 18-15962S, the Czech Ministry of Education (ERD Funds, project OPVVV/0000759) to JL and the Wellcome Trust (204697/Z/16/Z) to MCF.

References

- Adl SM, Bass D, Lane CE, Lukeš J, Schoch CL, Smirnov A, Agatha S, Berney C, Brown MW, Burki F et al. 2019. Revisions to the Classification, Nomenclature, and Diversity of Eukaryotes. *Journal of Eukaryotic Microbiology* 66(1):4-119.
- Ali V, Nozaki T. 2013. Iron-Sulphur Clusters, Their Biosynthesis, and Biological Functions in Protozoan Parasites. *Advances in Parasitology* 83:1-92.
- Allen JWA, Ferguson SJ, Ginger ML. 2008. Distinctive biochemistry in the trypanosome mitochondrial intermembrane space suggests a model for stepwise evolution of the MIA pathway for import of cysteine-rich proteins. *FEBS Letters* 582(19):2817-2825.
- Andreeßen C, Gerlt V, Steinbüchel A. 2017. Conversion of cysteine to 3-mercaptopyruvic acid by bacterial aminotransferases. *Enzyme and Microbial Technology* 99:38-48.
- Aslett M, Aurrecochea C, Berriman M, Brestelli J, Brunk BP, Carrington M, Depledge DP, Fischer S, Gajria B, Gao X et al. 2010. TriTrypDB: a functional genomic resource for the Trypanosomatidae. *Nucleic Acids Research* 38:D457-D462.
- Attea A, Adrait A, Brugière S, Tardif M, van Lis R, Deusch O, Dagan T, Kuhn L, Gontero B, Martin W et al. 2009. A Proteomic Survey of *Chlamydomonas reinhardtii* Mitochondria Sheds New Light on the Metabolic Plasticity of the Organelle and on the Nature of the alpha-Proteobacterial Mitochondrial Ancestor. *Molecular Biology and Evolution* 26(7):1533-1548.
- Ba ANN, Pogoutse A, Provart N, Moses AM. 2009. NLStradamus: a simple Hidden Markov Model for nuclear localization signal prediction. *BMC Bioinformatics* 10.

- Benz C, Kovářová J, Králová-Hromadová I, Pierik AJ, Lukeš J. 2016. Roles of the Nfu Fe-S targeting factors in the trypanosome mitochondrion. *International Journal for Parasitology* 46(10):641-651.
- Boussau B, Karlberg EO, Frank AC, Legault BA, Andersson SGE. 2004. Computational inference of scenarios for alpha-proteobacterial genome evolution. *Proceedings of the National Academy of Sciences of the United States of America* 101(26):9722-9727.
- Braymer JJ, Lill R. 2017. Iron-sulfur cluster biogenesis and trafficking in mitochondria. *Journal of Biological Chemistry* 292(31):12754-12763.
- Bridwell-Rabb J, Fox NG, Tsai CL, Winn AM, Barondeau DP. 2014. Human Frataxin Activates Fe-S Cluster Biosynthesis by Facilitating Sulfur Transfer Chemistry. *Biochemistry* 53(30):4904-4913.
- Burbaum JJ, Schimmel P. 1991. Structural relationships and the classification of aminoacyl-transfer RNA-synthetases. *Journal of Biological Chemistry* 266(26):16965-16968.
- Calvo SE, Clauser KR, Mootha VK. 2016. MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins. *Nucleic Acids Research* 44(D1):D1251-D1257.
- Carrie C, Small I. 2013. A reevaluation of dual-targeting of proteins to mitochondria and chloroplasts. *Biochimica Et Biophysica Acta-Molecular Cell Research* 1833(2):253-259.
- Casaletti L, Lima PS, Oliveira LN, Borges CL, Báo SN, Bailão AM, Soares CMA. 2017. Analysis of *Paracoccidioides lutzii* mitochondria: a proteomic approach. *Yeast* 34(4):179-188.
- Chang JH, Tong L. 2012. Mitochondrial poly(A) polymerase and polyadenylation. *Biochimica et Biophysica Acta-Genetic Regulatory Mechanisms* 1819(9-10):992-997.
- Chang JS, Van Remmen H, Cornell J, Richardson A, Ward WF. 2003. Comparative proteomics: characterization of a two-dimensional gel electrophoresis system to study the effect of aging on mitochondrial proteins. *Mechanisms of Ageing and Development* 124(1):33-41.
- Changmai P, Horáková E, Long SJ, Černotikova-Stříbrná E, McDonald LM, Bontempi EJ, Lukeš J. 2013. Both human ferredoxins equally efficiently rescue ferredoxin deficiency in *Trypanosoma brucei*. *Molecular Microbiology* 89(1):135-151.
- Chaudhuri M, Ott RD, Hill GC. 2006. Trypanosome alternative oxidase: from molecule to function. *Trends in Parasitology* 22(10):484-491.
- Combet C, Blanchet C, Geourjon C, Deléage G. 2000. NPS@: Network Protein Sequence Analysis. *Trends in Biochemical Sciences* 25(3):147-150.
- Cox J, Mann M. 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology* 26(12):1367-1372.
- de Duve C. 1988. Transfer RNAs: the second genetic code. *Nature* 333(6169):117-8.
- Desmond E, Brochier-Armanet C, Forterre P, Gribaldo S. 2011. On the last common ancestor and early evolution of eukaryotes: reconstructing the history of mitochondrial ribosomes. *Research in Microbiology* 162(1):53-70.
- Desy S, Schneider A, Mani J. 2012. *Trypanosoma brucei* has a canonical mitochondrial processing peptidase. *Molecular and Biochemical Parasitology* 185(2):161-164.
- Dobáková E, Flegontov P, Skalický T, Lukeš J. 2015. Unexpectedly Streamlined Mitochondrial Genome of the Euglenozoan *Euglena gracilis*. *Genome Biology and Evolution* 7(12):3358-3367.
- Dolan SK, Welch M. 2018. The Glyoxylate Shunt, 60 Years On. In: Gottesman S, editor. *Annual Review of Microbiology*, Vol 72. Palo Alto: Annual Reviews. p. 309-330.
- Ebenezer TE, Carrington M, Lebert M, Kelly S, Field MC. 2017. *Euglena gracilis* Genome and Transcriptome: Organelles, Nuclear Genome Assembly Strategies and Initial Features. *Euglena: Biochemistry, Cell and Molecular Biology* 979:125-140.
- Ebenezer TE, Zoltner M, Burrell A, Nenarokova A, Vanclova A, Prasad B, Soukal P, Santana-Molina C, O'Neill E, Nankissoor NN et al. 2019. Transcriptome, proteome and draft genome of *Euglena gracilis*. *BMC Biology* 17:11.

- Eckers E, Cyrklaff M, Simpson L, Deponte M. 2012. Mitochondrial protein import pathways are functionally conserved among eukaryotes despite compositional diversity of the import machineries. *Biological Chemistry* 393(6):513-524.
- Eddy SR. 2009. A new generation of homology search tools based on probabilistic inference. *Genome Inform. Japan*. p. 205-11.
- El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A et al. 2019. The Pfam protein families database in 2019. *Nucleic Acids Research* 47(D1):D427-D432.
- Emanuelsson O, Brunak S, von Heijne G, Nielsen H. 2007. Locating proteins in the cell using TargetP, SignalP and related tools. *Nature Protocols* 2(4):953-971.
- 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(4):1005-1016.
- Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* 16.
- Faktorová D, Valach M, Kaur B, Burger G, Lukeš J. 2018. Mitochondrial RNA Editing and Processing in Diplonemid Protists. *RNA Metabolism in Mitochondria* 34:145-176.
- Fang J, Beattie DS. 2003. Alternative oxidase present in procyclic *Trypanosoma brucei* may act to lower the mitochondrial production of superoxide. *Archives of Biochemistry and Biophysics* 414(2):294-302.
- Faou P, Hoogenraad NJ. 2012. Tom34: A cytosolic cochaperone of the Hsp90/Hsp70 protein complex involved in mitochondrial protein import. *Biochimica Et Biophysica Acta-Molecular Cell Research* 1823(2):348-357.
- Flegontov P, Gray MW, Burger G, Lukeš J. 2011. Gene fragmentation: a key to mitochondrial genome evolution in Euglenozoa? *Current Genetics* 57(4):225-232.
- Flegontova O, Flegontov P, Malviya S, Audic S, Wincker P, de Vargas C, Bowler C, Lukeš J, Horák A. 2016. Extreme Diversity of Diplonemid Eukaryotes in the Ocean. *Current Biology* 26(22):3060-3065.
- Fraga H, Papaleo E, Vega S, Velazquez-Campoy A, Ventura S. 2013. Zinc induced folding is essential for TIM15 activity as an mtHsp70 chaperone. *Biochimica Et Biophysica Acta-General Subjects* 1830(1):2139-2149.
- Francklyn C, Perona JJ, Puetz J, Hou YM. 2002. Aminoacyl-tRNA synthetases: Versatile players in the changing theater of translation. *RNA* 8(11):1363-1372.
- Fukasawa Y, Oda T, Tomii K, Imai K. 2017. Origin and Evolutionary Alteration of the Mitochondrial Import System in Eukaryotic Lineages. *Molecular Biology and Evolution* 34(7):1574-1586.
- Fukasawa Y, Tsuji J, Fu SC, Tomii K, Horton P, Imai K. 2015. MitoFates: Improved Prediction of Mitochondrial Targeting Sequences and Their Cleavage Sites. *Molecular & Cellular Proteomics* 14(4):1113-1126.
- Gakh E, Cavadini P, Isaya G. 2002. Mitochondrial processing peptidases. *Biochimica et Biophysica Acta-Molecular Cell Research* 1592(1):63-77.
- Gawryluk RMR, Chisholm KA, Pinto DM, Gray MW. 2014. Compositional complexity of the mitochondrial proteome of a unicellular eukaryote (*Acanthamoeba castellanii*, supergroup Amoebozoa) rivals that of animals, fungi, and plants. *Journal of Proteomics* 109:400-416.
- Gawryluk RMR, Gray MW. 2010. An Ancient Fission of Mitochondrial cox1. *Molecular Biology and Evolution* 27(1):7-10.
- Gawryluk RMR, Gray MW. 2019. A split and rearranged nuclear gene encoding the iron-sulfur subunit of mitochondrial succinate dehydrogenase in Euglenozoa. 2:16.
- Gibson W. 2017. Kinetoplastea. *Handbook of Protists*. Springer International Publishing. p. 1089-1138.
- Gonczarowska-Jorge H, Zahedi RP, Sickmann A. 2017. The proteome of baker's yeast mitochondria. *Mitochondrion* 33:15-21.

- Gray MW. 2015. Mosaic nature of the mitochondrial proteome: Implications for the origin and evolution of mitochondria. *Proceedings of the National Academy of Sciences of the United States of America* 112(33):10133-10138.
- Green LS, Li YZ, Emerich DW, Bergersen FJ, Day DA. 2000. Catabolism of alpha-ketoglutarate by a *sucA* mutant of *Bradyrhizobium japonicum*: Evidence for an alternative tricarboxylic acid cycle. *Journal of Bacteriology* 182(10):2838-2844.
- Guda C, Guda P, Fahy E, Subramaniam S. 2004. MITOPRED: a web server for the prediction of mitochondrial proteins. *Nucleic Acids Research* 32:W372-W374.
- Harsman A, Oeljeklaus S, Wenger C, Huot JL, Warscheid B, Schneider A. 2016. The non-canonical mitochondrial inner membrane presequence translocase of trypanosomatids contains two essential rhomboid-like proteins. *Nature Communications* 7:12.
- Heazlewood JL, Howell KA, Whelan J, Millar AH. 2003. Towards an analysis of the rice mitochondrial proteome. *Plant Physiology* 132(1):230-242.
- 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(1):241-256.
- Hochholdinger F, Guo L, Schnable PS. 2004. Cytoplasmic regulation of the accumulation of nuclear-encoded proteins in the mitochondrial proteome of maize. *Plant Journal* 37(2):199-208.
- Hoffmeister M, Piotrowski M, Nowitzki U, Martin W. 2005. Mitochondrial trans-2-enoyl-CoA reductase of wax ester fermentation from *Euglena gracilis* defines a new family of enzymes involved in lipid synthesis. *Journal of Biological Chemistry* 280(6):4329-4338.
- Hoffmeister M, van der Klei A, Rotte C, van Grinsven KWA, van Hellemond JJ, Henze K, Tielens AGM, Martin W. 2004. *Euglena gracilis* Rhodoquinone : Ubiquinone ratio and mitochondrial proteome differ under aerobic and anaerobic conditions. *Journal of Biological Chemistry* 279(21):22422-22429.
- Horváth A, Berry EA, Maslov DA. 2000. Translation of the edited mRNA for cytochrome b in trypanosome mitochondria. *Science* 287(5458):1639-1640.
- 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 Physiology* 149(2):719-734.
- Inui H, Ishikawa T, Tamoi M. 2017. Wax Ester Fermentation and Its Application for Biofuel Production. *Euglena: Biochemistry, Cell and Molecular Biology* 979:269-283.
- Inwongwan S, Kruger NJ, Ratcliffe RG, O'Neill EC. 2019. *Euglena* Central Metabolic Pathways and Their Subcellular Locations. *Metabolites* 9(6):115.
- Jedelský PL, Doležal P, Rada P, Pyrih J, Smid O, Hrdý I, Sedinová M, Marčinciková M, Voleman L, Perry AJ et al. 2011. The Minimal Proteome in the Reduced Mitochondrion of the Parasitic Protist *Giardia intestinalis*. *Plos One* 6(2).
- Kanehisa M. 2017. Enzyme Annotation and Metabolic Reconstruction Using KEGG. *Protein Function Prediction: Methods and Protocols* 1611:135-145.
- Kanehisa M, Sato Y, Morishima K. 2016. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *Journal of Molecular Biology* 428(4):726-731.
- Karnkowska A, Vacek V, Zubáčová Z, Treitli SC, Petrželková R, Eme L, Novák L, Žárský V, Barlow LD, Herman EK et al. 2016. A eukaryote without a mitochondrial organelle. *Current Biology* 26(10):1274-1284.
- Kaur B, Záhonová K, Valach M, Faktorová D, Prokopchuk G, Burger G, Lukeš J. Forthcoming. Gene fragmentation and RNA editing without borders: eccentric mitochondrial genomes of diplomonads. *Nucleic Acids Research*.

- Kaurov I, Vancová M, Schimanski B, Cadena LR, Heller J, Bílý T, Potěšil D, Eichenberger C, Bruce H, Oeljeklaus S et al. 2018. The Diverged Trypanosome MICOS Complex as a Hub for Mitochondrial Cristae Shaping and Protein Import. *Current Biology* 28(21):3393-+.
- Kim M, Lee U, Small I, des Francs-Small CC, Vierling E. 2012. Mutations in an *Arabidopsis* Mitochondrial Transcription Termination Factor-Related Protein Enhance Thermotolerance in the Absence of the Major Molecular Chaperone HSP101. *Plant Cell* 24(8):3349-3365.
- Kleine T, Leister D. 2015. Emerging functions of mammalian and plant mTERFs. *Biochimica Et Biophysica Acta-Bioenergetics* 1847(9):786-797.
- Kosugi S, Hasebe M, Tomita M, Yanagawa H. 2009. Systematic identification of cell cycle-dependent yeast nucleocytoplasmic shuttling proteins by prediction of composite motifs. *Proceedings of the National Academy of Sciences of the United States of America* 106(25):10171-10176.
- Kozjak-Pavlovic V. 2017. The MICOS complex of human mitochondria. *Cell and Tissue Research* 367(1):83-93.
- Krajčovič J, Vesteg M, Schwartzbach SD. 2015. Euglenoid flagellates: A multifaceted biotechnology platform. *Journal of Biotechnology* 202:135-145.
- Kroph A, Rapacki K. TMHMM Server v. 2.00 prediction of transmembrane helices in proteins [Internet]. Available from: <http://www.cbs.dtu.dk/services/TMHMM-2.0/>
- Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Molecular Biology and Evolution* 35(6):1547-1549.
- Lane N, Martin W. 2010. The energetics of genome complexity. *Nature* 467(7318):929-934.
- Leander BS, Lax G, Karnkowska A, Simpson AGB. 2017. Euglenida. In: Archibald JM, Simpson AGB, Slamovits CH, editors. *Handbook of the Protists*. Cham: Springer International Publishing. p. 1047-1088.
- Lee CP, Taylor NL, Millar AH. 2013. Recent advances in the composition and heterogeneity of the *Arabidopsis* mitochondrial proteome. *Frontiers in Plant Science* 4:8.
- Léon S, Touraine B, Ribot C, Briat JF, Lobreaux S. 2003. Iron-sulphur cluster assembly in plants: distinct NFU proteins in mitochondria and plastids from *Arabidopsis thaliana*. *Biochemical Journal* 371:823-830.
- Li J, Cai TX, Wu P, Cui ZY, Chen XL, Hou JJ, Xie ZS, Xue P, Shi LA, Liu PS et al. 2009. Proteomic analysis of mitochondria from *Caenorhabditis elegans*. *Proteomics* 9(19):4539-4553.
- Lill R. 2009. Function and biogenesis of iron-sulphur proteins. *Nature* 460(7257):831-838.
- Lin JR, Hu JJ. 2013. SeqNLS: Nuclear Localization Signal Prediction Based on Frequent Pattern Mining and Linear Motif Scoring. *Plos One* 8(10).
- Lukeš J, Archibald JM, Keeling PJ, Doolittle WF, Gray MW. 2011. How a Neutral Evolutionary Ratchet Can Build Cellular Complexity. *Iubmb Life* 63(7):528-537.
- Lukeš J, Basu S. 2015. Fe/S protein biogenesis in trypanosomes - A review. *Biochimica et Biophysica Acta-Molecular Cell Research* 1853(6):1481-1492.
- Maio N, Singh A, Uhrigshardt H, Saxena N, Tong WH, Rouault TA. 2014. Cochaperone Binding to LYR Motifs Confers Specificity of Iron Sulfur Cluster Delivery. *Cell Metabolism* 19(3):445-457.
- Mani J, Meisinger C, Schneider A. 2016. Peeping at TOMs-Diverse Entry Gates to Mitochondria Provide Insights into the Evolution of Eukaryotes. *Molecular Biology and Evolution* 33(2):337-351.
- Martin J. 1997. Molecular chaperones and mitochondrial protein folding. *Journal of Bioenergetics and Biomembranes* 29(1):35-43.
- McShea DW, Hordijk W. 2013. Complexity by Subtraction. *Evolutionary Biology* 40(4):504-520.
- Mi-ichi F, Abu Yousuf M, Nakada-Tsukui K, Nozaki T. 2009. Mitosomes in *Entamoeba histolytica* contain a sulfate activation pathway. *Proceedings of the National Academy of Sciences of the United States of America* 106(51):21731-21736.
- Millar AH, Heazlewood JL, Kristensen BK, Braun HP, Moller IM. 2005. The plant mitochondrial proteome. *Trends in Plant Science* 10(1):36-43.

- Miranda-Astudillo HV, Yadav KNS, Colina-Tenorio L, Bouillenne F, Degand H, Morsomme P, Boekema EJ, Cardol P. 2018. The atypical subunit composition of respiratory complexes I and IV is associated with original extra structural domains in *Euglena gracilis*. *Scientific Reports* 8:9698.
- Mitchell AL, Attwood TK, Babbitt PC, Blum M, Bork P, Bridge A, Brown SD, Chang HY, El-Gebali S, Fraser MI et al. 2019. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Research* 47(D1):D351-D360.
- Mokranjac D, Neupert W. 2009. Thirty years of protein translocation into mitochondria: Unexpectedly complex and still puzzling. *Biochimica Et Biophysica Acta-Molecular Cell Research* 1793(1):33-41.
- Mueller SJ, Lang D, Hoernstein SNW, Lang EGE, Schuessele C, Schmidt A, Fluck M, Leisibach D, Niegler C, Zimmer AD et al. 2014. Quantitative Analysis of the Mitochondrial and Plastid Proteomes of the Moss *Physcomitrella patens* Reveals Protein Macrocompartmentation and Microcompartmentation. *Plant Physiology* 164(4):2081-2095.
- Müller M, Mentel M, van Hellemond JJ, Henze K, Woehle C, Gould SB, Yu RY, van der Giezen M, Tielens AGM, Martin WF. 2012. Biochemistry and Evolution of Anaerobic Energy Metabolism in Eukaryotes. *Microbiology and Molecular Biology Reviews* 76(2):444-495.
- Nakazawa M, Nishimura M, Inoue K, Ueda M, Inui H, Nakano Y, Miyatake K. 2011. Characterization of a Bifunctional Glyoxylate Cycle Enzyme, Malate Synthase/Isocitrate Lyase, of *Euglena gracilis*. *Journal of Eukaryotic Microbiology* 58(2):128-133.
- Nogales J, Gudmundsson S, Knight EM, Palsson BO, Thiele I. 2012. Detailing the optimality of photosynthesis in cyanobacteria through systems biology analysis. *Proceedings of the National Academy of Sciences of the United States of America* 109(7):2678-2683.
- Novák Vanclová AMG, Zoltner M, Kelly S, Soukal P, Záhonová K, Füßy Z, Ebenezer TE, Lacová Dobáková E, Eliáš M, Lukeš J et al. 2019. Metabolic quirks and the colourful history of the *Euglena gracilis* secondary plastid. *New Phytol.*
- O'Neill EC, Trick M, Hill L, Rejzek M, Dusi RG, Hamilton CJ, Zimba PV, Henrissat B, Field RA. 2015. The transcriptome of *Euglena gracilis* reveals unexpected metabolic capabilities for carbohydrate and natural product biochemistry. *Molecular Biosystems* 11(10):2808-2820.
- Ollagnier-de-Choudens S, Mattioli T, Tagahashi Y, Fontecave M. 2001. Iron-sulfur cluster assembly - Characterization of IscA and evidence for a specific and functional complex with ferredoxin. *Journal of Biological Chemistry* 276(25):22604-22607.
- Panigrahi AK, Ogata Y, Ziková A, Anupama A, Dalley RA, Acestor N, Myler PJ, Stuart KD. 2009. A comprehensive analysis of *Trypanosoma brucei* mitochondrial proteome. *Proteomics* 9(2):434-450.
- Patron NJ, Durnford DG, Kopriva S. 2008. Sulfate assimilation in eukaryotes: fusions, relocations and lateral transfers. *BMC Evolutionary Biology* 8:14.
- Peikert CD, Mani J, Morgenstern M, Käser S, Knapp B, Wenger C, Harsman A, Oeljeklaus S, Schneider A, Warscheid B. 2017. Charting organellar importomes by quantitative mass spectrometry. *Nature Communications* 8.
- Peña-Díaz P, Lukeš J. 2018. Fe-S cluster assembly in the supergroup Excavata. *Journal of Biological Inorganic Chemistry* 23(4):521-541.
- Perez E, Lapaille M, Degand H, Cilibrasi L, Villavicencio-Queijeiro A, Morsomme P, González-Halphen D, Field MC, Remacle C, Baurain D et al. 2014. The mitochondrial respiratory chain of the secondary green alga *Euglena gracilis* shares many additional subunits with parasitic *Trypanosomatidae*. *Mitochondrion* 19:338-349.
- Petrov AS, Wood EC, Bernier CR, Norris AM, Brown A, Amunts A. 2019. Structural Patching Fosters Divergence of Mitochondrial Ribosomes. *Molecular Biology and Evolution* 36(2):207-219.
- Quesada V. 2016. The roles of mitochondrial transcription termination factors (MTERFs) in plants. *Physiologia Plantarum* 157(3):389-399.

- Ramrath DJF, Niemann M, Leibundgut M, Bieri P, Prange C, Horn EK, Leitner A, Boehringer D, Schneider A, Ban N. 2018. Evolutionary shift toward protein-based architecture in trypanosomal mitochondrial ribosomes. *Science* 362(6413):7735.
- Read LK, Lukeš J, Hashimi H. 2016. Trypanosome RNA editing: the complexity of getting U in and taking U out. *Wiley Interdisciplinary Reviews-Rna* 7(1):33-51.
- Roberti M, Polosa PL, Bruni F, Manzari C, Deceglie S, Gadaleta MN, Cantatore P. 2009. The MTERF family proteins: Mitochondrial transcription regulators and beyond. *Biochimica Et Biophysica Acta-Bioenergetics* 1787(5):303-311.
- Robles P, Navarro-Cartagena S, Ferrández-Ayela A, Núñez-Delegido E, Quesada V. 2018. The Characterization of *Arabidopsis* mterf6 Mutants Reveals a New Role for mTERF6 in Tolerance to Abiotic Stress. *International Journal of Molecular Sciences* 19(8):2388.
- Roger AJ, Muñoz-Gómez SA, Kamikawa R. 2017. The Origin and Diversification of Mitochondria. *Current Biology* 27(21):R1177-R1192.
- Rotte C, Stejskal F, Zhu G, Keithly JS, Martin W. 2001. Pyruvate : NADP(+) oxidoreductase from the mitochondrion of *Euglena gracilis* and from the apicomplexan *Cryptosporidium parvum*: A biochemical relic linking pyruvate metabolism in mitochondriate and amitochondriate protists. *Molecular Biology and Evolution* 18(5):710-720.
- Saidha T, Na SQ, Li JY, Schiff JA. 1988. A sulfate metabolizing center in *Euglena* mitochondria. *Biochemical Journal* 253(2):533-539.
- Salvato F, Havelund JF, Chen MJ, Rao RSP, Rogowska-Wrzesinska A, Jensen ON, Gang DR, Thelen JJ, Moller IM. 2014. The Potato Tuber Mitochondrial Proteome. *Plant Physiology* 164(2):637-653.
- Santos HJ, Makiuchi T, Nozaki T. 2018. Reinventing an Organelle: The Reduced Mitochondrion in Parasitic Protists. *Trends in Parasitology* 34(12):1038-1055.
- Schneider AP. 2018. Mitochondrial protein import in trypanosomatids: Variations on a theme or fundamentally different? *Plos Pathogens* 14(11).
- Schneider RE, Brown MT, Shiflett AM, Dyall SD, Hayes RD, Xie YM, Loo JA, Johnson PJ. 2011. The *Trichomonas vaginalis* hydrogenosome proteome is highly reduced relative to mitochondria, yet complex compared with mitosomes. *International Journal for Parasitology* 41(13-14):1421-1434.
- Schön A, Kannangara CG, Gough S, Söll D. 1988. Protein biosynthesis in organelles requires misaminoacylation of tRNA. *Nature* 331(6152):187-90.
- Schönfeld C, Wobbe L, Borgstadt R, Kienast A, Nixon PJ, Kruse O. 2004. The nucleus-encoded protein MOC1 is essential for mitochondrial light acclimation in *Chlamydomonas reinhardtii*. *Journal of Biological Chemistry* 279(48):50366-50374.
- Sharma S, Singha UK, Chaudhuri M. 2010. Role of Tob55 on mitochondrial protein biogenesis in *Trypanosoma brucei*. *Molecular and Biochemical Parasitology* 174(2):89-100.
- Sheftel AD, Wilbrecht C, Stehling O, Niggemeyer B, Elsässer HP, Mühlhoff U, Lill R. 2012. The human mitochondrial ISCA1, ISCA2, and IBA57 proteins are required for 4Fe-4S protein maturation. *Molecular Biology of the Cell* 23(7):1157-1166.
- Sickmann A, Reinders J, Wagner Y, Joppich C, Zahedi R, Meyer HE, Schönfisch B, Perschil I, Chacinska A, Guiard B et al. 2003. The proteome of *Saccharomyces cerevisiae* mitochondria. *Proceedings of the National Academy of Sciences of the United States of America* 100(23):13207-13212.
- Škodová-Sveráková I, Horváth A, Maslov DA. 2015. Identification of the mitochondrially encoded subunit 6 of F-1 F-0 ATPase in *Trypanosoma brucei*. *Molecular and Biochemical Parasitology* 201(2):135-138.
- Smith DGS, Gawryluk RMR, Spencer DF, Pearlman RE, Siu KWM, Gray MW. 2007. Exploring the mitochondrial proteome of the ciliate protozoon *Tetrahymena thermophila*: Direct analysis by tandem mass spectrometry. *Journal of Molecular Biology* 374(3):837-863.

- Stoltzfus A. 1999. On the possibility of constructive neutral evolution. *Journal of Molecular Evolution* 49(2):169-181.
- Taylor SW, Fahy E, Zhang B, Glenn GM, Warnock DE, Wiley S, Murphy AN, Gaucher SP, Capaldi RA, Gibson BW et al. 2003. Characterization of the human heart mitochondrial proteome. *Nature Biotechnology* 21(3):281-286.
- Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, Mann M, Cox J. 2016. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature Methods* 13(9):731-740.
- Valach M, Moreira S, Hoffmann S, Stadler PF, Burger G. 2017. Keeping it complicated: Mitochondrial genome plasticity across diplomonads. *Scientific Reports* 7.
- Verner Z, Basu S, Benz C, Dixit S, Dobáková E, Faktorová D, Hashimi H, Horáková E, Huang ZQ, Paris Z et al. 2015. Malleable Mitochondrion of *Trypanosoma brucei*. *International Review of Cell and Molecular Biology* 315:73-151.
- Vesteg M, Hadariová L, Horváth A, Estraño CE, Schwartzbach SD, Krajčovič J. 2019. Comparative molecular cell biology of phototrophic euglenids and parasitic trypanosomatids sheds light on the ancestor of Euglenozoa. *Biological Reviews*.
- Vizcaíno JA, Csordas A, del-Toro N, Dianas JA, Griss J, Lavidas I, Mayer G, Perez-Riverol Y, Reisinger F, Ternent T et al. 2016. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Research* 44(D1):D447-D456.
- Werhahn W, Braun HP. 2002. Biochemical dissection of the mitochondrial proteome from *Arabidopsis thaliana* by three-dimensional gel electrophoresis. *Electrophoresis* 23(4):640-646.
- Wiedemann N, Frazier AE, Pfanner N. 2004. The protein import machinery of mitochondria. *Journal of Biological Chemistry* 279(15):14473-14476.
- Wredenber A, Lagouge M, Bratic A, Metodiev MD, Spähr H, Mourier A, Freyer C, Ruzzenente B, Tain L, Grönke S et al. 2013. MTERF3 Regulates Mitochondrial Ribosome Biogenesis in Invertebrates and Mammals. *PLOS Genetics* 9(1):16.
- Yoshida Y, Tomiyama T, Maruta T, Tomita M, Ishikawa T, Arakawa K. 2016. De novo assembly and comparative transcriptome analysis of *Euglena gracilis* in response to anaerobic conditions. *BMC Genomics* 17:182.
- Zakryš B, Milanowski R, Karnkowska A. 2017. Evolutionary Origin of *Euglena*. In: Schwartzbach SD, Shigeoka S, editors. *Euglena: Biochemistry, Cell and Molecular Biology*. Cham: Springer. p. 3-17.
- Zhang SY, Qian X, Chang SN, Dismukes GC, Bryant DA. 2016. Natural and Synthetic Variants of the Tricarboxylic Acid Cycle in Cyanobacteria: Introduction of the GABA Shunt into *Synechococcus* sp PCC 7002. *Frontiers in Microbiology* 7.
- Zimorski V, Rauch C, van Hellemond JJ, Tielens AGM, Martin WF. 2017. The Mitochondrion of *Euglena gracilis*. In: Schwartzbach SD, Shigeoka S, editors. *Euglena: Biochemistry, Cell and Molecular Biology*. Cham: Springer. p. 19-37.
- Záhonová K, Füssy Z, Birčák E, Novák Vanclova AMG, Klimeš V, Vesteg M, Krajčovič J, Oborník M, Eliáš M. 2018. Peculiar features of the plastids of the colourless alga *Euglena longa* and photosynthetic euglenophytes unveiled by transcriptome analyses. *Scientific Reports* 8:17012.
- Zíková A, Panigrahi AK, Dalley RA, Acestor N, Anupama A, Ogata Y, Myler PJ, Stuart K. 2008. *Trypanosoma brucei* mitochondrial ribosomes. *Molecular & Cellular Proteomics* 7(7):1286-1296.

Figure legends

Figure 1. Diagram of identifying strategies used for majority ID transcripts of *E. gracilis* mitoproteome. List of proteins grouped by identification strategy are available in Suppl. Table 3.

Figure 2. Functional categories of 743 mitochondrially enriched transcripts with predicted function (42% of the entire proteome) including the logarithmic enrichment ratio of the mitochondrial versus chloroplast preparation ($\log_{10}MT/CP$) ratio for each category, shown in shades of orange (infinite, 0-1 representing 1-10x greater protein amount in mitochondrial fraction, 1-2 for 10-100x; full category names are listed in Materials and Methods).

Figure 3. Mitochondrial protein import apparatus of *E. gracilis*. Blue for proteomically confirmed sequences, light blue for sequences identified in transcriptome, grey for absent sequences and purple for ‘putative’ sequences with weak homology that nonetheless show presence within the mitoproteome. Sequences taken from translocation apparatus of *T. brucei*, or *S. cerevisiae* in the case of opisthokont TIM22 complex and transmembrane proteins, Mba 1, Mia40, IMP1.

Figure 4. Core metabolic pathways and their proteomic presence within *E. gracilis*. (A) TCA cycle showing the conventional pathway rendered non-functional by absence of key protein subunits along with alpha-proteobacterial shunt and aminobutyric (GABA) shunt. Components involved only in the glyoxylate cycle are transparent. (B) Glyoxylate cycle, highlighting associated enzymes from TCA cycle. Blue for proteomically confirmed sequences, light blue for sequences identified in transcriptome, grey for absent sequences. Duplicated sequences are available in Suppl. Table 6.

Figure 5. Iron-sulphur cluster biosynthesis in *E. gracilis* mitochondrial matrix and intermembrane space (IMS). Blue for proteomically confirmed sequences, light blue for sequences identified in transcriptome, white for proteins requiring iron-sulphur cluster insertion for functionality.

Figure 6. Evolutionary schematic showing unique mitochondrial features to euglenida, with features observed only in *Euglena gracilis* denoted by *. Common features discovered in this study with the *Trypanosoma brucei* mitoproteome are listed, which were likely present in the mitochondria of the last euglenozoan common ancestor. Over 30 RNA editing enzymes

appear to be present in this ancestor, which led to development of different RNA editing pathways in diplonemid and kinetoplastid lineages, denoted by †.

Figures:

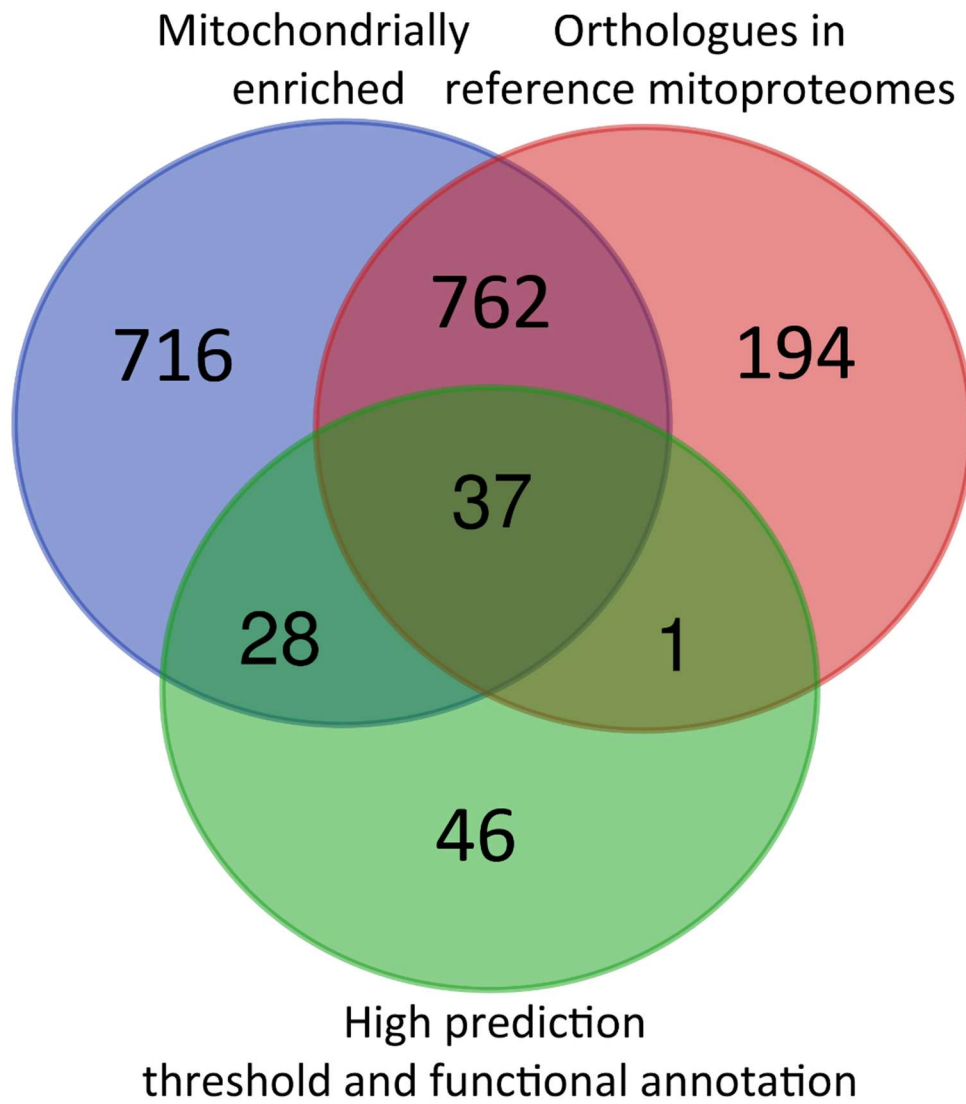


Figure 1

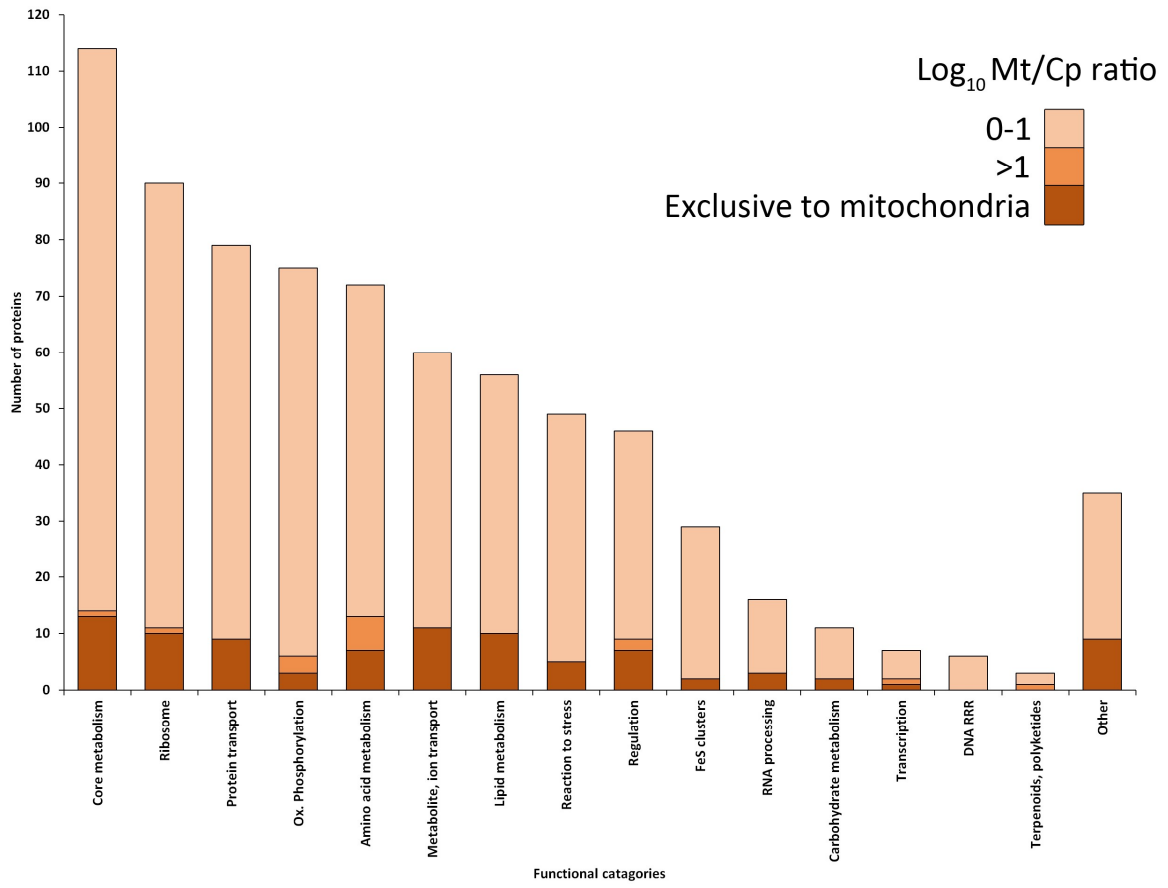


Figure 2

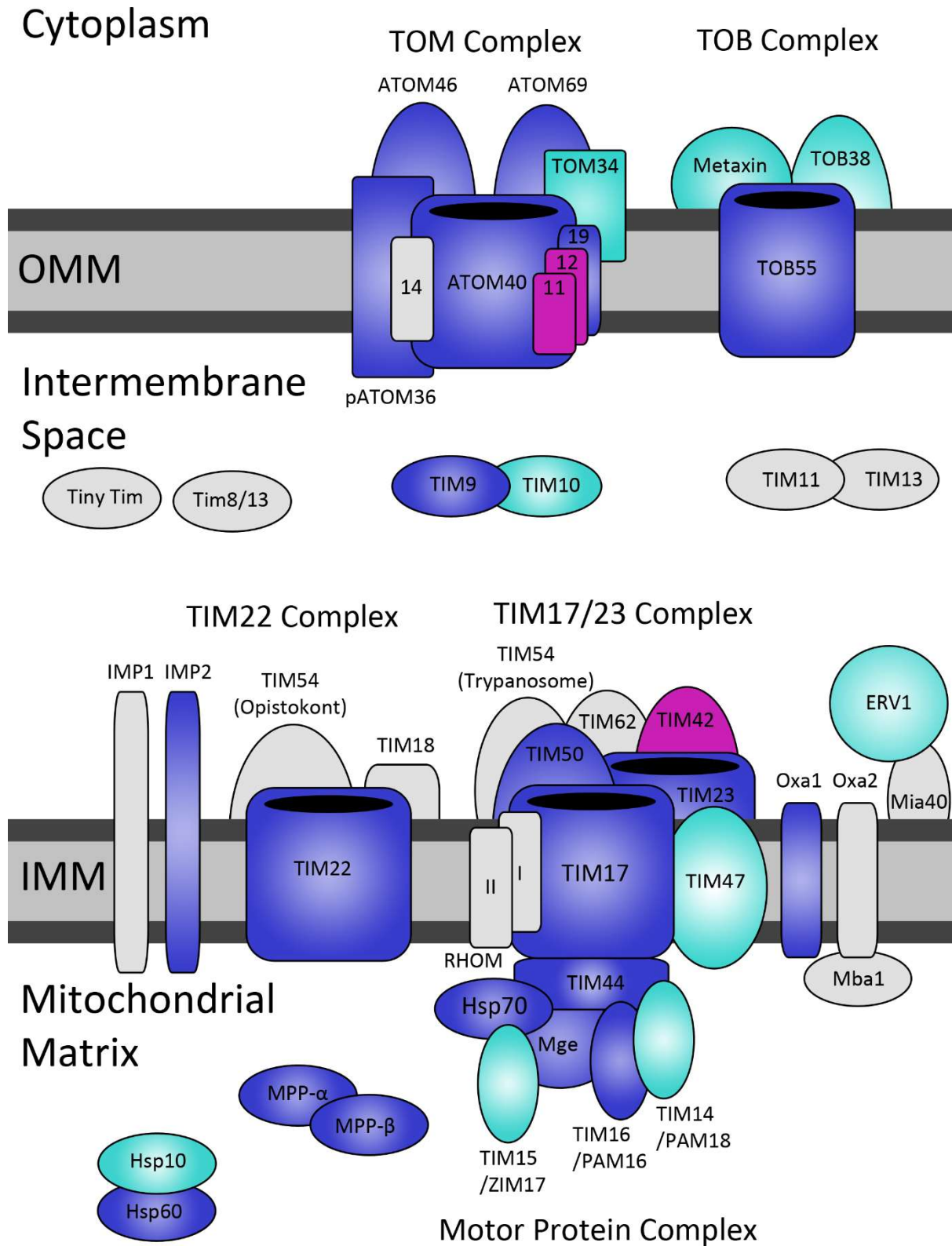


Figure 3

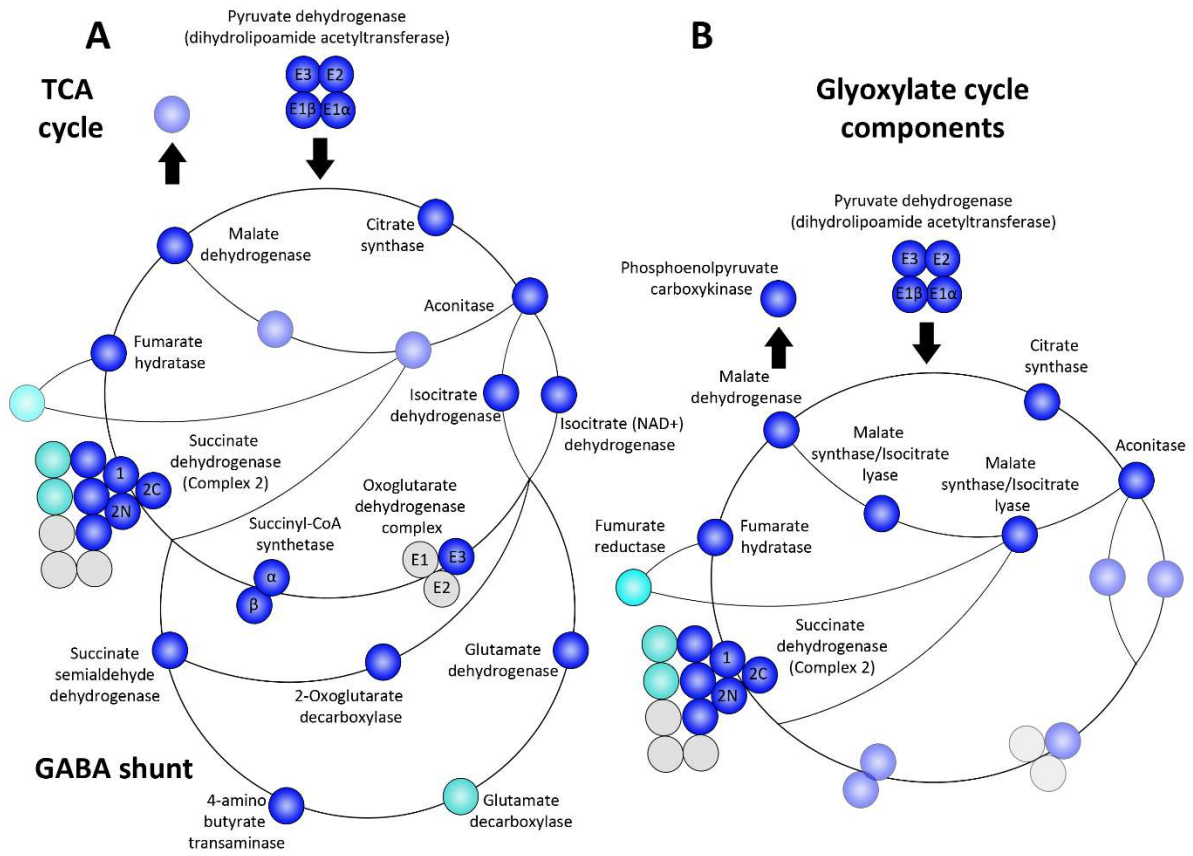


Figure 4

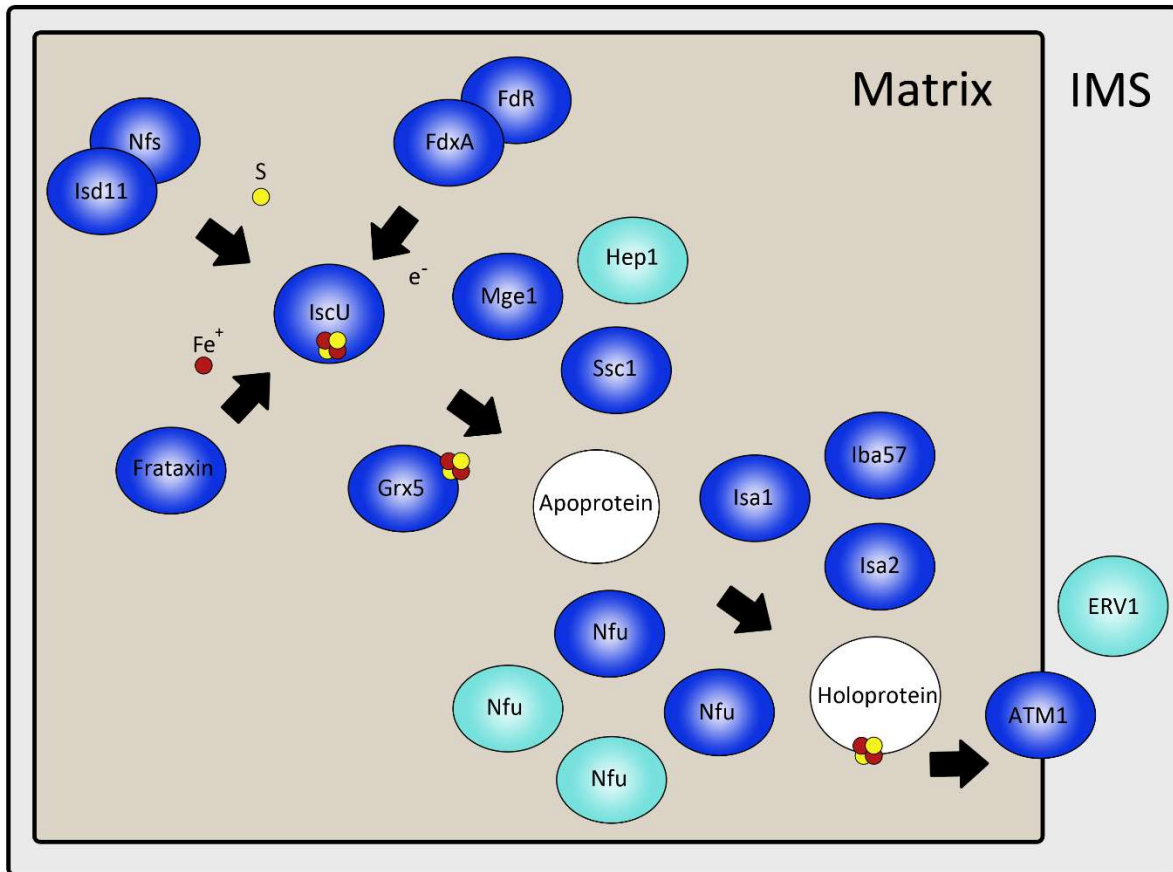


Figure 5

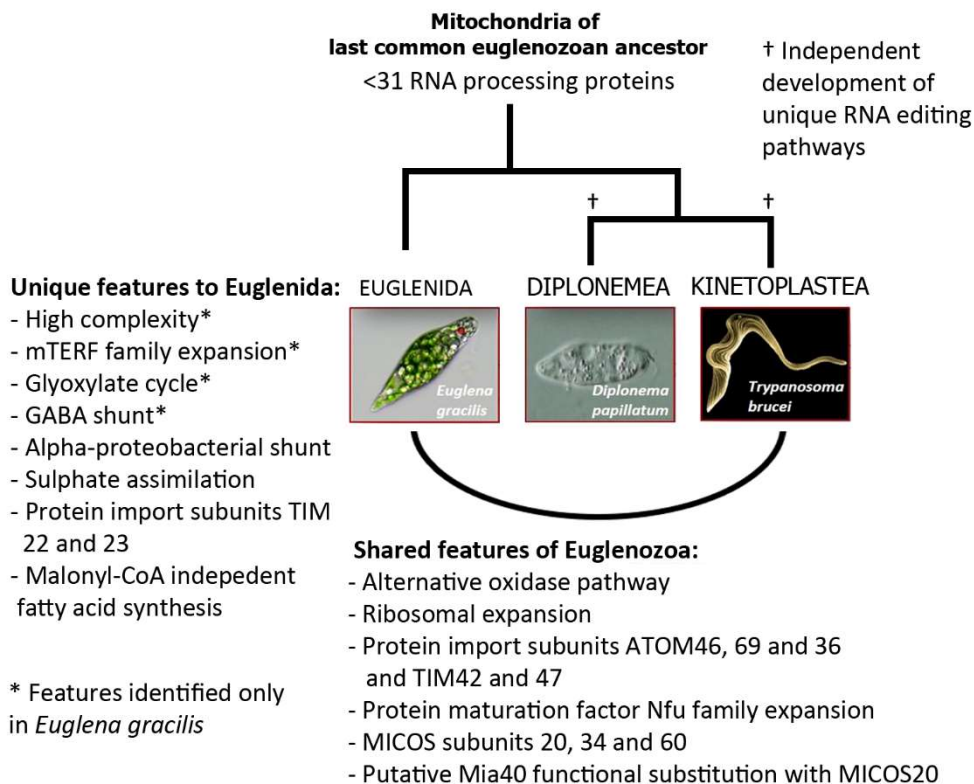


Figure 6

Supplementary figure legends

Supplementary Figure 1. Isolation of *E. gracilis* organellar fractions in sucrose gradient. *E. gracilis* whole cell lysate was loaded onto a discontinuous sucrose discontinuous density gradient. After centrifugation, the mitochondrial fraction used for mitoproteome identification was located at the interphase of 1.7 M and 1.5 M sucrose layer. Other organellar fractions separated by ultracentrifugation were chloroplasts (interphase of 1.5 M and 1.25 M sucrose) and peroxisomes (interphase of 1.25 M and 1 M sucrose).

Supplementary Figure 2. Chloroplast enrichment of transcripts in the experimentally verified mitoproteome, displaying 211 chloroplast-enriched sequences, log₁₀MT/CP ratios shown in shades of green (0-1 representing 1-10x greater protein level in chloroplast fraction, 1-2 for 10-100x, 2-3 for 100x-1000x, >3 for greater than 1000x).

Supplementary Figure 3. Volcano plots from *p*-values versus the corresponding t-test difference of 8,216 protein groups quantified in the two organellar fractions and whole cell lysate. Green and blue dots represent proteins assigned to “photosynthetic” and “mitochondrial” GO categories, respectively. The remaining colours represent other selected GO categories (indicated at the top right) associated with other cellular compartments.

Supplementary Figure 4. Functionally annotated sequences of mitoproteome transcripts, sorted into the following groups: “core metabolic pathways”, “oxidative phosphorylation and electron transport”, “carbohydrate metabolism”, “lipid metabolism”, “amino acid metabolism”, “metabolism of cofactors and vitamins”, “metabolism of terpenoids and polyketides”, “DNA replication, recombination and repair”, “transcription and transcription regulation”, “RNA processing and degradation”, “ribosome, aminoacyl-tRNA biosynthesis and translation”, “protein transport, folding, processing and degradation”, “Fe-S cluster assembly and sulphur metabolism”, “metabolite and ion transport”, “regulation and signal transduction”, “reaction to oxidative and toxic stress”, and “other”.

Supplementary Figure 5. Summary of MICOS sequences in mitochondrial proteome showing domain architecture. Key to colour-coding of motifs on right. Probability scores from CC predictions are indicated by *.

Supplementary Figure 6. Map of amino acid biosynthesis in *E. gracilis* mitochondrion.

Supplementary File 1. Extended information on sulphate assimilation and fatty acid synthesis, as well as detail on ubiquinone synthesis pathway.

Supplementary Table 1. 2,704 mitochondrial proteins filtered by false discovery rate of 0.01.

Supplementary Table 2. Mass spectrometry data for transcript sequences experimentally verified as mitochondrial, and peptide evidence for mitochondrially-encoded proteins.

Supplementary Table 3. Table of reference mitochondrial proteomes used in this study, including *Trypanosoma brucei*, *Arabidopsis thaliana*, *Mus musculus* and *Saccharomyces cerevisiae*.

Supplementary Table 4. Verified transcripts constituting *Euglena gracilis* mitoproteome, displaying orthologous group, mitochondria to chloroplast ratio (Mt/Cp), predicted function, functional category, KEGG and Blast2Go annotation and orthologues present in reference mitochondrial proteomes.

Supplementary Table 5. Summary table of number of experimentally verified sequences from *E. gracilis* mitoproteome, as well as additional *in silico* predicted sequences.

Supplementary Table 6. Assessment of the purity of isolated fractions using selected marker proteins and their relative abundance compared against the whole cell lysate and between the organellar fractions (Mt=mitochondrial fraction, CP=chloroplast fraction, W=whole cell lysate).

Supplementary Table 7. Functional characterisation of mitochondrial sequences as defined by transcriptome from Field's lab (Ebenezer et al. 2019), identified *in silico* and found in mitochondrial proteome.

Supplementary Table 8. Aminoacyl-tRNA synthetases (aaRS) and related enzymes identified in the transcriptome of *E. gracilis* and their putative localization (M-mitochondria; P-plastid, C-cytoplasm). AaRSs for which mitochondrion-targeted paralogue (or isoform) could be detected are marked in green.

Supplementary Table 9. Orthologues of *Eutreptiella gymnastica* transcriptome against verified mitoproteome of *Euglena gracilis* with functional classification of specific pathways and Hidden Markov Model search results.