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

Investigation of the Subunit Composition of Mitochondrial Dehydrogenase Complexes, Putative Kinetochores, and Localization of Paraflagellar Rod Proteins in Marine Diplonemids

Ph.D. Thesis

Pragya Tripathi

Supervisor: Prof. RNDr. Julius Lukeš, CSc.

Institute of Parasitology,

Czech Academy of Sciences and Biology Centre,

České Budějovice

2023

This thesis should be cited as:
Tripathi P., 2023: Investigation of the subunit composition of mitochondrial dehydrogenase complexes, putative kinetochores, and localization of paraflagellar rod proteins in marine diplonemids. Ph.D. Thesis. University of South Bohemia, Faculty of Science, School of Doctoral Studies in Biological Sciences, České Budějovice, Czech Republic, 243 pp.

Annotation

In this study I have studied various aspects of the biology of diplonemids, a group of marine planktonic protists that were previously understudied due to technical constraints but have now been recognized as one of the most diverse in terms of species richness. I aimed to provide insight in the metabolism of *Paradiplonema papillatum* by investigating the composition of its pyruvate dehydrogenase (PDH) complex. In association with the recent annotation of the complete nuclear genome of *P. papillatum*, I decided to examine the composition of its PDH complex. Furthermore, I have started an exploration of the kinetochore and kinetochore-related proteins, which seem to be extremely derived in diplonemids. Finally, I have investigated the re-location of conserved paraflagellar rod proteins in these flagellates, which can construct flagella with or without morphologically distinguishable paraflagellar rods, contingent upon their life cycle stage.

Declaration

I hereby declare that I am the author of this dissertation and that I have used only those sources and literature detailed in the list of references.

In České Budějovice May 23, 2023 Pragya Tripathi

List of papers and author's contribution

1. Kristína Záhonová[1,2,3,4,#], Matus Valach[5,#], **Pragya Tripathi**[1,6,#], Corinna Benz[1], Fred R. Opperdoes[7], Peter Barath[8,9], Veronika Lukáčová[9], Maksym Danchenko[8], Drahomíra Faktorová[1,6], Anton Horváth[10], Gertraud Burger[5], Julius Lukeš[1,6,*] and

Ingrid Škodová-Sveráková[1,2,10,*] **;** Subunit composition of mitochondrial dehydrogenase complexes in widespread marine diplonemids. *Biochimica et Biophysica Acta - General Subjects (re-submitted)*

P. T. participated in cell cultivation, transfection, production of transgenic cell lines, and investigation, and contributed to the writing of the manuscript. Her contribution was 50 %.

2. Matus Valach[1]*, Sandrine Moreira[1], Celine Petitjean[2], Corinna Benz[3], Anzhelika Butenko[3,4,5], Olga Flegontova[3,5], Anna Nenarokova[2,3], Galina Prokopchuk[3,4], Tom Batstone[2,6], Pascal Lapébie[7], Lionnel Lemogo[1,8], Matt Sarrasin[1], Paul Stretenowich[1,9], **Pragya Tripathi[3,4]**, Euki Yazaki[10], Takeshi Nara[11], Bernard Henrissat[7,12], B. Franz Lang[1], Michael W. Gray[13], Tom A. Williams[2], Julius Lukeš[3,4] and Gertraud Burger[1]*; Recent expansion of metabolic versatility in *Diplonema papillatum*, the model species of a highly speciose group of marine eukaryotes.*BMC Biology*. (IF-7.364) https://doi.org/10.1186/s12915-023-01563-9

P. T. participated in cell cultivation, performing experiment using an expansion microscopy. Her contribution was 15 %.

3. **Pragya Tripathi[1,2]** Michael Hammond[1,2,] and Julius Lukeš[1,2], An investigation into elusive paraflagellar rod proteins in *Paradiplonema papillatum*. (Manuscript in preparation)

P. T. participated in cell cultivation, transfection, production of transgenic cell lines, and investigation, and contributed to the writing of the manuscript. Her contribution was 85 %.

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 program.

Financial support

This work was supported by Grant agency of the University of South Bohemia in České Budějovice (grant 050/2016/P and 094/2018/P to Pragya Tripathi), the Czech Grant Agency (15–21974S and 16–18699S to Julius Lukeš), the ERC CZ grant (LL1601 to J.L.), the Czech Ministry of Education (ERD Funds OPVVV16_019/ 0000759 to J.L.).

Acknowledgments

For the completion of my Ph.D. journey, I am grateful to my supervisor, **Julius Lukeš** for his invaluable role in shaping me into a scientific scholar. Observing Julius's enthusiasm for science, even after spending so many years in the same field, is an inspiration to me and helps me to better grasp scientific principles in my life. Despite his numerous achievements, he remained deeply involved and committed to pursuing his passion for scientific exploration. One particular instance that I remember, and which deeply inspired me, occurred after Julius published a significant breakthrough. Instead of celebration, he immediately posed the question to the lab, **"What is next?"**. After spending five and a half years under Julius's guidance and closely observing his activities, I have come to believe that in the life of science, external factors hold little significance compared to the depth of knowledge. Handling a substantial laboratory and shouldering numerous responsibilities is no easy task. However, he created an environment that was both welcoming and accessible to everyone, encouraging open discussions about science while also occasionally social gatherings. I am grateful for the opportunity to pursue my Ph.D. studies under Julius's guidance.

I am grateful to you, **Corinna Benz**, for your support, guidance, and encouragement during the challenging times of my Ph.D. journey. Your presence, mentorship, and timely assistance have been invaluable in helping me understand science, overcome challenges, and navigate through the complexities of my research. I am thankful for all your suggestions while I was writing my thesis.

I would like to express my gratitude to my colleague, **Michael John Hammond** for his kind assistance and valuable advice during the period of my research. In addition to helping me generate and analyze the diplonema data, his knowledge and ease of communication have been helpful. In addition, I appreciate

the insightful conversations and helpful suggestions you offered me while I was writing my thesis.

I was overwhelmed to share my office with **Faktorová Drahomíra** during my Ph.D. journey. You were there for me whenever I needed any suggestions, and I am truly thankful to you. Additionally, I am grateful for **Michaela Svobodová**, a very important member in our lab who takes care of many things and is hardworking.

I consider myself fortunate to have had not just colleagues and friends like you guys' **Durante Ignacio, Galina Prokopchuk, Butenko Anzhelika, and Nina shaghayegh sheikh.** I will miss your music, **Durante Ignacio.** I wish good luck to you all.

I would like to express my sincere appreciation to all the people from Jula's lab, both former and present members, for their tremendous help and support. I would also like to express my gratitude to **Alena Panicucci Zíková**'s lab members and **Zdeněk Paris**'s lab members.

Additionally, I thank all my co-authors especially, **Ingrid Sveráková** and **Matus Valach,** whose invaluable support played a significant role in shaping my Ph.D. Their enthusiasm and dedication in selecting delightful science topics have truly enriched our work. I am grateful for their insightful perspectives and collaborative efforts, which have made this project both engaging and meaningful.

I want to thank my husband, **Rahul Kumar Mishra** for all his support, inspiration, and encouragement throughout my Ph.D. journey. He has assisted me in finding a balance between work and home life, and his advice has been essential to my success. I want to thank my 4-year-old son, **Kaulyam Kumar Mishra**, for making the journey easier and more meaningful with his love,

laughter, and innocent curiosity. I am grateful for his understanding of my responsibility and sacrifices, and I am proud to have him as my biggest support.

I want to express my sincere gratitude to my friend, a nature enthusiast, **Ambar Kachale**. I have never encountered someone as passionate and observant as you when it comes to appreciating the smallest details of our surroundings. Your curiosity for science is equally remarkable, and it has been a joy to witness your inquisitive nature in action. It's wonderful to hear that the name PT, which you gave me, carries personal connections, and brings me even more joy. I wish you good luck in your future.

I extend my heartfelt gratitude to my dear friend, **Ayush Sharma**, who has been a constant source of encouragement, support, and inspiration throughout my studies. Thanks to you, I have gained immense confidence throughout my journey in science and in my personal life. Your presence and encouragement have had a significant impact on my understanding of life. My best wishes to you and my dear Lu for a better future.

I would also like to thank **Komal Varsadiya** and **Milan Varsadiya.** Their presence has made the journey much more enjoyable, and I could not have made it this far without them. Lastly, I am grateful for all my friends from České Budějovice.

I would like to express my deepest gratitude to my family My mother, my father, and my In-laws, and especially my sister **Shraddha Tripathi**, for their love, support, and encouragement throughout my Ph.D. journey. Their constant belief in me, their willingness to listen and offer advice, and their sacrifices have been essential to my success.

In the end, I want to dedicate this thesis to my husband, **Rahul Kumar Mishra**, for his belief in me and his encouragement. His presence has made this journey much more meaningful, and I am truly grateful for his love and support.

List of papers and author's contribution

1. Kristína Záhonová[1,2,3,4,#], Matus Valach[5,#], **Pragya Tripathi[1,6,#]**, Corinna Benz[1], Fred R. Opperdoes[7], Peter Barath[8,9], Veronika Lukáčová[9], Maksym Danchenko[8], Drahomíra Faktorová[1,6], Anton Horváth[10], Gertraud Burger[5], Julius Lukeš[1,6,*] and Ingrid Škodová-Sveráková[1,2,10,*] **;** Subunit composition of mitochondrial dehydrogenase complexes in widespread marine diplonemids. *Biochimica et Biophysica Acta - General Subjects (re-submitted)*

P. T. participated in cell cultivation, transfection, production of transgenic cell lines, and investigation, and contributed to the writing of the manuscript. Her contribution was 50 %.

2. Matus Valach[1]*, Sandrine Moreira[1], Celine Petitjean[2], Corinna Benz[3], Anzhelika Butenko[3,4,5], Olga Flegontova[3,5], Anna Nenarokova[2,3], Galina Prokopchuk[3,4], Tom Batstone[2,6], Pascal Lapébie[7], Lionnel Lemogo[1,8], Matt Sarrasin[1], Paul Stretenowich[1,9], **Pragya Tripathi[3,4]**, Euki Yazaki[10], Takeshi Nara[11], Bernard Henrissat[7,12], B. Franz Lang[1], Michael W. Gray[13], Tom A. Williams[2], Julius Lukeš[3,4] and Gertraud Burger[1]*; Recent expansion of metabolic versatility in *Diplonema papillatum*, the model species of a highly speciose group of marine eukaryotes.*BMC Biology*. (IF-7.364) https://doi.org/10.1186/s12915-023-01563-9

P. T. participated in cell cultivation, performing experiment using an expansion microscopy. Her contribution was 15 %.

3. **Pragya Tripathi[1,2]** Michael Hammond[1,2,] and Julius Lukeš[1,2], An investigation into elusive paraflagellar rod proteins in *Paradiplonema papillatum*. (Manuscript in preparation)

P. T. participated in cell cultivation, transfection, production of transgenic cell lines, and investigation, and contributed to the writing of the manuscript. Her contribution was 85 %.

Co-author agreement

Julius Lukeš, the supervisor of this Ph.D. thesis and co-author of all presented papers, fully acknowledges the contribution of Pragya Tripathi.

Prof. RNDr. Julius Lukeš

Table of contents

Abbreviations

ATCC	American type culture collection
BCKDH	branched chain ketoacid dehydrogenase
CRISPER	clustered regularly interspaced short palindromic repeats
DSPD	deep-sea pelagic diplonemids
E1	pyruvate dehydrogenase
E2	dihydrolipoyl acetyltransferase
E3	dihydrolipoyl dehydrogenase
OTS	operational taxonomy Units
OXPHOS	oxidative phosphorylation
OXDH	2-oxoglutarate dehydrogenase
PDH	pyruvate dehydrogenase complex
PFR	paraflagellar rod protein
SSDH	succinate-semialdehyde dehydrogenase
TCA	tricarboxylic acid cycle
NADH	nicotinamide adenine dinucleotide hydrogen
18S rRNA	18S ribosomal RNA

Summary of chapters

Diplonemids were once thought to be a small and uncommon group of flagellates, with only two genera and fewer than a dozen species identified (Massana 2011; Simpson 1997; Vickerman 2000; von der Heyden et al. 2004). The recent comprehensive analysis of the eukaryotic planktonic diversity in samples obtained by the Tara Oceans expedition, with the dataset made up of almost 800 million V9 18S rRNA barcodes, has, however, generated a great deal of interest in this group. Surprisingly, the research revealed that diplonemids are the sixth most abundant eukaryotic group inhabiting oceans worldwide and exhibit the highest diversity among marine eukaryotes, surpassing metazoans, and dinoflagellates de (Vargas *et al*. 2015; Lukeš *et al*. 2015; Flegontova *et al*. 2016). Moreover, all the identified species are from clades that are infrequently encountered in the marine samples that are currently available, and there is only limited sequence data and no information on the morphology of the members of the Eupelagonemidae, a diplonemid group (Gawryluk *et al*. 2016). In fact, this significant clade is still largely unfamiliar despite comprising up to 97% of all oceanic members within this group (Flegontova *et al*. 2016; Okamoto *et al*. 2019). The vast majority of the morphological, genomic, and physiological information currently available on diplonemids has been gathered from a single species, *Paradiplonema papillatum (*Kiethega *et al* 2011; Maslov *et al.* 1999; Valach *et al.*2016). This is largely attributed to its ability to endure cryopreservation, cultivate in a cost-effective medium at high densities, and the lack of availability of other species in culture collections. Furthermore, due to its genomic and transcriptomic data, *P. papillatum* is a suitable model for genetic manipulation and subsequent functional studies (Faktorová *et al.*, 2020; Valach, Moreira, Petitjean, Benz, Butenko, *et al.*, 2023).

In the first chapter of this thesis, we are describing the first nuclear genome sequence of the type species *P. papillatum*, with a genome assembly of around 280 Mb that includes approximately 32,000 protein-coding genes. Gene clusters are separated by long repetitive regions with transposable elements. Analysis of gene family evolution reveals considerable metabolic expansion in the last common ancestor of diplonemids. *P. papillatum* is shown to be predatory, capable of degrading live microeukaryotes, macroalgae, and water plants. The organism is also able to use cell-wall carbohydrates as an energy source and may play a role in bioremediation of eutrophication in temperate coastal waters. Horizontal gene transfer appears to have contributed to carbohydrate-degradation capability in this organism.

The second chapter focuses on the pyruvate dehydrogenase (PDH) complex and its role in eukaryotic metabolism. The PDH complex is responsible for converting pyruvate into acetyl-coenzyme A, a key molecule in the tricarboxylic acid cycle. We found through comparative transcriptome analyses of diplonemids, a diverse group of oceanic protists, the conventional subunits of the PDH complex were absent. Instead, an archaeal-type AceE protein replaced the E1 subunit in the diplonemid ancestor, a substitution also observed in dinoflagellates. The study demonstrates that the mitochondria of *P. papillatum,* a model diplonemid, exhibit pyruvate and 2-oxoglutarate dehydrogenase activities. Mass spectrometry analysis reveals that the AceE protein is as abundant as the E1 subunit of the branched chain ketoacid dehydrogenase complex. We hypothesized that in diplonemids, the PDH complex is composed of the AceE subunit partnering with E2 and E3 subunits from the branched-chain ketoacid dehydrogenase complex and/or the 2-oxoglutarate dehydrogenase complex. This suggests a unique configuration and

functionality of the PDH complex in diplonemids compared to other euglenozoans.

The third chapter of the thesis focuses on the paraflagellar rod, a lattice-like structure found in Euglenozoa, which is absent in the diplonemid *P. papillatum*. Genome analysis of *P. papillatum* revealed homologues to paraflagellar components, prompting investigation into their localization. Five paraflagellar proteins were found to be absent from flagella regions but localized to the cytosol and strongly to the papilla, an apical projecting cell structure of unknown function. The study suggests that relocation events have contributed to the prominence of the papilla in *P. papillatum*. This represents the first efforts to characterize the functionality of the papilla structure within the diplonemid clade and demonstrates the persistent conservation of paraflagellar components within Euglenozoa.

The fourth chapter of this thesis focuses on the identification of homologs of kinetochore and kinetochore-related proteins in diplonemids using bioinformatics techniques. The study found widely conserved Mad2 and homologs of kinetoplastids-specific KKT proteins in diplonemids. Additionally, the study also tagged KKT-10/19, KKT-17/18, CENP-A, and Mad2 in *P. papillatum* and confirmed their localization to the nucleus through immunofluorescence. The kinetochore is a multi-protein complex that designates the binding region of spindle microtubules to chromosomes, and this study provides insight into the conserved nature of kinetochore proteins among eukaryotes, including diplonemids.

Aims of the study

Diplonemids are a group of highly abundant and diverse marine microeukaryotes that belong to the phylum Euglenozoa and are a sister clade to the well-studied groups of medically relevant parasitic kinetoplastids as well as ecologically important photosynthetic euglenids. Limited information is available concerning diplonemids, since so far only a few species have been formally described. However, *P. papillatum* has recently emerged as a tractable model organism, due to its ability to be genetically modified with ease. It is now possible to integrate exogenous DNA into its genome, as well as to tag target proteins and to detect their interacting partners. Using this technology, we sought to investigate a selection of exciting protein candidates, which are either widespread throughout eukaryotes but display peculiar features within diplonemids, or have the potential to offer evolutionary insights into innovations specific to the Euglenozoa clade.

The primary objectives of this study were:

1. **Recent expansion of metabolic versatility in the model diplonemid *Paradiplonema papillatum***

2. **Endogenous tagging of subunit composition for mitochondrial dehydrogenase complexes in marine diplonemids**

3. **Relocation of paraflagellar rod proteins in *Paradiplonema papillatum***

4. **Characterization of a putative kinetochore in diplonemids**

1.0 Introduction

1.1 General overview of Euglenozoa

The phylum Euglenozoa contains a diverse group of flagellated protists that exhibit significant departures from a typical eukaryotic cell (Adl *et al.*, 2019a; Godeanu, 2020). This phylum includes organisms such as parasitic kinetoplastids, free-living marine and freshwater diplonemids, freshwater euglenids, as well as rare marine symbiontids (Faktorová *et al.*, 2016; Adl *et al.*, 2019b). Kinetoplastids are well known for containing the genera *Trypanosoma*, *Leishmania,* and *Phytomonas*, which are pathogenic for vertebrates including humans, as well as plants (Simpson, Lukeš and Roger, 2002a; Stuart *et al.*, 2008).

Kinetoplastids visually manifest as colorless flagellates with unique and prominent morphological features, such as the paraflagellar rod (PFR) in their flagellum (Simpson, 1997) and an unusual mitochondrial DNA, referred to as the kinetoplast (Mensa-Wilmot *et al.*, 2019). The related and well-studied euglenids are best known for their phototrophic representatives, such as the model organism *Euglena gracilis*, which can carry out both photosynthesis and heterotrophic feeding (Suzuki, 2017; Zoltner and Field, 2022) while only very limited information is available for symbiontids (Yubuki, Čepička and Leander, 2016a) (Figure 1). Finally, Diplonemea are heterotrophic marine flagellates initially associated with Euglenida, but now distinguished as a well-defined separate clade (Lukeš, Flegontova and Horák, 2015a). Diplonemids were only recently shown to be one of the most speciose and abundant groups of marine eukaryotes within planktonic communities. As a result, they are also of research interest as they represent a relatively understudied group possessing unique features that can provide insights into the evolution of eukaryotes (Gawryluk *et al.*, 2016a). Recent molecular data suggests that diplonemids constitute tens of thousands of operational taxonomic units are subdivided into four clades represented by just a

handful of morphologically described species that are also available in culture (Flegontova *et al.*, 2020).

1.2 Diplonemids

Diplonemids were first discovered in the early 20th century, but received more consistent attention only about 30 years ago, when they were documented as relatively frequent marine planktonic organisms off the coast of California (Porter, 1973). However, until the beginning of 21st century, only morphological descriptions were available for a few diplonemid species. These studies identified a number of shared features that unify these marine flagellates with kinetoplastids and euglenids (Cavalier-Smith, 2016).

Only relatively recently, molecular techniques were used to identify diplonemids in diverse marine environment, based on high-throughput sequencing of the 18S rRNA gene, which has allowed for their more accurate identification and classification (Schoenle *et al.*, no date; de Vargas *et al.*, 2015; Gawryluk *et al.*, 2016a). At first, only two genera of diplonemids, namely *Diplonema* and *Rhynchopus*, were available in culture (Skuja, 1948). Their morphological features, such as the microtubule-reinforced feeding apparatus, flagellar apparatus with two basal bodies, microtubular root system, and reticulated mitochondria were initially considered diplonemid-specific (Roy *et al.*, 2007; Lukeš, Flegontova and Horák, 2015b).

Additional features from early studies, however, such as the seeming lack of the PFR (see Chapter 3), absence of gliding motility, highly reduced C-shaped collar, and well-developed apical papilla subsequently justified the establishment of a distinct group Diplonemea within Euglenozoa (Simpson, 1997; Simpson, Lukeš and Roger, 2002b; Tashyreva *et al.*, 2022a). While their nuclear genome remained unstudied, their mitochondrial genome was found to be highly unusual, being divided into more than a hundred chromosomes (Marande, Lukeš and Burger,

2005) with the mitochondrial transcripts *trans*-spliced and edited in an extremely complicated manner (Marande and Burger, 2007; Kaur *et al.*, 2020).

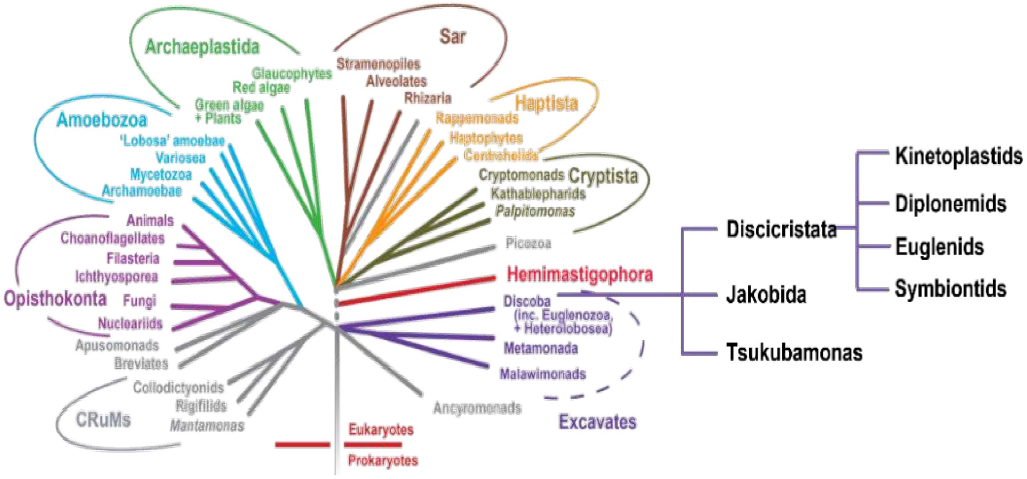


Figure 1. The eukaryotic tree of life. This tree is based on eukaryotic phylogenomic studies. Defunct supergroup of Excavates indicated by purple with broken lines contains the phylum Euglenozoa. The taxonomic sub-division of Euglenozoa has been corroborated and is shown with purple dash (Burki *et al*. 2019).

Despite these discoveries, diplonemids were considered an ecologically marginal and evolutionarily rather uninformative group, until they were relatively recently, and thus unexpectedly, identified as extremely abundant and diverse eukaryotes of the world ocean (Lukeš, Flegontova and Horák, 2015b).

1.3 Abundance and diversity of diplonemids

The *Tara Oceans* expedition aimed to explore the diversity and distribution of marine planktonic organisms. It proved crucial for advancing our understanding of diplonemids, as it provided a large dataset of diplonemid DNA sequences (de Vargas *et al*. 2015; Lukeš *et al*. 2015; Flegontova *et al*. 2016). This data allowed us to study the diversity, distribution, and evolution of diplonemids on a global

scale and led to the discovery of their unanticipated diversity and species richness. Indeed, the global survey by the *Tara Oceans* circumnavigation expedition identified diplonemids as the sixth most abundant eukaryotic group based on the V9 region, which is a specific region of the 18S rRNA gene (De Vargas *et al.*, 2015; Flegontova *et al.*, 2016a; Boeuf *et al.*, 2019; Obiol *et al.*, 2020) (Figure 2A). The majority of V9 DNA barcodes, grouped into operational taxonomic units, which were identified as diplonemids came from deep sea samples, indicating that these organisms have adapted to survive in the extreme conditions found in the deep ocean (Lukeš, Flegontova and Horák, 2015b) (Figure 2A). As a result, this huge group turned into an important player in the marine ecosystem (Sauvadet, Gobet and Guillou, 2010). We have very little information about the nutrition of diplonemids, and thus their ecological functions, but latest data identified them as efficient bacteriovores (Prokopchuk *et al.*, 2019), although parasitism is most likely the life strategy of at least some diplonemids (Tashyreva *et al.*, 2022a).

Based on 18S rRNA sequences, diplonemids are currently divided into four major lineages: (i) the "classical" diplonemids (Diplonemidae), which comprise both benthic and planktonic species; (ii) the hemistasiids (Hemistasiidae), a small planktonic clade; (iii) a highly diverse clade of deep-sea pelagic diplonemids, known as DSPD I, which has recently been named Eupelagonemidae; and (iv) DSPD II, a relatively small clade of deep-sea pelagic diplonemids (Lara *et al.*, 2009; Flegontova *et al.*, 2016a; Okamoto *et al.*, 2019) (Figure 2B). The diversity of diplonemids in the DSPD I clade is known only from a single-cell genomic survey (Gawryluk *et al.*, 2016), while some morphological and molecular data are available from members of the genera *Diplonemida*, *Hemistasia,* and *Eupelagonema* (Okamoto *et al.*, 2019; Tashyreva *et al.*, 2022a). The best-known

representative is *Paradiplonema papillatum*, a sack-shaped cell equipped with two short, thin flagella and a well-developed apical papilla (Prokopchuk *et al.*, 2019).

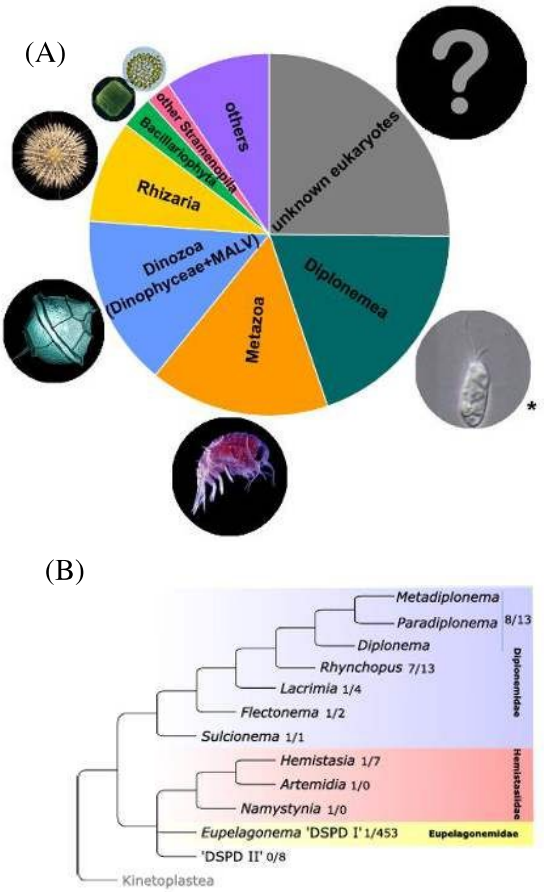


Figure 2. (**A**) Pie chart showing the greatest number of operational taxonomic units corresponding to diplonemids from the *Tara Oceans* expedition (Lukeš, Flegontova and Horák, 2015b). **(B)** A cladogram demonstrating the phylogenetic relationships between various diplonemid lineages. The analysis is based on 18S rRNA gene sequences, with Kinetoplastea serving as an outgroup. The numerical values displayed next to each taxon signify the number of formally described species and the number of 18S rRNA gene sequences that have been deposited in the EukRef database (Tashyreva *et al.*, 2022a).

1.4 Different morphological stages of *P. papillatum*

Some diplonemid species exhibit distinct morphological stages, classified as the trophic, sessile, and swimming phases (Tashyreva *et al.*, 2022a). Their morphology and behavior are defined by the availability of nutrients in their environment (Simpson, 1997). *Paradiplonema* species exhibit a distinct swimming behavior, since unlike *Diplonema*, *Metadiplonema*, *Rhynchopus*, and *Hemistasia* species, they do not show the typical 'spinning lasso' pattern of the anterior flagellum (Skuja, 1948; Roy *et al.*, 2007; Tashyreva *et al.*, 2018). In contrast, their flagella bend differently to enable rotational swimming in nutrition and nutrient-deprived conditions (Figure 3). As nutrient availability declines, the elongated trophic cells gradually become more compact and acquire a diamond-like shape. Eventually, they enter a rounded mucilage-enwrapped stage that resembles a cyst, with all morphologies of this species lacking the PFR and extrusomes (Tashyreva *et al.*, 2018).

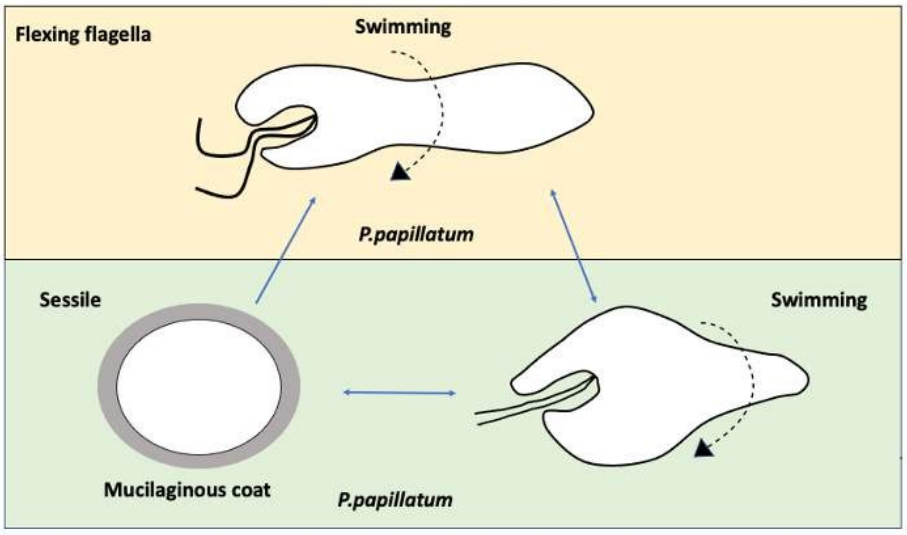


Figure 3. A schematic diagram is presented to depict the life stages manifested by cultured *P. papillatum* showcasing minor behavioral and morphological variations in

response to changing conditions, represented by a lack of nutrition (green background) (Tashyreva *et al.*, 2022a).

1.5 Comparison of diplonemids with other euglenozoans

The primary limiting factor for a thorough comparative analysis of diplonemids with other euglenozoans was the absence of a whole nuclear genome sequence, which has changed only very recently (Valach, Moreira, Petitjean, Benz and Butenko, 2023). Prior to this, molecular characters, such as polycistronic transcription, *trans*-splicing, giant mitochondrial genomes, and others have been identified that served to further unify these marine flagellates with kinetoplastids and euglenids (Butenko *et al.*, 2020). Diplonemids possess huge amount of mitochondrial DNA, organized in the form of inter-catenated circles similar to kinetoplastids (Valach *et al.*, 2020) (Figure 4). However, unlike kinetoplastids, the DNA is not concentrated in a densely packed region but rather evenly distributed throughout the lumen (Cavalcanti and De Souza, 2018).

Another shared feature of Euglenozoa is the microtubule corset or spiral array of microtubules that lies just beneath the plasma membrane (Cavalier-Smith, 2017). This structure provides support and shape to the cell and is also likely involved in its movement (Simpson, 2016). Euglenozoans lack a rigid cell wall, with only euglenids possessing a flexible pellicle that seemingly allows for changes in cell shape and movement (Kostygov *et al.*, 2021). While lacking this pellicle, diplonemids also display an undulating cell movement initially seen in euglenids, referred to as ‘metaboly’ (Tashyreva *et al.*, 2022b). It is noteworthy that diplonemids invariably possess two flagella, whereas kinetoplastids mostly develop a single flagellum, and in euglenids, the number of flagella is even more variable (Gawryluk *et al.*, 2016). While euglenids are found in a variety of environments, including freshwater, marine, and soil habitats, with their life style varying from photosynthetic to heterotrophic (Von Der Heyden *et al.*, 2004;

Cavalier-Smith, 2010; Kostygov *et al.*, 2021), diplonemids are found almost exclusively in marine environments and are strictly heterotrophic (Flegontova *et al.*, 2016a).

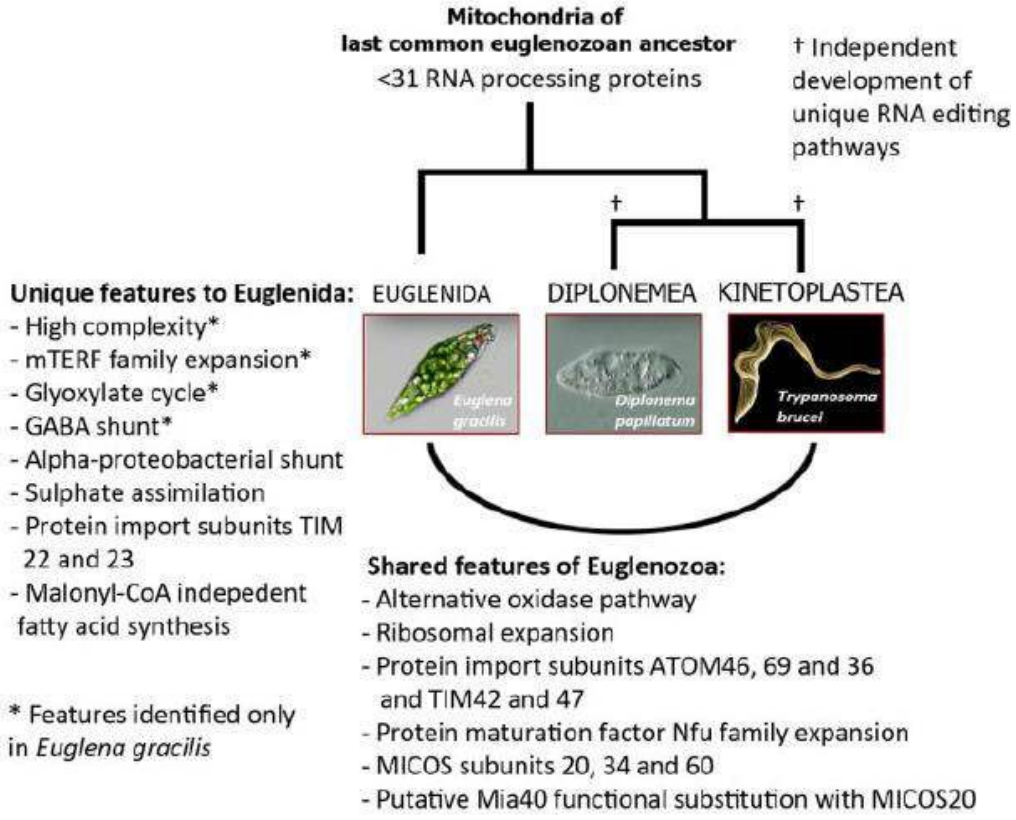


Figure 4. An evolutionary diagram highlighting distinctive mitochondrial characteristics shared among Euglenozoa, while also highlighting specific innovations of euglenids (Hammond *et al.*, 2020).

2. Genetically tractable protists as emerging model organisms

Advanced technology like high-throughput sequencing holds the potential to boost our knowledge of the diversity of marine protists. Indeed, sequences of

millions of genes are now available but almost none have been functionally studied (Carradec *et al.*, 2018). The ability to conduct a systematic analysis of gene regulation, expression, and protein-protein interactions is contingent upon the availability of genetically tractable model organisms, which can provide insights into the biology of the studied organisms (Teichert *et al.*, 2014; Teichert, Pöggeler and Nowrousian, 2020). Functional studies through reverse genetic approaches such as gene replacement *via* homologous recombination, RNA interference, and CRISPR/Cas9 methods represent essential tools for this process. Until recently, only very few genetically tractable marine protists have been developed, such as *Thalassiosira pseudonana, Chaetoceros gracilis,* and *Parabodo caudatus* (Ifuku *et al.*, 2015; Hopes *et al.*, 2016; Gomaa *et al.*, 2017).Recent discoveries demonstrated the advancement of a technique for targeted gene integration *via* homologous recombination in the best-studied diplonemid *P. papillatum*. This approach was used to generate *in situ* tagging and targeting of sequences into its genome, leading to the establishment of stable transformants, providing a platform for functional studies (Faktorová et al., 2020a).

In the process of generating a model organism, relevant characteristics such as the accessibility of corresponding genome and transcriptome data, short generation time, and ease of cultivation in laboratory settings are highly desirable (Faktorová *et al.,* 2020b). While a few diplonemid species have been cataloged in the American Type Culture Collection, they still lack both morphological descriptions and sequenced genomes. *P. papillatum*, however, represents an exception, as a substantial amount of information is available for this flagellate and its nuclear genome information has now been published, allowing systematic functional and cell biology studies (Valach, Moreira, Petitjean, Benz and Butenko, 2023). Its rapid growth in artificial seawater, high cell density (6×10^6 cells/ml) with a ~12-

hours doubling time, and ease of cryopreservation further suggest its potential to serve as a model species for this group of protists (Kaur *et al.*, 2018a).

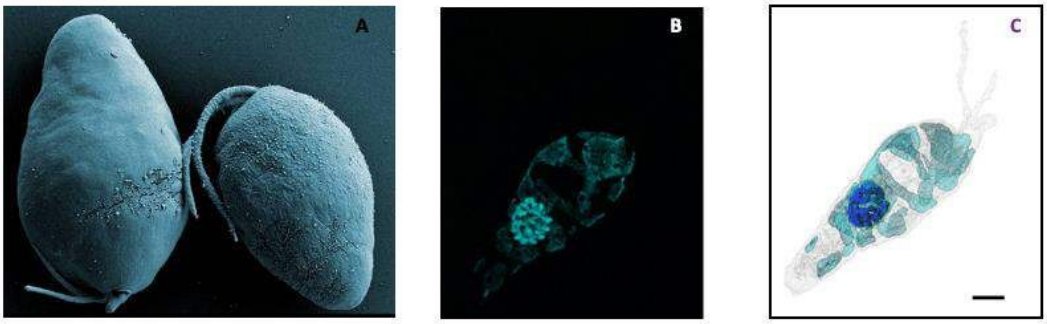


Figure 5. Ultrastructure of *P. papillatum*: The scanning electron microscopy image illustrates the anterior end of the cell, which exhibits two flagella emerging from the flagellar pocket on the left, a cytopharynx on the right, and a prominent lip-like papillum between them (Marande, Lukeš and Burger, 2005), Panel B shows a three-dimensional model generated using expansion microscopy, with DNA stained with DAPI (cyan) revealing the reticulated mitochondrion and nucleus. In panel C, the light gray indicates cell-surface tubulin, and nuclear heterochromatin (in blue) which was visualized through staining with both DAPI and the anti-tubulin antibody (Valach, Moreira, Petitjean, Benz and Butenko, 2023).

3. Genomic view of *P. papillatum*

The kinetoplastids represent the only euglenozoans for which high-quality nuclear genomes are available (Kostygov *et al.*, 2021). Until very recently, the genome and transcriptome information available for diplonemids was insufficient, with only a few gene and partial single-cell genome sequences available for studies related to phylogenies, biodiversity, and metabolic adaptations (Flegontova *et al.*, 2016b; Butenko *et al.*, 2020). Although this data is useful for addressing specific research questions, they are inadequate for comprehensively understanding the functions encoded in, and the

evolutionary patterns of diplonemid nuclear genomes. Additionally, the currently assembled and annotated nuclear genome of *P. papillatum* of estimated size approximately 280 Mbp is characterized by high levels of repetition, particularly in intergenic regions, and is estimated to contain around 32,000 protein-coding genes. These genes are thought to be co-transcribed in groups of up to 100 (Valach *et al.*, 2016; Kaur *et al.*, 2020; Valach, Moreira, Petitjean, Benz and Butenko, 2023). Based on the transcriptomic data, the coding capacity of other diplonemid genomes is estimated to largely vary, with the gene number of *Rhynchopus humris* and *Hemistasia phaeocysticola* ranging from 37,000 to 52,000, respectively (Kaur *et al.*, 2020). The recent availability of the nuclear genome of *P. papillatum* (Valach et al., 2023) enables comprehensive functional and cell biology investigations.

4.1 Exploring intriguing metabolic pathways of diplonemids

A combination of transcriptomic, proteomic, and metabolomic approaches supplemented with biochemical experiments have been utilized to examine the metabolism of *P. papillatum* (Škodová-Sveráková *et al.*, 2021). This species is currently the only diplonemid that has been extensively studied at a metabolic level. *P. papillatum* possesses the complete enzymatic machinery required for fundamental metabolic pathways such as glycolysis, gluconeogenesis, the pentose phosphate pathway, the tricarboxylic acid (TCA) cycle, synthesis, β-oxidation of fatty acids, and oxidative phosphorylation (OXPHOS) (Morales *et al.*, 2016; Škodová-Sveráková *et al.*, 2021) (Figure 6). These pathways collectively enable *P. papillatum* to generate ATP through both substrate-level phosphorylation and OXPHOS. Moreover, it was demonstrated that in this protist gluconeogenesis is more prevalent than glycolysis (Škodová-Sveráková *et al.*, 2020b).

However, the glucose anabolic pathway was recently reconstructed up to glucose synthesis, revealing that while *P. papillatum* possesses the enzymatic machinery required for glycolysis, it lacks crucial protein domains necessary for the functionality of the annotated glucose-6-phosphatase protein sequence (Škodová-Sveráková *et al.*, 2020b). As a result, its glycolysis may only proceed up to the generation of glucose-6-phosphate, which can be utilized by the pentose phosphate pathway or converted into paramylon, a type of β-1,3 glucan commonly found in euglenids (Škodová-Sveráková *et al.*, 2020) (Figure 6). The presence of transcripts of glucan synthase, a key enzyme responsible for the synthesis of paramylon, has been identified in the transcriptomes of several diplonemids, indicating that these organisms may also have the ability to produce this macromolecule (Škodová-Sveráková *et al.*, 2021). The function of paramylon in diplonemids appears to differ from its function in *Euglena gracilis*, where it accumulates under excess nutrients (Barsanti *et al.*, 2001). In *P. papillatum*, paramylon appears to be synthesized under nutrient-poor conditions and may serve a role in the development of resting cysts or increasing cell density, enabling the organisms to sink into deeper ocean strata where nutrient availability is potentially more favorable (Škodová-Sveráková, Prokopchuk, *et al.*, 2020). Additionally, diplonemids can synthesize all nucleotides, 20 amino acids, and vitamins, and their fatty acids are synthesized using the cytosolic fatty acid synthesis pathway (Butenko *et al.*, 2020). Although the majority of fatty acid synthesis II enzymes have been identified, the crucial malonyl-coenzyme A-ACP transacylase enzyme responsible for the transfer of malonyl to ACP in the initial step is missing from all transcriptomes (Zimorski *et al.*, 2017; Škodová-Sveráková *et al.*, 2020a). The enzymes involved in anaerobic processes, such as pyruvate NADP+ oxidoreductase, fumarate reductase, enoyl-coenzyme A reductase, lactate dehydrogenase, and opine dehydrogenase have also been identified in *P.*

papillatum (Nakazawa *et al.*, 2017; Škodová-Sveráková *et al.*, 2021). However, the exact mechanisms of their functionality require further investigation. Interestingly, when deprived of oxygen, the metabolic and division rates of *P. papillatum* decreased (Škodová-Sveráková *et al.*, 2021). Overall, these findings provide insight into the unique features and metabolic capabilities of diplonemids.

4.2 Insights into pyruvate dehydrogenase complex and tricarboxylic acid cycle: A metabolic perspective

Pyruvate is the end product of glycolysis, the process by which glucose is broken down for energy metabolism (Cazzulo, 1992). After glycolysis, pyruvate is converted into acetyl-coenzyme A, which serves as the primary substrate for a series of biochemical reactions known as the TCA cycle (Meléndez-Hevia, Waddell and Montero, 1994; Nakazawa *et al.*, 2017). Pyruvate is converted into acetyl-coenzyme A by the pyruvate dehydrogenase (PDH) complex in the mitochondrial matrix (Reed and Oliver, 1982; Danson, 1988; Nakazawa *et al.*, 2017). Evolutionarily related to PDH are two other dehydrogenase complexes, namely 2-oxoglutarate dehydrogenase (OGDH) and branched-chain ketoacid dehydrogenase (BCKDH), which were likely present in the last eukaryotic common ancestor (Whittle *et al.*, 2023) Figure 6).

The functioning of the TCA cycle in *P. papillatum* is complex due to the presence of alternative enzymes, 2-oxoglutarate decarboxylase (OGDC) and succinate-semialdehyde dehydrogenase (SSDH) (Škodová-Sveráková *et al.*, 2021) (Figure 6). Both pairs of enzymes convert 2-oxoglutarate to succinate but with different metabolic outcomes. The classical TCA cycle produces NADH and ATP, whereas the alternative shunt produces NADPH instead of NADH and generates no ATP or GTP (Morales *et al.*, 2016; Škodová-Sveráková *et al.*, 2021). OGDC catalyzes an irreversible decarboxylation reaction, while OGDH catalyzes a reversible

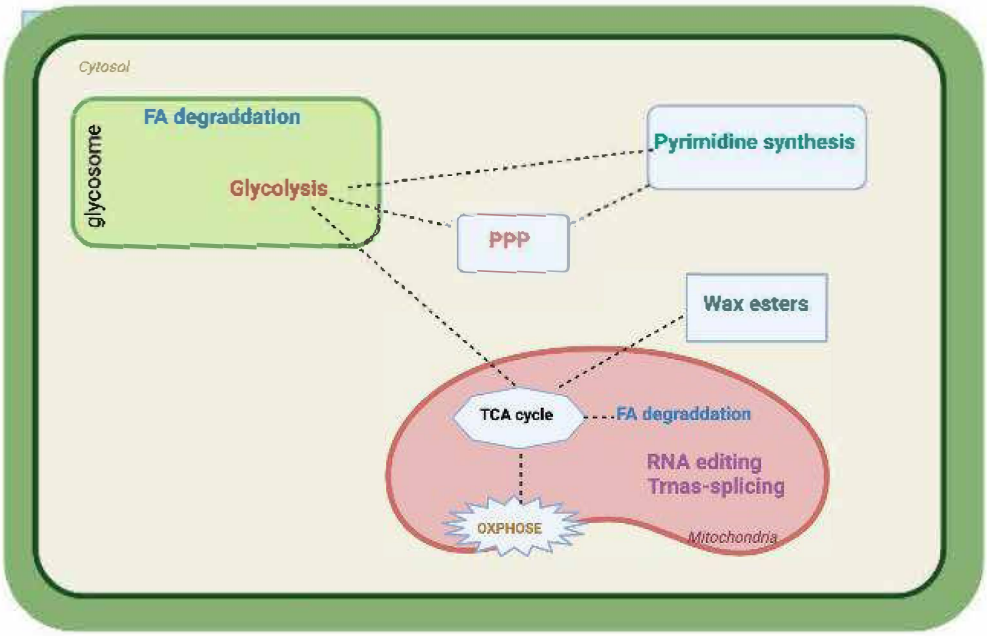


Figure 6. Major metabolic pathways of *P. papillatum*. Transcriptome and proteome were analyzed to predict metabolic pathways. Different boxes indicate pathways so far identified in *P. papillatum*. FA - fatty acids; PPP - pentose phosphate pathway; TCA - tricarboxylic acid.

reaction, which could potentially lead to both catabolic and anabolic outcomes. The specific conditions under which each pathway operates are unknown, as *P. papillatum* is the first organism where both branches have been identified simultaneously (Novák Vanclová *et al.*, 2020; Butenko *et al.*, 2021; Škodová-Sveráková *et al.*, 2021).

The prevalence of OGDC in the presence of oxygen implies its involvement in the aerobic metabolism, much like in *E. gracilis* where it, together with SSDH, constitutes the exclusive TCA cycle pathways (Shigeoka *et al.*, 1986; He *et al.*, 2021; Ouyang *et al.*, 2021). Acetyl-CoA is predominantly generated from carbohydrate metabolism in *E. gracilis*, while in *P. papillatum*, pyruvate most likely arises from amino acid metabolism due to its preference for gluconeogenesis (Figure 6). *P. papillatum* makes use of oxygen-rich metabolism,

employing all respiratory complexes and alternative enzymes to shuttle electrons from reduced equivalents to oxygen. Its anaerobic enzymes are probably widespread among diplonemids, and may occur also in other groups of marine protists (Škodová-Sveráková *et al.*, 2021).

References

Absalon, S., Blisnick, T., Kohl, L., Toutirais, G., Doré, G., Julkowska, D., Tavenet, A. and Bastin, P., 2008. Intraflagellar transport and functional analysis of genes required for flagellum formation in trypanosomes. *Molecular Biology of the Cell*, *19*(3), 929-944.

Adl, S.M., Bass, D., Lane, C.E., Lukeš, J., Schoch, C.L., Smirnov, A., Agatha, S., Berney, C., Brown, M.W., Burki, F. and Cárdenas, P., 2019. Revisions to the classification, nomenclature, and diversity of eukaryotes. *Journal of Eukaryotic Microbiology*, *66*(1), 4-119.

Barsanti, L., Vismara, R., Passarelli, V. and Gualtieri, P., 2001. Paramylon (β-1, 3-glucan) content in wild type and WZSL mutant of Euglena gracilis. Effects of growth conditions. *Journal of Applied Phycology*, *13*, 59-65.

Bastin, P., MacRae, T.H., Francis, S.B., Matthews, K.R. and Gull, K., 1999. Flagellar morphogenesis: protein targeting and assembly in the paraflagellar rod of trypanosomes. *Molecular and Cellular Biology*, *19*(12), 8191-8200.

Bastin, P., Matthews, K.R. and Gull, K., 1996. The paraflagellar rod of kinetoplastida: solved and unsolved questions. *Parasitology Today*, *12*(8), 302-307.

Berg, H.C., 2003. The rotary motor of bacterial flagella. *Annual Review of Biochemistry*, *72*(1), 19-54.

Boeuf, D., Edwards, B.R., Eppley, J.M., Hu, S.K., Poff, K.E., Romano, A.E., Caron,

D.A., Karl, D.M. and DeLong, E.F., 2019. Biological composition and microbial dynamics of sinking particulate organic matter at abyssal depths in the oligotrophic open ocean. *Proceedings of the National Academy of Sciences*, *116*(24), 11824-11832.

Butenko, A., Opperdoes, F.R., Flegontova, O., Horák, A., Hampl, V., Keeling, P., Gawryluk, R.M., Tikhonenkov, D., Flegontov, P. and Lukeš, J., 2020. Evolution of metabolic capabilities and molecular features of diplonemids, kinetoplastids, and euglenids. *BMC Biology*, *18*(1), 1-28.

Butenko, A., Hammond, M., Field, M.C., Ginger, M.L., Yurchenko, V. and Lukeš, J., 2021. Reductionist pathways for parasitism in euglenozoans? Expanded datasets provide new insights. *Trends in Parasitology*, *37*(2), 100-116.

Cachon, J., Cachon, M., Cosson, M.P. and Cosson, J., 1988. The paraflagellar rod: a structure in search of a function. *Biology of the Cell*, *63*(2), 169-181.

Carradec, Q., Pelletier, E., Da Silva, C., Alberti, A., Seeleuthner, Y., Blanc-Mathieu, R., Lima-Mendez, G., Rocha, F., Tirichine, L., Labadie, K. and Kirilovsky, A., 2018. A global ocean atlas of eukaryotic genes. *Nature Communications*, *9*(1), 373.

Cavalcanti, D.P. and de Souza, W., 2018. The kinetoplast of trypanosomatids: from early studies of electron microscopy to recent advances in atomic force microscopy. *Scanning*, *2018*.

Cavalier-Smith, T., 2010. Kingdoms Protozoa and Chromista and the eozoan root of the eukaryotic tree. *Biology Letters*, *6*(3), 342-345.

Cavalier-Smith, T., 2016. Higher classification and phylogeny of Euglenozoa. *European Journal of Protistology*, *56*, 250-276.

Cavalier-Smith, T., 2017. Euglenoid pellicle morphogenesis and evolution in light of comparative ultrastructure and trypanosomatid biology: Semi-conservative microtubule/strip duplication, strip shaping and transformation. *European Journal of Protistology*, *61*, 137-179.

Cazzulo, J.J., 1992. Aerobic fermentation of glucose by trypanosomatids. *The FASEB Journal*, *6*(13), 3153-3161.

Coceres, V.M., Iriarte, L.S., Miranda-Magalhães, A., Santos de Andrade, T.A., de Miguel, N. and Pereira-Neves, A., 2021. Ultrastructural and functional analysis of a novel extra-axonemal structure in parasitic trichomonads. *Frontiers in Cellular and Infection Microbiology*, 1101.

Danson, M.J., 1988. Dihydrolipoamide dehydrogenase: a 'new'function for an old enzyme?. *Biochemical Society Transactions*, *16*(2), 87-89.

Elbrächter, M., Schnepf, E. and Balzer, I., 1996. Hemistasia phaeocysticola (Scherffel) comb. nov., redescription of a free-living, marine, phagotrophic kinetoplastid flagellate. *Archiv für Protistenkunde*, *147*(2),125-136.

Drahomíra, F., Eva, D., Peña-Diaz, P. and Julius, L., 2016. From simple to supercomplex: mitochondrial genomes of euglenozoan protists. *F1000Research*, *5*.

Faktorová, D., Kaur, B., Valach, M., Graf, L., Benz, C., Burger, G. and Lukeš, J., 2020. Targeted integration by homologous recombination enables in situ tagging and replacement of genes in the marine microeukaryote Diplonema papillatum. *Environmental Microbiology*, *22*(9), 3660-3670.

Faktorová, D., Nisbet, R.E.R., Fernández Robledo, J.A., Casacuberta, E., Sudek, L., Allen, A.E., Ares Jr, M., Aresté, C., Balestreri, C., Barbrook, A.C. and Beardslee, P.,

2020. Genetic tool development in marine protists: emerging model organisms for experimental cell biology. *Nature Methods*, *17*(5), 481-494.

Flegontova, O., 2017. Diversity and biogeography of diplonemid and kinetoplastid protists in global marine plankton.

Flegontova, O., Flegontov, P., Malviya, S., Audic, S., Wincker, P., de Vargas, C., Bowler, C., Lukeš, J. and Horák, A., 2016. Extreme diversity of diplonemid eukaryotes in the ocean. *Current Biology*, *26*(22), 3060-3065.

Flegontova, O., Flegontov, P., Londoño, P.A.C., Walczowski, W., Šantić, D., Edgcomb, V.P., Lukeš, J. and Horák, A., 2020. Environmental determinants of the distribution of planktonic diplonemids and kinetoplastids in the oceans. *Environmental Microbiology*, *22*(9), 4014-4031.

Gadelha, C., Wickstead, B., de Souza, W., Gull, K. and Cunha-e-Silva, N., 2005. Cryptic paraflagellar rod in endosymbiont-containing kinetoplastid protozoa. *Eukaryotic Cell*, *4*(3), 516-525.

Gadelha, C., Wickstead, B., McKean, P.G. and Gull, K., 2006. Basal body and flagellum mutants reveal a rotational constraint of the central pair microtubules in the axonemes of trypanosomes. *Journal of Cell Science*, *119*(12), 2405-2413.

Gawryluk, R.M., Del Campo, J., Okamoto, N., Strassert, J.F., Lukeš, J., Richards, T.A., Worden, A.Z., Santoro, A.E. and Keeling, P.J., 2016. Morphological identification and single-cell genomics of marine diplonemids. *Current Biology*, *26*(22), 3053-3059.

Gomaa, F., Garcia, P.A., Delaney, J., Girguis, P.R., Buie, C.R. and Edgcomb, V.P., 2017. Toward establishing model organisms for marine protists: Successful transfection protocols for Parabodo caudatus (Kinetoplastida: Excavata). *Environmental Microbiology*, *19*(9), 3487-3499.

Hammond, M.J., Nenarokova, A., Butenko, A., Zoltner, M., Dobáková, E.L., Field, M.C. and Lukeš, J., 2020. A uniquely complex mitochondrial proteome from Euglena gracilis. *Molecular Biology and Evolution*, *37*(8), 2173-2191.

He, J., Liu, C., Du, M., Zhou, X., Hu, Z., Lei, A. and Wang, J., 2021. Metabolic responses of a model green microalga Euglena gracilis to different environmental stresses. *Frontiers in Bioengineering and Biotechnology*, *9*, 662655.

von der Heyden, S., Chao, E.E., Vickerman, K. and CAVALIER-SMITH, T.H.O.M.A.S., 2004. Ribosomal RNA phylogeny of bodonid and diplonemid flagellates and the evolution of Euglenozoa. *Journal of Eukaryotic Microbiology*, *51*(4), 402-416.

Hopes, A., Nekrasov, V., Kamoun, S. and Mock, T., 2016. Editing of the urease gene by CRISPR-Cas in the diatom Thalassiosira pseudonana. *Plant Methods*, *12*(1), 1-12.

Ifuku, K., Yan, D., Miyahara, M., Inoue-Kashino, N., Yamamoto, Y.Y. and Kashino, Y., 2015. A stable and efficient nuclear transformation system for the diatom Chaetoceros gracilis. *Photosynthesis Research*, *123*, 203-211.

Kaur, B., Valach, M., Peña-Diaz, P., Moreira, S., Keeling, P.J., Burger, G., Lukeš, J. and Faktorová, D., 2018. Transformation of Diplonema papillatum, the type species of the highly diverse and abundant marine microeukaryotes Diplonemida (Euglenozoa). *Environmental Microbiology*, *20*(3), 1030-1040.

Kaur, B., Záhonová, K., Valach, M., Faktorová, D., Prokopchuk, G., Burger, G. and Lukeš, J., 2020. Gene fragmentation and RNA editing without borders: eccentric mitochondrial genomes of diplonemids. *Nucleic Acids Research*, *48*(5), 2694-2708.

Klena, N. and Pigino, G., 2022. Structural biology of cilia and intraflagellar transport. *Annual Review of Cell and Developmental Biology*, *38*, 103-123.

Kohil, L., Sherwin, T. and Gull, K., 1999. Assembly of the paraflagellar rod and the flagellum attachment zone complex during the Trypanosoma brucei cell cycle. *Journal of Eukaryotic Microbiology*, *46*(2), 105-109.

Kostygov, A.Y., Karnkowska, A., Votýpka, J., Tashyreva, D., Maciszewski, K., Yurchenko, V. and Lukeš, J., 2021. Euglenozoa: taxonomy, diversity and ecology, symbioses and viruses. *Open Biology*, *11*(3), 200407.

Krueger, T. and Engstler, M., 2015, October. Flagellar motility in eukaryotic human parasites. In *Seminars in Cell & Developmental Biology* (Vol. 46, 113-127). Academic Press.

Lacomble, S., Vaughan, S., Gadelha, C., Morphew, M.K., Shaw, M.K., McIntosh, J.R. and Gull, K., 2009. Three-dimensional cellular architecture of the flagellar pocket and associated cytoskeleton in trypanosomes revealed by electron microscope tomography. *Journal of Cell Science*, *122*(8), 1081-1090.

Lara, E., Moreira, D., Vereshchaka, A. and López-García, P., 2009. Pan-oceanic distribution of new highly diverse clades of deep-sea diplonemids. *Environmental Microbiology*, *11*(1), 47-55.

Lehninger, A. L. (1993) 'Lehninger principles of biochemistry, Second Edition', *W H Freeman*. (Accessed: 13 May 2023).

Lukeš, J., Flegontova, O. and Horák, A. (2015) 'Diplonemids', *Current Biology*. Cell Press, 25(16).

Maharana, B.R., Rao, J.R., Tewari, A.K., Singh, H., Allaie, I.M. and Varghese, A., 2014. Molecular characterisation of paraflagellar rod protein gene (PFR) of Trypanosoma evansi. *Journal of Applied Animal Research*, *42*(1),1-5.

Marande, W. and Burger, G., 2007. Mitochondrial DNA as a genomic jigsaw puzzle. *Science*, *318*(5849), 415-415.

Marande, W., Lukeš, J. and Burger, G., 2005. Unique mitochondrial genome structure in diplonemids, the sister group of kinetoplastids. *Eukaryotic Cell*, *4*(6),1137-1146.

Maric, D., Epting, C.L. and Engman, D.M., 2010. Composition and sensory function of the trypanosome flagellar membrane. *Current Opinion in Microbiology*, *13*(4), 466-472.

Meléndez-Hevia, E., Waddell, T.G. and Montero, F., 1994. Optimization of metabolism: the evolution of metabolic pathways toward simplicity through the game of the pentose phosphate cycle. *Journal of Theoretical Biology*, *166*(2), 201-220.

Mensa-Wilmot, K., Hoffman, B., Wiedeman, J., Sullenberger, C. and Sharma, A., 2019. Kinetoplast division factors in a trypanosome. *Trends in Parasitology*, *35*(2),119-128.

Morales, J., Hashimoto, M., Williams, T.A., Hirawake-Mogi, H., Makiuchi, T., Tsubouchi, A., Kaga, N., Taka, H., Fujimura, T., Koike, M. and Mita, T., 2016. Differential remodelling of peroxisome function underpins the environmental and metabolic adaptability of diplonemids and kinetoplastids. *Proceedings of the Royal Society B: Biological Sciences*, *283*(1830), 20160520.

Mukherjee, I., Salcher, M.M., Andrei, A.Ş., Kavagutti, V.S., Shabarova, T., Grujčić, V., Haber, M., Layoun, P., Hodoki, Y., Nakano, S.I. and Šimek, K., 2020. A freshwater radiation of diplonemids. *Environmental Microbiology*, *22*(11), 4658-4668.

Mul, W., Mitra, A. and Peterman, E.J., 2022. Mechanisms of regulation in intraflagellar transport. *Cells*, *11*(17), 2737.

Nakazawa, M., 2017. C2 metabolism in Euglena. *Euglena: Biochemistry, Cell and Molecular Biology*, 39-45.

Nakazawa, M., Hayashi, R., Takenaka, S., Inui, H., Ishikawa, T., Ueda, M., Sakamoto, T., Nakano, Y. and Miyatake, K., 2017. Physiological functions of pyruvate: NADP+ oxidoreductase and 2-oxoglutarate decarboxylase in Euglena gracilis under aerobic and anaerobic conditions. *Bioscience, Biotechnology, and Biochemistry*, *81*(7),1386-1393.

Novák Vanclová, A.M., Zoltner, M., Kelly, S., Soukal, P., Záhonová, K., Füssy, Z., Ebenezer, T.E., Lacová Dobáková, E., Eliáš, M., Lukeš, J. and Field, M.C., 2020. Metabolic quirks and the colourful history of the Euglena gracilis secondary plastid. *New Phytologist*, *225*(4),1578-1592.

Obiol, A., Giner, C.R., Sánchez, P., Duarte, C.M., Acinas, S.G. and Massana, R., 2020. A metagenomic assessment of microbial eukaryotic diversity in the global ocean. *Molecular Ecology Resources*, *20*(3), 718-731.

Okamoto, N., Gawryluk, R.M., Del Campo, J., Strassert, J.F., Lukeš, J., Richards, T.A., Worden, A.Z., Santoro, A.E. and Keeling, P.J., 2019. A Revised Taxonomy of Diplonemids Including the Eupelagonemidae n. fam. and a Type Species, Eupelagonema oceanica n. gen. & sp. *Journal of Eukaryotic Microbiology*, *66*(3), 519-524.

Ouyang, Y., Chen, S., Zhao, L., Song, Y., Lei, A., He, J. and Wang, J., 2021. Global

metabolomics reveals that Vibrio natriegens enhances the growth and paramylon synthesis of Euglena gracilis. *Frontiers in Bioengineering and Biotechnology*, *9*, 652021.

Pinskey, J.M., Lagisetty, A., Gui, L., Phan, N., Reetz, E., Tavakoli, A., Fu, G. and Nicastro, D., 2022. Three-dimensional flagella structures from animals' closest unicellular relatives, the Choanoflagellates. *Elife*, *11*, 78133.

Porter, D., 1973. Isonema papillatum sp. n., a new colorless marine flagellate: a light-and electronmicroscopic study. *The Journal of Protozoology*, *20*(3), 351-356.

Portman, N. and Gull, K., 2010. The paraflagellar rod of kinetoplastid parasites: from structure to components and function. *International Journal for Parasitology*, *40*(2), 135-148.

Prokopchuk, G., Tashyreva, D., Yabuki, A., Horák, A., Masařová, P. and Lukeš, J., 2019. Morphological, ultrastructural, motility and evolutionary characterization of two new Hemistasiidae species. *Protist*, *170*(3),259-282.

Prokopchuk, G., Korytář, T., Juricová, V., Majstorović, J., Horák, A., Šimek, K. and Lukeš, J., 2022. Trophic flexibility of marine diplonemids-switching from osmotrophy to bacterivory. *The ISME Journal*, *16*(5),1409-1419.

Reed, L.J. and Oliver, R.M., 1982. Structure-function relationships in pyruvate and α-ketoglutarate dehydrogenase complexes. *Structure and Function Relationships in Biochemical Systems*, 231-241.

Rosati, G., Verni, F., Barsanti, L., Passarelli, V. and Gualtieri, P., 1991. Ultrastructure of the apical zone of Euglena gracilis: photoreceptors and motor apparatus. *Electron Microscopy Reviews*, *4*(2), 319-342.

Rotureau, B., Ooi, C.P., Huet, D., Perrot, S. and Bastin, P., 2014. Forward motility is essential for trypanosome infection in the tsetse fly. *Cellular Microbiology*, *16*(3), 425-433.

Roy, J., Faktorova, D., BENADA, O., LUKEŠ, J. and Burger, G., 2007. Description of Rhynchopus euleeides n. sp.(Diplonemea), a free-living marine euglenozoan. *Journal of Eukaryotic Microbiology*, *54*(2), 137-145.

Satir, P. and Christensen, S.T., 2007. Overview of structure and function of mammalian cilia. *Annu. Rev. Physiol.*, *69*, 377-400.

Sauvadet, A.L., Gobet, A. and Guillou, L., 2010. Comparative analysis between protist communities from the deep-sea pelagic ecosystem and specific deep hydrothermal habitats. *Environmental Microbiology*, *12*(11), 2946-2964.

Schoenle, A., Hohlfeld, M., Hermanns, K., Mahé, F., de Vargas, C., Nitsche, F. and Arndt, H., 2021. High and specific diversity of protists in the deep-sea basins dominated by diplonemids, kinetoplastids, ciliates and foraminiferans. *Communications Biology*, *4*(1), 501.

Shigeoka, S., Onishi, T., Maeda, K., Nakano, Y. and Kitaoka, S., 1986. Occurrence of thiamin pyrophosphate-dependent 2-oxoglutarate decarboxylase in mitochondria of Euglena gracilis. *FEBS Letters*, *195*(1-2), 43-47.

Sigman, D.M. and Hain, M.P., 2012. The biological productivity of the ocean. *Nature Education Knowledge*, *3*(10), 21.

Simpson, A.G., 1997. The identity and composition of the Euglenozoa. *Archiv Für Protistenkunde*, *148*(3), 318-328.

Simpson, A.G., Lukeš, J. and Roger, A.J., 2002. The evolutionary history of kinetoplastids and their kinetoplasts. *Molecular biology and evolution*, *19*(12), pp.2071-2083.

Simpson, A.G., Lukeš, J. and Roger, A.J., 2002. The evolutionary history of kinetoplastids and their kinetoplasts. *Molecular Biology and Evolution*, *19*(12), 2071-2083.

Škodová-Sveráková, I., Záhonová, K., Bučková, B., Füssy, Z., Yurchenko, V. and Lukeš, J., 2020. Catalase and ascorbate peroxidase in euglenozoan protists. *Pathogens*, *9*(4), 317.

Škodová-Sveráková, I., Prokopchuk, G., Peña-Diaz, P., Záhonová, K., Moos, M., Horváth, A., Šimek, P. and Lukeš, J., 2020. Unique dynamics of paramylon storage in the marine euglenozoan Diplonema papillatum. *Protist*, *171*(2),125717.

Škodová-Sveráková, I., Záhonová, K., Juricová, V., Danchenko, M., Moos, M., Baráth, P., Prokopchuk, G., Butenko, A., Lukáčová, V., Kohútová, L. and Bučková, B., 2021. Highly flexible metabolism of the marine euglenozoan protist Diplonema papillatum. *BMC biology*, *19*(1), 1-21.

Skuja, H., 1948. Taxonomie des phytoplanktons einiger seen in Uppland, Schweden. *Symb. Bot. Ups.*, *9*, pp.1-399.

Stuart, K., Brun, R., Croft, S., Fairlamb, A., Gürtler, R.E., McKerrow, J., Reed, S. and Tarleton, R., 2008. Kinetoplastids: related protozoan pathogens, different diseases. *The Journal of Clinical Investigation*, *118*(4), 1301-1310.

Sunter, J.D. and Gull, K., 2016. The flagellum attachment zone:'the cellular ruler'of trypanosome morphology. *Trends in Parasitology*, *32*(4),309-324.

Strauch, S.M., Schuster, M., Lebert, M., Richter, P., Schmittnagel, M. and Häder, D.P.,

2008, June. A closed ecological system in a space experiment. In *Proceedings of the Symposium Life in Space for Life on Earth* (22-27).

Tashyreva, D., Prokopchuk, G., Votýpka, J., Yabuki, A., Horák, A. and Lukeš, J., 2018. Life cycle, ultrastructure, and phylogeny of new diplonemids and their endosymbiotic bacteria. *MBio*, *9*(2), 02447-17.

Tashyreva, D., Prokopchuk, G., Yabuki, A., Kaur, B., Faktorová, D., Votýpka, J., Kusaka, C., Fujikura, K., Shiratori, T., Ishida, K.I. and Horák, A., 2018. Phylogeny and morphology of new diplonemids from Japan. *Protist*, *169*(2), 158-179.

Tashyreva, D., Simpson, A.G., Prokopchuk, G., Škodová-Sveráková, I., Butenko, A., Hammond, M., George, E.E., Flegontova, O., Zahonova, K., Faktorova, D. and Yabuki, A., 2022. Diplonemids–A Review on" New" Flagellates on the Oceanic Block. *Protist*, *173*(2), 125868.

Teichert, I., Pöggeler, S. and Nowrousian, M., 2020. Sordaria macrospora: 25 years as a model organism for studying the molecular mechanisms of fruiting body development. *Applied Microbiology and Biotechnology*, *104*, 3691-3704.

Valach, M., Moreira, S., Faktorová, D., Lukeš, J. and Burger, G., 2016. Post-transcriptional mending of gene sequences: looking under the hood of mitochondrial gene expression in diplonemids. *RNA Biology*, *13*(12), 1204-1211.

Valach, M., Moreira, S., Petitjean, C., Benz, C., Butenko, A., Flegontova, O., Nenarokova, A., Prokopchuk, G., Batstone, T., Lapébie, P. and Lemogo, L., 2023. Recent expansion of metabolic versatility in Diplonema papillatum, the model species of a highly speciose group of marine eukaryotes. *BMC Biology*, *21*(1), 99.

Wan, K.Y., 2018. Coordination of eukaryotic cilia and flagella. *Essays in*

Biochemistry, *62*(6), 829-838.

Whittle, E.F., Chilian, M., Karimiani, E.G., Progri, H., Buhas, D., Kose, M., Ganetzky, R.D., Toosi, M.B., Torbati, P.N., Badv, R.S. and Shelihan, I., 2023. Biallelic variants in OGDH encoding oxoglutarate dehydrogenase lead to a neurodevelopmental disorder characterized by global developmental delay, movement disorder, and metabolic abnormalities. *Genetics in Medicine*, *25*(2), 100332.

Yubuki, N., Čepička, I. and Leander, B.S., 2016. Evolution of the microtubular cytoskeleton (flagellar apparatus) in parasitic protists. *Molecular and Biochemical Parasitology*, *209*(1-2), 26-34.

Yubuki, N., Čepička, I. and Leander, B.S., 2016. Evolution of the microtubular cytoskeleton (flagellar apparatus) in parasitic protists. *Molecular and Biochemical Parasitology*, *209*(1-2), 26-34.

Yubuki, N., Simpson, A.G. and Leander, B.S., 2013. Reconstruction of the feeding apparatus in Postgaardi mariagerensis provides evidence for character evolution within the Symbiontida (Euglenozoa). *European Journal of Protistology*, *49*(1),32-39.

Zimorski, V., Rauch, C., van Hellemond, J.J., Tielens, A.G. and Martin, W.F., 2017. The mitochondrion of Euglena gracilis. *Euglena: Biochemistry, Cell and Molecular Biology*,19-37.

Zoltner, M. and Field, M.C., 2022. Microbe Profile: Euglena gracilis: photogenic, flexible and hardy. *Microbiology*, *168*(9), 001241.

Chapter 1

Recent expansion of metabolic versatility in the model diplonemid *Paradiplonema papillatum*

Valach *et al. BMC Biology* (2023) 21:99
https://doi.org/10.1186/s12915-023-01563-9

BMC Biology

RESEARCH ARTICLE Open Access

Recent expansion of metabolic versatility in *Diplonema papillatum*, the model species of a highly speciose group of marine eukaryotes

Matus Valach[1*], Sandrine Moreira[1], Celine Petitjean[2], Corinna Benz[3], Anzhelika Butenko[3,4,5], Olga Flegontova[3,5], Anna Nenarokova[2,3], Galina Prokopchuk[3,4], Tom Batstone[2,6], Pascal Lapébie[7], Lionnel Lemogo[1,8], Matt Sarrasin[1], Paul Stretenowich[1,9], Pragya Tripathi[3,4], Euki Yazaki[10], Takeshi Nara[11], Bernard Henrissat[7,12,13], B. Franz Lang[1], Michael W. Gray[14], Tom A. Williams[2], Julius Lukeš[3,4] and Gertraud Burger[1*]

Abstract

Background Diplonemid flagellates are among the most abundant and species-rich of known marine microeukaryotes, colonizing all habitats, depths, and geographic regions of the world ocean. However, little is known about their genomes, biology, and ecological role.

Results We present the first nuclear genome sequence from a diplonemid, the type species *Diplonema papillatum*. The ~ 280-Mb genome assembly contains about 32,000 protein-coding genes, likely co-transcribed in groups of up to 100. Gene clusters are separated by long repetitive regions that include numerous transposable elements, which also reside within introns. Analysis of gene-family evolution reveals that the last common diplonemid ancestor underwent considerable metabolic expansion. *D. papillatum*-specific gains of carbohydrate-degradation capability were apparently acquired via horizontal gene transfer. The predicted breakdown of polysaccharides including pectin and xylan is at odds with reports of peptides being the predominant carbon source of this organism. Secretome analysis together with feeding experiments suggest that *D. papillatum* is predatory, able to degrade cell walls of live microeukaryotes, macroalgae, and water plants, not only for protoplast feeding but also for metabolizing cell-wall carbohydrates as an energy source. The analysis of environmental barcode samples shows that *D. papillatum* is confined to temperate coastal waters, presumably acting in bioremediation of eutrophication.

Conclusions Nuclear genome information will allow systematic functional and cell-biology studies in *D. papillatum*. It will also serve as a reference for the highly diverse diplonemids and provide a point of comparison for studying gene complement evolution in the sister group of Kinetoplastida, including human-pathogenic taxa.

Keywords *Paradiplonema papillatum*, Protists, Genome, Transcriptome, Proteome, Gene-family evolution, Lateral gene transfer, CAZymes, Feeding strategy, Geographical distribution, Ecological distribution

*Correspondence:
Matus Valach
matus.a.valach@gmail.com
Gertraud Burger
gertraud.burger@umontreal.ca
Full list of author information is available at the end of the article

Background

Diplonemids are heterotrophic, flagellated, unicellular eukaryotes. Overlooked for decades, they have recently been characterized as the most species-rich group of known marine protists [1, 2]. Global metabarcoding surveys have estimated at least 67,000 species [3], revealing that diplonemids populate not only all biogeographic and pelagic zones of the oceans [4, 5], but also thrive in anoxic zones [3] and dominate deep-sea sediments [6]. Diplonemids inhabit fresh water as well, but in moderate abundance and diversity, suggesting recent habitat transitions [7].

Due to their abundance, distribution, and diversity, diplonemids are thought to play an important role in the marine food web. However, we have very little data regarding their nutrition. Views about their predominant feeding strategy are controversial, ranging from parasitism [6] to epibiosis of water plants and invertebrates, to predation of diverse algae including diatoms and dinoflagellates, to saprotrophy [8, 9]. In addition, a few diplonemid species seem to be bacterivorous [10, 11]. New research also indicates that diplonemids may significantly contribute to the cycling of certain heavy metals [12], though the actual extent and relevance for the marine ecosystem remains to be determined.

Diplonemids (Diplonemea) are subdivided into four monophyletic lineages, the classical diplonemids (Diplonemidae), hemistasiids (Hemistasiidae), and the Deep-sea pelagic diplonemid clades I and II (DSPDI and II) [13], the former now classified as Eupelagonemidae [14]. Currently, about nine diplonemid genera comprising nearly two dozen species are formally recognized and morphologically characterized [15]. However, axenic cultures have been established for only a handful of species that mostly belong to the Diplonemidae [16, 17], including the type species *Diplonema papillatum* (Fig. 1) (alternatively referred to as *Paradiplonema papillatum* [18]), whose genome is described here. From the Eupelagonemidae, the ecologically most prominent diplonemid group, just a few cells have been examined by microscopy and single-cell sequencing, while the vast majority of taxa is only known from environmental barcoding surveys [1, 19, 20].

In global eukaryotic phylogenies, diplonemids are placed together with euglenids, kinetoplastids, and symbiontids within the phylum Euglenozoa. Diplonemids form the sister group to kinetoplastids, which include the human pathogens *Trypanosoma* and *Leishmania* alongside several free-living taxa (e.g., *Bodo saltans*). Euglenozoa belong to the deeply diverging eukaryotic supergroup Discoba [21], which differs in essentially all aspects of biology from the familiar and best-studied eukaryotes—animals, fungi, and plants.

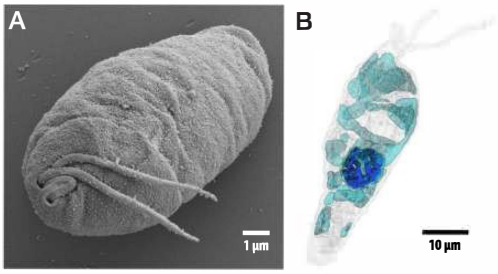


Fig. 1 Morphology and ultrastructure of *D. papillatum*. **A** Scanning electron microscopy image. The anterior end of the cell exposes the two flagella emanating from the flagellar pocket (left), the cytopharynx (right), and the conspicuous lip-like papillum between these two openings. Measurements taken from this image (using the ImageJ software): cell length > 10.5 µm (the exact size cannot be measured from this image because the cell does not lie flat); cell width ~ 5.9 µm; flagella length 6.5 µm; and cytostome width 0.31 µm. **B** Expansion microscopy-based model showing the nucleus and mitochondrion of a typical cell. Light gray, cell-surface tubulin; cyan, reticulated mitochondrion; blue, reticulated nuclear heterochromatin. The three-dimensional model was built from the Z-stack series of images after staining with DAPI and anti-tubulin antibodies. For details see Additional file 1: Sect. 1. Physical structure and size of the *Diplonema papillatum* nuclear genome

Diplonemids have also attracted interest because of their highly unusual mitochondrial genome, which has been most intensively investigated in *D. papillatum*. In contrast to conventional mitochondrial DNAs, that of *D. papillatum* not only makes up an overwhelming portion of cellular DNA [22], but also consists of hundreds of small chromosomes, each of which carries a single gene piece. Consequently, the assembly of full-length mRNAs and ribosomal RNAs (rRNAs) requires a range of RNA ligation events, which are accompanied by RNA editing [23, 24]. Studies across Diplonemidae and Hemistasiidae have revealed similar mitochondrial gene fragmentation and RNA editing features as seen in *D. papillatum*, reaching unprecedented degrees in certain species [25–27].

Establishing a diplonemid model system required not only a reasonably fast-growing axenic culture but also procedures to genetically modify the corresponding organism. In the past few years, we have developed protocols to transform *D. papillatum* with exogenous DNA [28] leading to homologous integration, which allows efficient knock-ins of tagged genes [29]. Now we have a powerful toolbox at hand for efficiently investigating the cellular and molecular biology of *D. papillatum*.

Available genome and transcriptome data from diplonemids are currently limited, with single gene and partial single-cell genome sequences generated for phylogenies [30, 31], biodiversity studies [19, 32, 33], or the investigation of metabolic adaptations in certain

diplonemids [34, 35]. While suited for the questions addressed in the corresponding studies, the data are too sparse to provide insight into the functions encoded in and the broader evolution of diplonemid nuclear genomes.

Here we present the nuclear genome and transcriptome sequence of *D. papillatum*. In addition to serving as a reference sequence for the diplonemids as a whole, our data provide insight into complex gene structures and expression modes. Analysis of the gene repertoire reveals a diverse metabolic potential of *D. papillatum*, but also, for a euglenozoan, unexpected conservatism of certain basic molecular machineries such as the cytosolic ribosome. Comparative genomics demonstrates that genes and pathways involved in carbohydrate degradation have undergone a major evolutionary expansion in diplonemids. The inferred metabolism, backed by feeding experiments, support the view that *D. papillatum* is extraordinarily versatile in using diverse carbon sources from myco-, phyto-, and zooplankton, as well as marine water plants. Taken together, our experiments and comparative genomic analyses strongly suggest that diplonemid protists play a crucial and previously unrecognized role in the food web of aquatic environments.

Results and discussion

Genome assembly, genome size, and ploidy

We generated ~ 900 million short paired-end reads (Illumina) and ~ 700,000 long reads (PacBio) summing to 187 Gbp. Reads were assembled into 6181 contigs ≥ 200 bp long totaling 280,293,864 bp, with an N50 value of 190,080 bp and a maximum contig length above 1 Mbp. The completeness of the assembly is estimated to be above 95%. Although BUSCO benchmarking [36] of conceptually translated *D. papillatum* proteins against the set of highly conserved eukaryotic core proteins recovered only 89%, we determined that more than half of the proteins reported as missing are too divergent to fall within BUSCO's inclusion threshold, and that a quarter are absent from all diplonemids examined (Additional file 1: Sect. 2. Assembly and annotation of the nuclear genome and transcriptome of *Diplonema papillatum*). We therefore consider the *Diplonema* genome assembly as *quasi* complete.

Genome assemblies are typically larger than the genome because of repeats. The actual nuclear genome size of *D. papillatum* was calculated by various methods. The estimate of 260 Mbp, based on k-mer frequencies in short reads, is deemed most accurate because this procedure is least affected by artifacts (Additional file 1: Sect. 3. The ploidy level of *Diplonema papillatum*). Assessment of the *Diplonema* nuclear genome size by pulsed-field gel electrophoresis was inconclusive as it yielded numerous unresolved molecules of length 1.1 to 1.8 Mbp, with only two distinct bands at 0.5 and 1.0 Mbp. It appears that the genome consists of hundreds of similarly sized chromosomes. A clear separation of individual chromosomes is probably impeded by the complex reticulated nuclear DNA structure observed by ultrathin-section and fluorescence microscopy of the *Diplonema* nucleus (Additional file 1: Sect. 1. Physical structure and size of the *Diplonema papillatum* nuclear genome).

The ploidy estimation of the *D. papillatum* nuclear genome is based on the frequency and distribution of k-mers in reads and single-nucleotide variants (SNVs) in the assembly. The extremely low SNV frequency (less than 600 sites in the 142 Mbp repeat-free genome portion) of the *Diplonema* genome and a symmetric, bell-shaped k-mer distribution frequency of short reads suggest haploidy (Additional file 1: Sect. 3. The ploidy level of *Diplonema papillatum*). While it is not possible to distinguish haploids from homozygous diploids (or higher ploidy levels) with computational methods alone, the most convincing confirmation for haploidy comes from gene replacement experiments, in which the transformation with engineered gene versions resulted reproducibly in single alleles [28, 29].

It should be noted that our assessment of haploidy refers to the standard laboratory *D. papillatum* strain, where the exclusive form of reproduction appears to be mitosis. Although sexual reproduction or a diploid stage has not yet been observed, the gene repertoire implies that *D. papillatum* has the potential to form diploid zygotes that undergo meiosis (Additional file 1: Sect. 11. Meiosis in *Diplonema papillatum*?).

Genome annotation and quality assessment

Genome annotation was performed by a pipeline developed in-house, combining gene model prediction with evidence-based and ab initio gene prediction (Additional file 1: Sect. 2. Assembly and annotation of the nuclear genome and transcriptome of *Diplonema papillatum*). Evidence-based prediction of protein-coding regions was guided by curated SwissProt sequences and unreviewed Discoba sequences available from public data repositories, as well as assembled *D. papillatum* transcripts. The start of the 5′ *U*n*T*ranslated *R*egion (UTR) was positioned at the site at which a Spliced Leader (SL) is added to pre-mRNA by trans-splicing—a feature shared by all Euglenozoa [37, 38].

The completeness and quality of automatically predicted protein-coding gene models were assessed by expert inspection of the three largest contigs in the assembly. Together these contigs represent 1% of the total assembly and contained initially 319 gene models. By scrutinizing the coverage of RNA-Seq reads and

transcripts assembled from these reads in the corresponding genome regions, we detected that 15 genes lacked corresponding models, 125 models had inaccurate gene structures, and 68 were false positives. Although the error rate appears high, it compares favorably with current automated annotations [39]. In *Diplonema*, most of the omitted or erroneous gene predictions are a consequence of the highly repetitive genome sequence in this organism as detailed below.

In sum, while the automated annotation procedure predicted ~ 37,000 protein-coding genes, the false positive and negative rates observed during manual curation indicate that their actual number in the *D. papillatum* nuclear genome assembly version 1.0 is rather ~ 32,000. *Diplonema*'s protein-coding genes contain on average 1.6 introns. Alternative splicing is estimated at 5% among multi-exonic genes, a proportion that is low compared to multicellular eukaryotes such as human (60%) or *Arabidopsis* (20%) [40], but in the range reported for other unicellular organisms [41].

Functional information was assigned to about 51% of the predicted protein-coding genes with an explicit molecular function available for about 35% of the models, and a conserved Pfam protein domain for an additional 15%. As in many other organisms, approximately 50% of the predicted protein-coding genes in *Diplonema* lack any indication as to their function.

Nuclear gene structure

About 41% of the protein-coding genes in the *D. papillatum* nuclear genome assembly contain introns, the large majority of which are canonical, bearing GT at their 5′-end and AG at the 3′-end (GT-AG type) (Additional file 1: Sect. 4. Intron splicing and structural RNAs). A few non-canonical introns with GC-AG splice-site combinations were detected as well. It was shown for GC-AG introns from animals, fungi, and plants that these introns are typically spliced by the same major U2 spliceosome as GT-AG introns [42]. The generally rare AT-AC (U12-type) spliceosomal introns seem to be absent from *Diplonema*, which is consistent with the lack of the U4atac, U6atac, U11, and U12 RNAs among the set of spliceosomal RNAs identified in this organism. Moreover, we did not detect unconventional introns such as the ones present in *Euglena* that lack conserved splice boundaries, have extensive base pairing to bring intron ends together, and are apparently removed in a spliceosome-independent fashion ([43] and references therein). Certain non-classified diplonemids reportedly possess *Euglena*-like introns; however, in the absence of transcriptome data, this inference cannot be validated [19].

While the median size of *Diplonema* introns is below 1 kbp, a small percentage are considerably longer, often comprising complete or partial transposable elements with several open reading frames (ORFs) (see following section). The longest expert-validated intron is 72 kbp in size and resides in the gene DIPPA_22195, predicted to encode a protein with a conserved kinesin-motor domain. This is the largest known *Diplonema* gene (186 kbp), containing the highest number (20) of introns as confirmed by expert validation. It is noteworthy that several genes with confirmed alternative splicing combine more than one splicing mode. For example, the expression of DIPPA_03285 involves occasional exon skipping, intron retention, and alternative splice-site selection. The corresponding protein sequence has moderate similarity with the Pfam domain TFIIα (Transcription initiation factor II alpha) and a common structural domain called PDZ found in numerous cell-signaling proteins.

In all domains of life, the coding regions of genes are usually bounded by untranslated regions. At ~ 70 bp, the 5′-UTRs of *Diplonema* nuclear protein-coding genes are within the size range commonly observed across eukaryotes (Additional file 1: Sect. 5. Untranslated regions of nuclear genes). In contrast, the observed 3′-UTRs are exceptionally large; they sometimes extend up to several kbp and have a median size of ~ 800 bp, which is two to ten times longer than in other eukaryotes. In diplonemid's sister group, the kinetoplastids, the 3′-UTR gene region is known to play a predominant role in the regulation of gene expression, in particular by controlling mRNA translation and decay rates [44]. Therefore, the long 3′-UTRs of *D. papillatum* genes may serve as a binding platform for numerous regulatory proteins. It would be worthwhile to investigate the identity of these postulated RNA-binding proteins experimentally, with those *Diplonema* genes possessing the longest 3′-UTRs presenting the most obvious first targets.

During expert validation of the structural annotation of the assembly (Additional file 1: Sect. 2. Assembly and annotation of the nuclear genome and transcriptome of *Diplonema papillatum*), we identified dozens of gene models with adjacent sequence repeats. In many of these instances, a portion of the gene's 5′-region, including parts of the 5′-UTR and coding sequence (CDS), is repeated in tandem. In other cases, the 3′-end of the first exon is repeated, forming a part of the first intron or—if the gene consists of a single exon—the 3′-UTR. The longest repeated gene extension was detected upstream of DIPPA_19968 encoding an ABC transporter. Here, a ~ 400-bp long sequence motif composed of a part of the gene's 5′-UTR and the preceding intergenic region occurs in 12 tandemly arranged, degenerate copies, constituting a tandem array of nearly 5 kbp. PacBio reads support

the assembly in this genome region, and RNA-Seq-read coverage indicates that the tandem array is not part of the mature mRNA (Fig. 2). Obviously, repeats adjacent to genes interfere with automated structural annotation, because RNA-Seq-reads can be aligned to multiple locations, occasionally resulting in gene models that are too long or include spurious introns.

Non-coding regions and repeats

Nuclear genomes of free-living Discoba, for which near-complete assemblies are available, are all below 50 Mbp in size, with 20 Mbp for *Andalucia godoyi* (Jakobida) [45], 40 Mbp for *B. saltans* (Kinetoplastida) [46], and 41 Mbp for *Naegleria gruberi* (Heterolobosea) [47]. With the exception of the *E. gracilis* genome, estimated at 330–500 Mbp [48], the *D. papillatum* nuclear genome is 6–13 times larger than those known from other discobids. Introns contribute to some extent to this genome size difference, but the additional material mainly comprises repeat regions—mostly dispersed repeats—which make up 52% of the assembly (Additional file 1: Sect. 6. Repetitive sequences in the nuclear genome of *Diplonema papillatum* (assembly v_1.0); Additional file 2). Among the nearly 10,000 distinct dispersed repeat motifs, the most abundant one occurs about 6000 times when considering copies of ≥ 90% sequence identity. The longest, nearly 20-kbp dispersed repeat motif, which is found 13 times in the genome, is particularly notable because it is itself composed of an array of approximately eight 2.5 kbp-long motifs arranged in tandem. Each of these tandem repeats contains a 283-amino acid-long ORF that is apparently not transcribed, nor does it share similarity to proteins or conserved domains in public databases.

While many dispersed repeat units in the *Diplonema* nuclear genome have no obvious origin or function, others derive from transposable elements and encode proteins known from retrotransposons and DNA transposons described for a wide range of eukaryotes. The *Diplonema* nuclear genome assembly includes as many as ~ 2500 gene models annotated as retrovirus-related polyproteins, LINE-1 (Long interspersed nuclear element) ORFs, SLACS (Spliced-leader-associated-conserved sequence) reverse transcriptases, and DNA-directed RNA polymerase from mobile element R2 and jockey. In addition, more than 60 ORFs from *Diplonema* resemble proteins residing in DNA transposons including TATE (Telomere associated transposable element), MULE (mutator-like element), and Helitron [49]. A rigorous identification of transposable elements including non-coding regions will be warranted once a chromosome-scale genome assembly becomes available to eliminate artificially duplicated or collapsed repeat regions.

As expected, certain dispersed repeat units contain regular genes, one of which is the ribosomal DNA (rDNA) cluster that is composed of 18S, 5.8S, and 28S rRNA genes. The *Diplonema* nuclear genome assembly contains 13 copies of this cluster at ≥ 99% sequence identity. In addition to dispersed copies of complete genes, we also found multiple copies of gene fragments. For

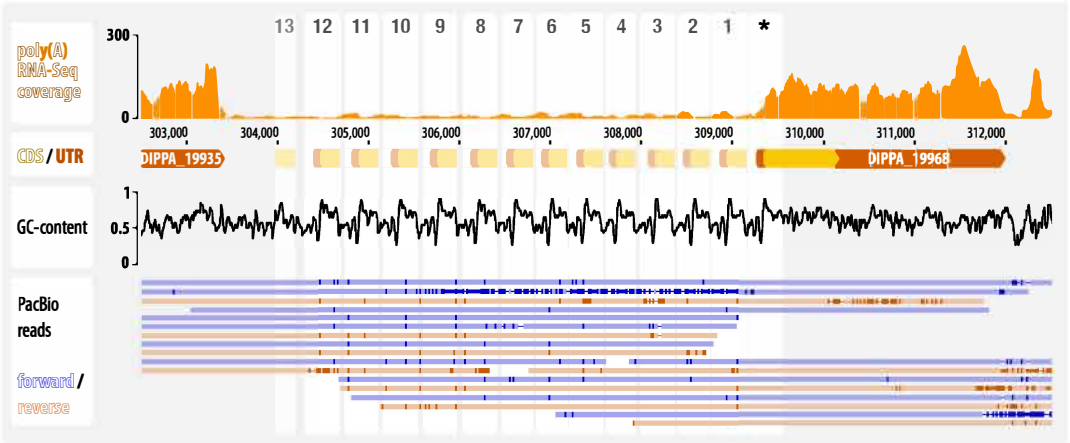


Fig. 2 Repeat-bounded gene structures in the *D. papillatum* nuclear genome. An example of a gene with a terminal region repeated multiple times adjacent to the expressed portion of the gene. DIPPA_19968 encodes a SufC homolog, a protein involved in iron-sulfur cluster synthesis. The 5′-terminal segment, including the 5′-UTR and part of CDS, is repeated 13 times upstream of the transcribed gene portion. Copies #1–#12 display 71–96% sequence identity, while the most distal repeat has only 47%. Middle pane: the G + C content plot reflects the repetitiveness of the region. Lower panel: long reads covering this region confirm the correctness of the genome assembly in this region

example, about 3600 28S rRNA gene pieces (up to 10% in length of the complete gene) are scattered throughout the genome. The nuclear genomes of human and other eukaryotes carry similar repeats that are referred to as terminal-repeat retrotransposons in miniature (TRIMs) and short interspersed elements (SINEs), but which contain gene portions of 5S rRNA and 28S rRNA [49].

Another source of extra sequence in the *Diplonema* nuclear genome are nuclear mitochondrial segments (NUMTs), i.e., portions of the mitochondrial genome [23, 24, 50] incorporated into the nuclear DNA, and which make up at least 343 kbp (1.2%) of the assembly (Additional file 1: Sect. 6. Repetitive sequences in the nuclear genome of *Diplonema papillatum* (assembly v_1.0)). We detected more than 1400 NUMTs (>100 bp), including 11 complete mitochondrial chromosomes. NUMTs are inserted predominantly in intergenic regions, but ~20% occur in introns and UTRs of nuclear genes (Additional file 2). Nearly 2% of NUMTs are arranged in tandem. The longest array of nearly 10 kbp consists of 164 copies of a 68-bp stretch from the B-class constant region of mitochondrial chromosomes. The total length and proportion that NUMTs contribute to the *Diplonema* nuclear genome compares with the situation in animals and plants [51–53].

Transcription, transcript maturation, and regulation of gene expression

Protein-coding genes in the *D. papillatum* genome assembly are conspicuously arranged in clusters with genes sharing the same transcriptional orientation (Additional file 1: Sect. 7. Polycistronic transcription units in the nuclear genome of *Diplonema papillatum*). Nearly 90% of all contigs larger than 50 kbp include unidirectional arrays of five up to 120 genes. The longest expert-validated gene cluster (in the contig tig00022654_12, which is 1,009,103 bp long) comprises as many as 108 genes. Inside clusters, genes are not particularly tightly spaced. For example, in tig00022654, several intergenic regions are longer than 10 kbp (Additional file 3). This gene arrangement is reminiscent of trypanosomes, where arrays of about 100 unidirectional genes are co-transcribed into several long primary polycistronic RNAs [54].

As already mentioned above, mRNAs of *Diplonema* and other Euglenozoa carry a spliced-leader (SL) sequence extension at their 5′-terminus that is encoded by a separate gene, transcribed independently and added by trans-splicing to pre-mRNAs [55]. Extrapolating from the set of expert-validated genes, essentially all mRNAs in *D. papillatum* carry an SL at their 5′ terminus (Additional file 1: Sect. 2. Assembly and annotation of the nuclear genome and transcriptome of *Diplonema papillatum*), strongly suggesting that the maturation of clustered genes proceeds as in kinetoplastids, involving the processing of long polycistronic RNAs to monocistronic units along with the posttranscriptional addition of an SL to the 5′-end [56].

A predominant co-transcription of *D. papillatum* nuclear genes implies that in contrast to most other eukaryotic groups, gene expression—probably in euglenozoans as a whole—is not primarily regulated by transcription initiation. Our finding of genes involved in DNA modification and transcript degradation points to alternative, gene-specific control mechanisms acting in the *Diplonema* nucleus. First, *D. papillatum* has the potential for synthesizing nucleobase J and 5mC, both reported to play an important role in gene regulation of model organisms (Additional file 1: Sect. 8. DNA modifications (5mC and J)). Base J (β-D-glucopyranosyl-oxymethyluracil) is a hyper-modified thymine derivative, which was detected early on in the nuclear DNA of Euglenozoa [57]. Its role in transcription termination has been demonstrated in trypanosomes [58]. The *D. papillatum* genome encodes counterparts of all proteins participating in the biosynthesis and proliferation of this nucleotide modification.

Similarly, we identified homologs of DNA methyltransferase genes known to synthesize 5-methyldeoxycytosine (*5mC*) in the *Diplonema* genome. This epigenetic mark mediates transcriptional repression, particularly of transposons and other repetitive elements in nuclear genomes of animals, plants, and fungi [59]. The presence of a dozen homologs of AlkB-type genes encoding oxidative demethylases in the *Diplonema* genome indicates that this organism uses methylation/demethylation to dynamically regulate gene expression.

Further, *D. papillatum* has the potential for RNA interference (RNAi) (Additional file 1: Sect. 9. RNA interference (RNAi)). We retrieved from the inferred proteome homologs of all components required for a functional RNAi pathway, two Dicer-like proteins, five Piwi proteins, three members of the Argonaute family, and one RNA-dependent RNA polymerase. In model organisms, RNAi has been shown to control RNA degradation and translational silencing of transposable elements and genuine nuclear genes [60, 61]. Key determinants of the RNAi machinery are also encoded in the nuclear genome of *Euglena gracilis* [62], but are incomplete or missing in many (but not all) kinetoplastid taxa [63].

Figure 3 summarizes the current knowledge of the shared and particular features related to gene expression as well as genome architecture across the euglenozoans.

Structural RNA genes

The *D. papillatum* nuclear genome contains a total of ~37,000 genes, of which about 1000 encode structural

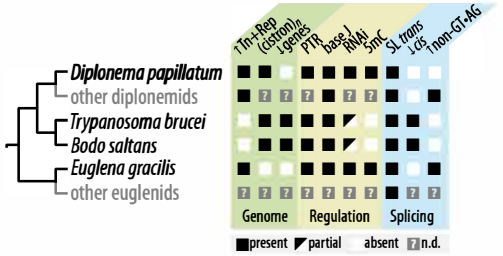


Fig. 3 Comparison of genome and gene expression features across Euglenozoa. ↑Tn + Rep, conspicuous abundance of transposons and repetitive sequences; (cistron)$_n$, polycistronic transcription; ↓genes, streamlined gene repertoire; PTR, posttranscriptional regulation of gene expression; base J, base J present in nuclear DNA; RNAi, RNA interference pathway; 5mC, 5-methyldeoxycytosine pathway; SL *trans*, spliced-leader *trans*-splicing; ↓*cis*, few cases of *cis*-splicing; ↑non-GT•AG, conspicuous abundance of unconventional *cis*-introns. n.d., not determined

RNAs, also referred to as non-(protein-)coding RNAs (ncRNAs).

Four rRNA species are associated with the cytosolic (cyto) ribosomes of *D. papillatum*, notably 28S, 5.8S, and 5S-sized rRNAs in the large subunit (LSU), and 18S rRNA in the small subunit (SSU) (Additional file 1: Sect. 4. Intron splicing and structural RNAs). We confirmed this number experimentally, because it is atypically low for cytoribosomes from euglenozoans. For example, the 28S rRNA is split into six or more pieces across kinetoplastids and euglenids [64, 65]. Despite the difference in the number of rRNA species, *Diplonema* cytoribosomes contain the same complement of canonical cytoribosomal proteins as kinetoplastids and euglenids, as well as one SSU ribosomal protein apparently unique to euglenozoans (Additional file 1: Sect. 10. The cytosolic ribosome of *Diplonema papillatum*). As in other eukaryotes, the *Diplonema* genome carries three of the rRNAs (28S, 5.8S, and 18S rRNAs) organized in a classical rDNA tandem unit. The genome assembly includes more than 20 of these ~ 7.6 kbp-long rDNA units arranged in clusters, but the actual length of these repeat arrays is not known. The 5S rRNA is not included in the rDNA repeat unit but rather in a separate repeat unit together with SL RNA, as discussed below.

The nuclear genome assembly contains a set of 211 high-scoring transfer RNA (tRNA) genes comprising up to 10 identical copies (tRNA-Lys$_{CUU}$), with some carrying an intron in the anticodon loop. Collectively, this ensemble of tRNA genes represents 46 out of 64 possible anticodons. Among the missing anticodons are all those reported absent from eukaryotes in general [66], but one missing specifically from *Diplonema* is that of tRNA-Leu$_{UAA}$ for decoding TTA codons. While this leucine codon is the one used most infrequently (2%) in protein-coding regions of *D. papillatum*, it does occur in vital genes. Assuming posttranscriptional modification of bases in the anticodon, UUA could be read by the anticodon of tRNA-Leu$_{CAA}$ after conversion of the wobble cytosine to uracil or, alternatively, by the tRNA-Leu$_{AAG}$ anticodon after deamination of adenine-34 to inosine. The genes from *Diplonema* that could catalyze such base-modification activities are the homologs of ADAT2 and ADAT3 encoding the two-component A-to-I tRNA editing enzyme known from other eukaryotes [67]. It is possible that in *D. papillatum*, the ADAT enzymes perform not only A-to-I but also C-to-U editing, since certain adenine deaminases have a relaxed nucleotide specificity [68].

We detected and manually validated the genes specifying five types of spliceosomal RNAs, namely U1, U2, U4, U5, and U6 small nuclear RNA (snRNA). U2 RNA occurs as often as 163 times in the assembly, with 151 identical copies. Most U2 RNA genes are part of a repeat region in which they alternate with the genes for 5S RNA and SL RNA, up to 27 times in a row.

Finally, we validated the predicted SL RNA genes of *Diplonema*, which are composed of a 39-bp long 5′-exon (the SL) and a 75-nt long intron. This gene occurs in 110 copies (at ≥ 90% sequence identity), forming a tandem repeat unit together with the 5S rRNA and U2 snRNA genes (Additional file 1: Sect. 6. Repetitive sequences in the nuclear genome of *Diplonema papillatum* (assembly v_1.0)). Among the eukaryotes possessing SL RNA, the gene is often part of a tandem repeat unit and associated with the 5S rRNA gene (e.g., in some animal and dinoflagellate groups, euglenids, and kinetoplastids [69, 70]. However, a repeat unit consisting of three alternating ncRNA genes as in *D. papillatum* (SL RNA–5S rRNA–U2 snRNA) is exceptional and also seems to be absent in other diplonemids. (For more details on structural RNAs, see Additional file 1: Sect. 4. Intron splicing and structural RNAs).

Genes involved in the general cellular metabolism

Among the ~ 37,000 *D. papillatum* protein-coding genes, at least 15% are predicted to be involved in metabolism. Biochemical studies of metabolic processes in diplonemids have investigated glycolysis and gluconeogenesis [35, 71], carbon storage [72], respiration [73], and free-radical detoxification [74]. In addition, recent in silico transcriptome analyses have provided an overview of basic metabolic pathways such as fatty acid synthesis and degradation, pyruvate metabolism, and pentose phosphate pathway across diplonemids [34] and in *D.*

papillatum specifically [75]. The limitation of transcriptome-based studies is that the data may include unrecognized contamination with mRNAs from other organisms or lack reads from genes poorly expressed under the examined conditions. Still, the metabolism of diplonemids inferred from the transcriptomes is overall in agreement with that inferred from the nuclear genome sequence presented here. In the following section, we will focus on the polycarbohydrate metabolism of *D. papillatum*, an aspect neglected in earlier work and, as we will show, one with important bearings on the ecological role of this protist in the marine environment.

Gene complement participating in polycarbohydrate metabolism of *D. papillatum* and other euglenozoans

Enzymes involved in the synthesis and degradation of polysaccharides (referred to here as Carbohydrate-active enzymes (CAZymes)) currently comprise ~350 distinct catalytic families and about 90 non-catalytic families (Additional file 1: Sect. 12. CAZyme-coding genes in *Diplonema papillatum*). The nuclear genome of *D. papillatum* encodes nearly 500 CAZymes from 52 families for metabolizing diverse polysaccharides (Fig. 4A). By far the most diverse and largest enzyme group is involved in the degradation of pectin (a heteropolysaccharide consisting mostly of methyl-esterified α-D-1,4-galacturonic acid units), with about 120 genes from nine distinct CAZyme families. The second largest group consists of 82 proteins that belong to CAZyme families breaking down the β-1,4-linked glucose polymers cellulose or hemicelluloses. Further, we retrieved 27 homologs of enzymes degrading sulfated glucuronomannan (α-1,3-mannan with β-D-glucuronic acid side chains), which is the main polysaccharide component in the cell wall of diatoms [76]. In addition, the presence of certain glycoside hydrolase homologs in the genome assembly suggests that *Diplonema* is most likely able to digest the β-1,3-glucan laminarin, which is the storage polysaccharide of numerous micro- and macroalgae [77]. Laminarin plays a major role in the marine carbon cycle representing ~10% of the carbon produced globally [78]. The *D. papillatum* genome assembly also revealed 18 genes which, in model organisms, were shown to break down chitin (polymer of N-acetylglucosamine) and glycosaminoglycans, both extracellular polysaccharides of animals and fungi. Finally, the *D. papillatum* genome encodes 90 CAZyme genes whose substrate cannot be inferred with confidence. Some of these genes might be involved in the breakdown of complex glycans such as the transparent exopolymer particles (TEPs) secreted by diverse marine eukaryotes [79].

CAZyme genes that are conspicuously lacking in the *D. papillatum* genome assembly are homologs of enzymes degrading bacterial cell-wall components, a finding that is corroborated by feeding experiments [11]. The seemingly strictly eukaryotic diet of *D. papillatum* contrasts with the food preference of, e.g., the diplonemid *Rhynchopus euleeides*, to our knowledge the first reported bacterivorous diplonemid [80]. Also missing in the *D. papillatum* genome are genes encoding poly-α-D-1,4-glucose-depolymerizing enzymes, which might suggest that this organism is unable to digest starch and glycogen, the carbon-storage compounds of Viridiplantae and Metazoa, respectively. However, the latter inference contradicts the results of our feeding experiments (see further below), which demonstrate that *Diplonema* readily utilizes amylopectin, the predominant constituent in starch (see Additional file 1: Sect. 13. Glycan and peptide assimilation by *Diplonema papillatum*). The gene(s) responsible for amylopectin degradation may be among the functionally unassigned CAZymes mentioned above or belong to novel families. Interestingly, polyglycan-degrading enzymes are one of the largest CAZyme class predicted to be secreted outside the *Diplonema* cell (Additional file 1: Sect. 14. Secretome prediction),

(See figure on next page.)

Fig. 4 Polycarbohydrate metabolism in *D. papillatum*. **A** Proteins containing at least one CAZyme domain. Proteins were grouped by their cognate substrate class. The subdivision of the bars by different color shades represents the number of enzymes in the following subgroups. *pectin*: pectin hydrolases, pectin lyases, pectin acetylesterases, and pectin methylesterases. *cellulose & related*: cellulases, xylan-α-glucuronidases, xylan/cellulose and xylan/xyloglucan hydrolases, hemicellulases, β-glucan/β-xylan hydrolases, and β-mannanases. *β-1,3-glucan*: no subgroups. *glucuronomannan*: α-mannanases and β-glucuronidases. *chitin & related*: chitinases, glycosaminoglycan and glucosamine hydrolases. *N-glycan*: no subgroups. *oligomers and others*: α-glycosidases, β-glycosidases, trehalases, an α-fucanase, and an invertase. **B** Distribution of the CAZyme families GH (glycoside hydrolase), PL (polysaccharide lyase), CE (carbohydrate esterase), and CBM (carbohydrate-binding module) across four free-living euglenozoans (*D. papillatum*, *D. japonicum*, *B. saltans*, and *E. gracilis*) and two representative fungi (*Trichoderma reesei* and *Phanerochaete chrysosporium*). Rows correspond to individual CAZyme families with heatmap shading indicating the number of family members in each genome as detailed in the key (bottom). **C** DlPPA_15749, a GH115-family member and its 12 paralogs, were most likely acquired specifically by *D. papillatum* via horizontal transfer from diverse bacteria. Sequences that belong to bacteria, fungi, and diplonemids other than *D. papillatum* are highlighted in shades of yellow, beige, and light blue, respectively. For details, see Additional file 1: Sect. 15. Genes horizontally transferred from bacteria to *Diplonema papillatum*. **D** Biomass of *D. papillatum* cells grown in various substrates. The cell sizes are represented as circles of different diameter and the predominance of the various sizes by the grey shade of the fill. Cells were counted in triplicate after six days and weighed to calculate their biomass (wet weight per 10^8 cells). Bars indicate the mean deviation of the cell counts for each of the four independent biological replicates. Note that the predominant cell size correlates with both biomass and the number of granules

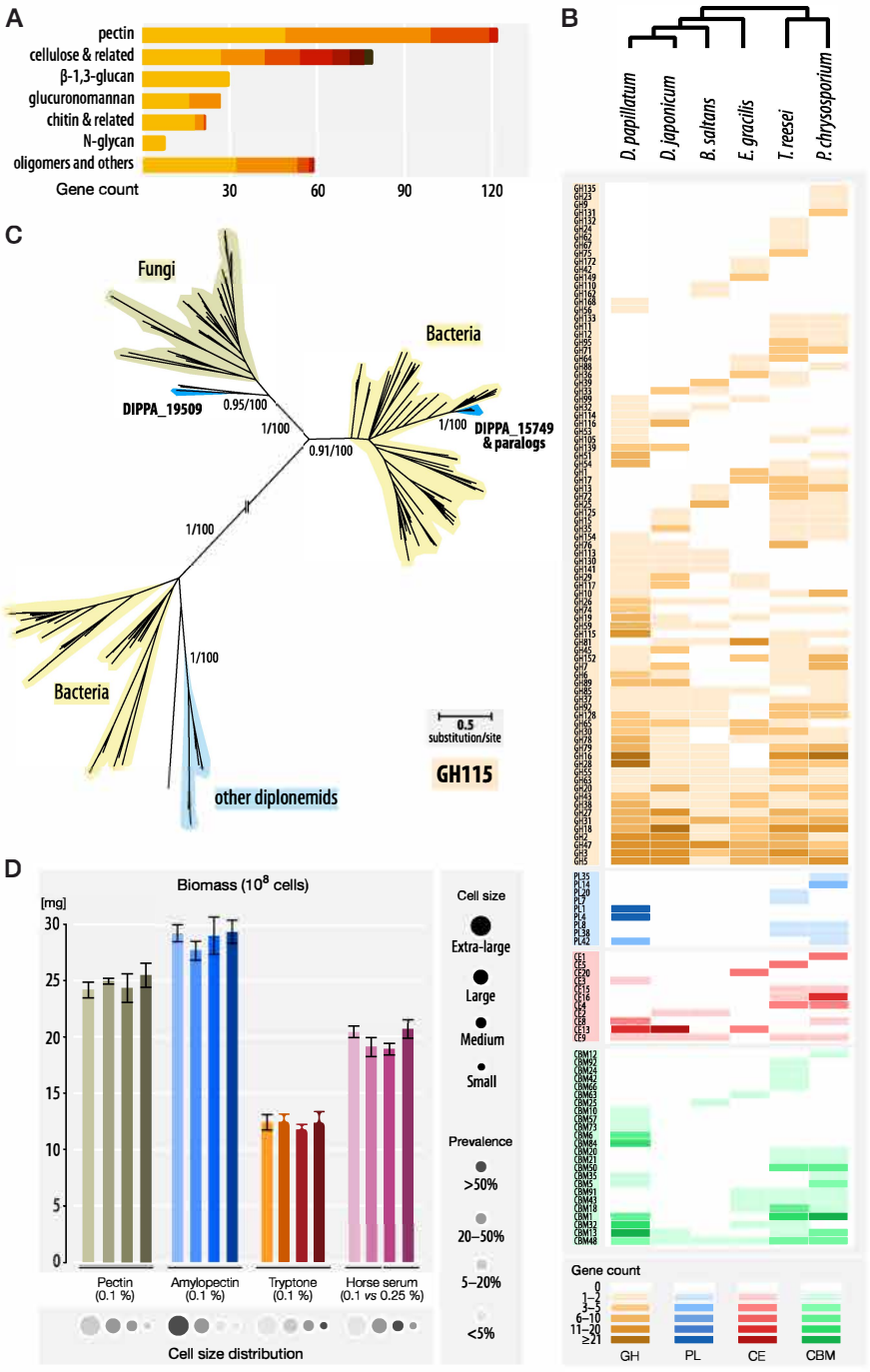


Fig. 4 (See legend on previous page.)

indicating an important role of this activity for the feeding strategy of this microeukaryote (see below).

In the search for *Diplonema* genes involved in the formation of carbon storage, we identified homologs of β-1,3-glucanase indicating the synthesis of paramylon, a polysaccharide long known from *Euglena* and recently identified experimentally also in *D. papillatum* [72]. As not only *Euglena* and *Diplonema* but also *Bodo* store their carbon in that form [81], paramylon was probably already synthesized by the last common ancestor of Euglenozoa.

Among the examined euglenozoans—i.e., *D. papillatum*, its closest described relative *D. japonicum* [8], the free-living kinetoplastid *B. saltans* [46], and the recently sequenced euglenid *E. gracilis* [48]—it is *D. papillatum* that carries the largest complement and diversity of CAZyme families. The carbohydrate-degrading enzymes (GH, PL, CE) and carbohydrate-binding modules (CBM) are particularly expanded in the diplonemid type species (Fig. 4B; Additional file 1: Sect. 12. CAZyme-coding genes in *Diplonema papillatum*, Additional file 4). Most notably, none of the other euglenozoans appears to possess nearly as many enzymes for pectin and β-1,3-glucan degradation (only 2–33% of the *D. papillatum* numbers). The exceptionally large repertoire of CAZyme genes in *D. papillatum* is comparable to that of saprophytic fungi and should allow this protist to feed on a multitude of algal and plant species occurring in the natural marine habitat of diplonemids [82]. Furthermore, the striking differences in CAZyme complement between the two closely related diplonemids that we examined provide a new window not only into the dynamic nature of diplonemid gene repertoires, but also an opportunity to begin to understand how the gene content impacts the varying lifestyles of diplonemids in general [18].

Horizontal gene transfer in *D. papillatum*

An important factor leading to differences in gene complements between closely related species is acquisition of genes by horizontal transfer (HGT). As bacteria-to-eukaryote gene transfers appear to be particularly frequent in marine ecosystems [83], we searched for similar signs of such HGTs in *D. papillatum*.

Genes that were likely acquired from bacteria by HGT (referred hereafter to as "HGT genes") were identified by best reciprocal blast hits against NCBI nr and a set of custom proteomes representing all domains of life, followed by phylogenetic inference and selection of well-supported tree topologies. Validation of candidate HGT genes included visual inspection of trees to assure that the *Diplonema* protein is nested within a bacterial clade. We also verified that the corresponding gene resides on a contig that also encodes typical, presumably endogenous nuclear genes and that the transcript carries an SL, which provides an extra layer of confidence that the gene is indeed expressed (Additional file 1: Sect. 15. Genes horizontally transferred from bacteria to *Diplonema papillatum*; Additional file 5).

The *D. papillatum* nuclear genome assembly includes at least 96 genes likely acquired horizontally from bacteria. These HGT genes form 56 families with up to 14 members; all are transcribed. Ten families have multiple members, with some expansion being a result of tandem gene duplication into up to six copies. Two out of the three largest gene families play a role in the detoxification of reactive oxygen species, but the majority of families participate in metabolic pathways. Four HGT families with a total of 17 members are predicted to be CAZymes, which apparently were acquired specifically by *D. papillatum* because they are not detected in the transcriptomes of the 10 other diplonemids examined (Fig. 4B). The largest HGT-CAZyme family (expanded to 12 members) encodes xylan-α-glucuronidases of the glycoside hydrolase family GH115, which comprise enzymes that break down hemicelluloses. Phylogenetic analysis places these *D. papillatum* proteins as a sister clade to Planctomycetes, Bacteroidetes, Gammaproteobacteria, or Verrucomicrobia, reflecting highly diverse donors as well as potential HGT among bacteria themselves (Fig. 4C).

As observed in other systems, most genes transferred from bacteria to eukaryotes expand or rewire the metabolic capabilities of the recipient [84]. Similarly to what has been documented in other eukaryotes (e.g., [85, 86]), in *Diplonema*, genes encoding CAZymes represent one of the most frequently horizontally acquired functional categories.

Gene-family evolution in *D. papillatum* and other diplonemids

In addition to gene acquisitions and losses, the *Diplonema* genome is also shaped by gene duplications followed by sequence divergence of copies, leading to multi-gene families that grow or shrink over time.

To investigate the gene-family evolution of the *Diplonema* genome, we established a proteome dataset comprising 30 eukaryotes that include *D. papillatum* and three other diplonemids, eight additional euglenozoans, and 12 eukaryotic species from major groups outside Euglenozoa. We used OrthoFinder 2 [87] to infer gene families and orthologous groups for subsequent analysis. OrthoFinder retrieved nearly 200 orthologs with representation in 25 or more taxa, which we concatenated in order to infer a species tree using the LG + C60 + F substitution model, the best-fitting model as determined by the BIC criterion in IQ-TREE [88]. The resulting tree furnished a well-resolved backbone phylogeny for the

euglenozoan clade (Fig. 5). The complete set of gene families was then used to estimate protein-family expansions, contractions, gains, and losses by using the phylogenetic birth–death model implemented in Count [89] (Additional file 1: Sect. 16. Evolution of gene families; Additional file 6).

Within the euglenozoan clade, gene family gains and expansions are much more frequent than losses and contractions. The highest count of gene-family gains in the entire tree is the ancestral diplonemid node, indicating a substantial diversification of the gene repertoire in the last diplonemid common ancestor (LDCA, Fig. 5). Similarly prolific is the expansion of protein families at that node. In addition to CAZymes and cytoribosomal proteins discussed in detail in previous sections, highly expanded families act in signal transduction, with a number of predicted protein kinases comparable to that in human [90] and plants [91]. Gene families involved in amino acid metabolism also expanded at the diplonemid node. A noteworthy finding is the gain of glycine amidination and methylation genes, which indicates that diplonemids are capable of converting glycine into creatine, a scarce compound in marine environments that is otherwise only supplied by metazoan and diatom excretion [92]. The gene families that expanded specifically in *D. papillatum* but not in the other diplonemids are involved in oxidative stress protection, including two families acquired horizontally from bacteria. This expansion might be an adaptation to life in the surface seawater layer penetrated by solar radiation that triggers the production of cytotoxic reactive oxygen species (ROS), an adaptation likely protecting also from ROS generated by man-made pollutants such as metals, polychlorinated biphenyls, and radioisotopes in coastal waters.

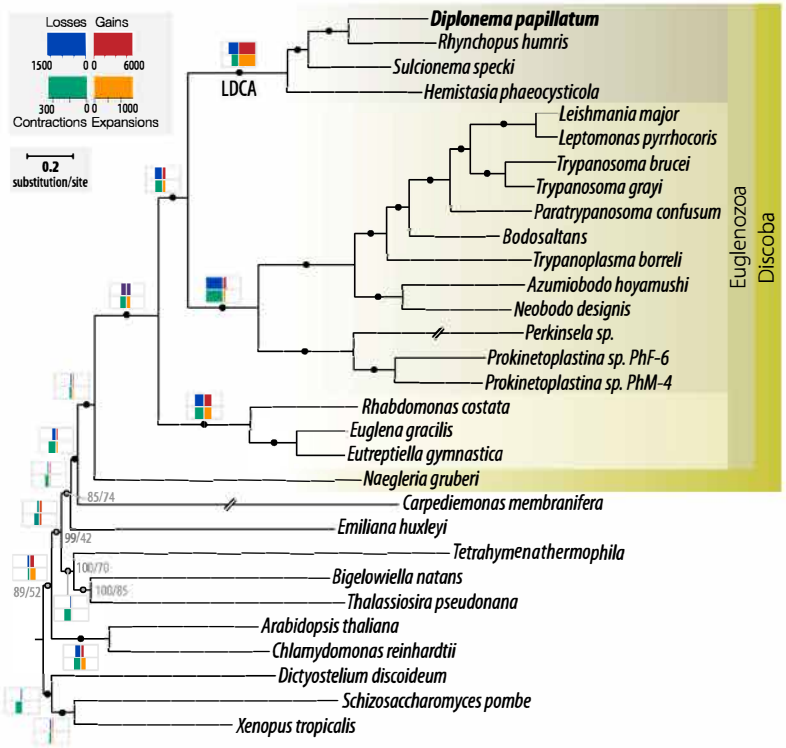


Fig. 5 Gene-family evolution in Euglenozoa. A maximum-likelihood phylogenetic tree based on the concatenated alignment of 167 proteins containing 57,565 amino acid positions. Nodes with maximal statistical support are indicated with black circles, for the remaining nodes the supports are in grey in the following format: bootstrap support/SH-aLRT value. Double-crossed branches were reduced to half of their original length. The black horizontal bar indicates the number of substitutions per site. The number of gene families lost, gained, expanded, and contracted at selected nodes (based on the sum of probabilities of the respective events at each node/tip) is indicated by the width of blue, red, orange, and green boxes, respectively. Key, event scales. Note the exceptionally large gain and expansion of more than 7000 and 1400 gene families, respectively, on the ancestral diplonemid branch. Diplonemid, kinetoplastid, and euglenid taxa are highlighted in beige background shades. LDCA, last diplonemid common ancestor

Carbon nutrition

Earlier experimental studies showed that *D. papillatum* does not import glucose in any significant amount from the medium, but instead readily takes up and metabolizes amino acids [35]. The authors concluded from this observation that in its natural habitat, the primary energy source of *D. papillatum* is not carbohydrates as is the case in most heterotrophic eukaryotes, but rather poly- and oligo-peptides. However, these earlier inferences about *Diplonema*'s nutrition are in conflict with our finding described here of a large ensemble of highly transcribed carbohydrate-metabolizing genes in the inferred proteome and secretome.

Certain diplonemids (though not *D. papillatum*) have been observed to feed on microalgae and decaying water plants (e.g., [8, 82, 93]), strongly suggesting that in its natural habitat *D. papillatum* uses its large CAZyme arsenal to break down cell-wall components of diverse prey (Additional file 1: Sect. 12. CAZyme-coding genes in *Diplonema papillatum*). The question arises whether cell-wall degradation serves *D. papillatum* solely for gaining access to proteins inside the prey's cell (referred to as protoplast feeding such as described recently for an amoeba [94]), or rather for assimilating the carbohydrates in the cell wall and/or intracellular storage glycans from starch to laminarins.

To test the hypothesis that *D. papillatum* is indeed able to assimilate cell-wall and storage glycans, we performed growth experiments in media of various compositions (Additional file 1: Sect. 13. Glycan and peptide assimilation by *Diplonema papillatum*). In agreement with previous studies [35, 72], our results confirm that this protist only poorly utilizes glucose as sole carbon source. However, our data also indicate that it does grow well on polycarbohydrates such as pectin and particularly amylopectin. Most importantly, *D. papillatum* utilizes carbohydrates as efficiently as peptides (Fig. 4D). Together with the identification of numerous CAZyme homologs and carbohydrate-transporter genes in the *D. papillatum* genome, we conclude that this organism degrades extracellular polysaccharides obtained from marine prey and imports oligomers into the cell for assimilation (Additional file 1: Sect. 17. Feeding strategy and food of *Diplonema papillatum*).

In addition, our findings question the view that *D. papillatum* is an exclusive osmotroph in marine environments, scavenging on debris of dead organisms. The results presented here suggest rather that in the wild, *D. papillatum* feeds mostly on living eukaryotes. We posit that this protist enzymatically pierces and ruptures live prey cells and then engulfs cell-wall particles and cytoplasm alike. This feeding strategy would allow *D. papillatum* to forage on living eukaryotes from a broad taxonomic range and of any physical size, from diatoms to dinoflagellates, macroalgae, and aquatic plants.

Fig. 6 Oceanic distribution of *D. papillatum*. The world map shows the distribution of sampling locations from the three datasets in which *D. papillatum* was detected, namely the Tara project, "Ocean Sampling Day 2014," and "Helgoland Roads 2016" (beige dots). Sites at which *D. papillatum*-representing OTUs (operational taxonomic units) were detected are highlighted in blue. For details see the main text and Additional file 7)

Environmental distribution of *D. papillatum*

Diplonemids are essentially omnipresent in marine environments, found from the tropics to the poles, in the top layer down to abyssal/hadal zones (> 6000 m below the surface), and in both pelagic (planktonic) and benthic (sediment) habitats [3]. We have a good understanding of the environmental and geographical distribution of the major diplonemid groups but not of the type species. Therefore, we searched available datasets of the V9 [4, 95–97] and V4 [96–99] hypervariable regions of the 18S rDNA for the presence of *D. papillatum* sequences (Additional file 1: Sect. 18. Environmental distribution of *Diplonema papillatum*). We detected no *D. papillatum* reads in datasets from samples collected in the open ocean, or from waters below ~ 10 m depth and 6 °C temperature. Instead, signature sequences of the type species were present in datasets of temperate coastal regions, from Helgoland to Japan, and the Americas (Fig. 6).

To summarize, the abundance of the type species and many other Diplonemidae members in marine habitats is relatively low. *D. papillatum* occurs sporadically in temperate coastal surface waters of the world ocean. We suggest that it preferentially populates coastal regions because they are more eutrophic than the open sea and thus much richer in plants and algae, the postulated major food sources of this organism.

Conclusions

Our analysis of the *D. papillatum* gene complement has provided insights into the previously unknown, central role of polysaccharide degradation in this organism, allowing inferences about its ecological role. However, these insights are not necessarily transferrable to other

diplonemids, because their CAZyme complement is different from that of *D. papillatum*. Moreover, all examined diplonemid species belong to experimentally tractable Diplonemidae and Hemistasiidae. We know nearly nothing about the metabolic capabilities and ecological role of the DSPDII group and of Eupelagomenidae in particular, which is the most abundant and diverse diplonemid clade. With recent advancements in addressing the challenges of single-cell technologies, from single-cell genomics to metabolomics [100], we should soon be able to fill this knowledge gap.

Methods

Strains and culture conditions

Diplonema papillatum (ATCC50162) was cultivated axenically at 15–22 °C in liquid medium containing 33 g/L Instant Ocean Sea Salt (Instant Ocean) and supplemented with 1% (v/v) horse serum as described earlier [73].

Extraction of nucleic acids and genome and transcriptome sequencing

Total cellular DNA was isolated from disrupted cells using Genomic-tip 100/G (Qiagen). Total cellular RNA was extracted using a home-made Trizol substitute [101], and residual DNA was removed by digestion with an RNase-free DNase. Poly(A) RNA was enriched by passage through oligo(dT)-cellulose. Library construction and Illumina and PacBio sequencing were performed by technology platforms. For details on strains, culture, nucleic acid extraction, and sequencing, see Additional file 1: Sect. 19. DNA and RNA preparation for high-throughput sequencing.

Assembly, structural, and functional annotation

We generated 462 million pairs of short reads (Illumina) and ~ 725,000 long reads (PacBio) totaling 126.4 Gbp raw data. Short and long reads were assembled separately with the Celera Assembler [102] and Canu [103], respectively, and non-redundant contigs were merged. Transcript sequences were obtained by de novo assembly of the ~ 645 million reads from the strand-specific poly(A) RNA libraries, and used in gene model prediction. Structural genome annotation was performed with an in-house developed tool [45]. For quality assessment, the gene models of the three largest contigs were expert-validated. Functional information was assigned by protein-sequence similarity to the SwissProt database and Hidden Markov Model (HMM) searches [104]. The proteins without SwissProt information were labelled as "hypothetical proteins." Transfer RNA genes were searched with tRNAscan-SE, rRNAs with Hmmerscan using profile HMMs from Rfam [105], and spliceosomal RNAs with Cmsearch [106] using home-built covariance models. For details on the assembly, annotation, and expert validation, see Additional file 1: Sect. 2. Assembly and annotation of the nuclear genome and transcriptome of *Diplonema papillatum* and Additional file 1: Sect. 4. Intron splicing and structural RNAs).

Otherwise, methods and data sources are described in detail in the corresponding Supplementary Information files with a focus on nuclear DNA structure and chromosome separation; ploidy; genomic repeats and NUMTs; RNA splicing, introns, and structural RNAs; cytoribosome; untranslated regions; polycistronic RNAs; DNA modifications; RNA interference machinery; gene complement; meiosis; CAZymes; nutrient assimilation; secretome; horizontal gene transfer; gene-family evolution; feeding behavior; and environmental distribution.

Abbreviations

DSPD	Deep-sea pelagic diplonemid
rRNA	Ribosomal RNA
SNV	Single-nucleotide variant
UTR	Untranslated region
SL	Spliced-leader
ORF	Open reading frame
TFIIα	Transcription initiation factor II alpha
CDS	Coding sequence
LINE-1	Long interspersed nuclear element-1
SLACS	Spliced-leader-associated-conserved sequence
TATE	Telomere associated transposable element
MULE	Mutator-like element
rDNA	Ribosomal DNA
TRIM	Terminal-repeat retrotransposon in miniature
SINE	Short interspersed element
NUMT	Nuclear mitochondrial segment
Base J	β-D-glucopyranosyl-oxymethyluracil
5mC	5-Methyldeoxycytosine
RNAi	RNA interference
ncRNA	Non-(protein-)coding RNA
tRNA	Transfer RNA
CAZyme	Carbohydrate-active enzyme
GH	Glycoside hydrolase
PL	Polysaccharide lyase
CE	Carbohydrate esterase
TEP	Transparent exopolymer particle
HGT	Horizontal gene transfer
LDCA	Last diplonemid common ancestor
ROS	Reactive oxygen species
OTU	Operational taxonomic unit

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12915-023-01563-9.

Additional file 1. Supporting information with additional details on the various topics described in the main text.

Additional file 2. Curated list of high-confidence nuclear mitochondrial segments (NUMTs) that can be anchored to non-repetitive, coding sequences of the mitochondrial genome. See also Additional file 1: Section 6. Repetitive sequences in the nuclear genome of *Diplonema papillatum* (assembly v_1.0).

Additional file 3. Gene count and orientation across contigs of the v1.0 assembly. Columns labelled 'Submitted genome annotation' show the data for the final assembly and annotation as submitted to NCBI GenBank. For the three longest contigs, we provide the corresponding information after expert curation and masking of genes derived from transposons (columns labelled 'Fully curated genome annotation and masked transposons'). See also Additional file 1: Section 7. Polycistronic transcription units in the nuclear genome of *Diplonema papillatum*.

Additional file 4. List of genes coding for carbohydrate-interacting proteins. A) CAZyme genes detected in *D. papillatum*. B) CAZyme genes detected in *D. japonicum*. See also Additional file 1: Section 12. CAZyme-coding genes in *Diplonema papillatum*.

Additional file 5. List of candidate genes horizontally transferred specifically from bacteria to *D. papillatum* or to the common ancestor of diplonemids. See also Additional file 1: Section 15. Genes horizontally transferred from bacteria to *Diplonema papillatum*.

Additional file 6. Detailed information on the evolution of gene families in diplonemids. A) Summary of gene-family gain/loss/expansion/contraction events. B) Diplonemid-specific gene-family gain events. Gene families gained on the ancestral diplonemid branch. C) Diplonemid-specific gene-family loss events. D) Diplonemid-specific gene-family expansion events. E) Diplonemid-specific gene-family contraction events. F) *D. papillatum*-specific gene-family gain events. Gene families gained on the ancestral diplonemid branch (posterior probability ≥0.5). G) *D. papillatum*-specific gene-family loss events. H) *D. papillatum*-specific gene-family expansion events. I) *D. papillatum*-specific gene-family contraction events. See also Additional file 1: Section 16. Evolution of gene families.

Additional file 7. List of samples from various locations of the world ocean investigated for the presence of *D. papillatum*. See also Additional file 1: Section 18. Environmental distribution of *Diplonema papillatum*.

Additional file 8. Uncropped gels and blots shown in the **Supplementary Figure S2.** See also Additional file 1: Section 1. Physical structure and size of the *D. papillatum* nuclear genome.

Acknowledgements
We thank Dr. Alastair Simpson (Dalhousie University, Halifax, Canada) for discussions on the diplonemid feeding apparatus; Dr. Daria Tashyreva (Institute of Parasitology, České Budějovice, Czech Republic) for scanning electron microscopy images of *D. papillatum* cells; Dr. Laura Landweber and her team (Columbia University, New York, USA) for initial help with the Celera assembler; and Dr. Fred Opperdoes (Université Catholique de Louvain, Brussels, Belgium) for validating the functional assignments of *D. papillatum* genes involved in metabolism.

Authors' contributions
Conceptualization, G.B., J.L., T.N., T.A.W.; methodology and software, C.P., M.S., P.S., S.M.; data curation, M.S., M.V., M.W.G., P.S.; resources: E.Y., T.B., T.N.; investigation, A.N., A.B., B.F.L., B.H., C.P., C.B., G.P., G.B., L.L., M.V., M.W.G., O.F., P.L., P.T., S.M.; formal analysis, A.B., A.N., B.F.L., B.H., C.P., G.B., L.L., M.V., M.W.G., O.F., S.M.; visualization, A.B., B.F.L., G.P., G.B., P.T., M.V.; writing—original draft / main text, G.B., M.V.; writing—original drafts / Supplementary Information, A.B., A.N., B.H., B.F.L., C.B., G.B., G.P., M.S., M.V., M.W.G., O.F.; writing—review and editing, B.F.L., G.B., J.L., M.V., M.W.G., T.A.W.; funding acquisition, B.F.L., G.B., J.L., T.A.W., T.N.; supervision, G.B., T.A.W. The authors read and approved the final manuscript.

Funding
This work was supported by grants from the *British Royal Society* for a University Research Fellowship (URF\R\201024 to T.A.W.); the *European Regional Development Fund* (ERDF 16_019/0000759; grant to J.L.); the *Fonds de Recherche du Québec*—Nature et Technologies (FRQNT; grant 2018-PR-206806 to B.F.L. and G.B.); the *Gordon and Betty Moore Foundation* (grants GBMF4983 to G.B. and J.L, GBMF#9354 to J.L., and GBMF9741 to T.A.W); the *Grant Agency of Czech Republic* (project 20-07186S and 23-06479X to J.L., and 23-07695S to A.B.); the *Japan Society for the Promotion of Science* (SPS KAKENHI; grant 25670205 to T.N.); the *Natural Environment Research Council* (NECP; grant NE/P00251X/1 to T.A.W); the *Natural Sciences and Engineering Research Council of Canada* (NSERC; grants RGPIN-2014–05286 and RGPIN-2019–04024 to G.B.; and RGPIN-2017–05411 to B.F.L.); and the *UK Biotechnology and Biological Sciences Research Council* (project BB/R016437/1 to T.A.W.). G.B. and M.W.G. acknowledge past support of research in this area by *CIHR* (grants MOP79309 and MOP4124, respectively).

Availability of data and materials
The datasets supporting the conclusions of this article are included as additional files or have been deposited under NCBI BioSample ID SAMN30986590 [107] and BioProject ID PRJNA883718 [108], including genome and transcriptome assemblies, genome annotations, and the inferred proteome.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Biochemistry, Robert-Cedergren Centre for Bioinformatics and Genomics, Université de Montréal, Montreal, QC, Canada. [2]School of Biological Sciences, University of Bristol, Bristol, UK. [3]Institute of Parasitology, Biology Centre, Czech Academy of Sciences, České Budějovice, Czech Republic. [4]Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic. [5]Faculty of Science, University of Ostrava, Ostrava, Czech Republic. [6]Present address: High Performance Computing Centre, Bristol, UK. [7]Architecture et Fonction des Macromolécules Biologiques (AFMB), CNRS, Aix Marseille Université, Marseille, France. [8]Present address: Environment Climate Change Canada, Dorval, QC, Canada. [9]Present address: Canadian Centre for Computational Genomics; McGill Genome Centre, McGill University, Montreal, QC, Canada. [10]RIKEN Interdisciplinary Theoretical and Mathematical Sciences Program (iTHEMS), Hirosawa, Wako, Saitama, Japan. [11]Laboratory of Molecular Parasitology, Graduate School of Life Science and Technology, Iryo Sosei University, Iwaki City, Fukushima, Japan. [12]Present address: DTU Bioengineering, Technical University of Denmark, Lyngby, Denmark. [13]Department of Biological Sciences, King Abdulaziz University, Jeddah, Saudi Arabia. [14]Department of Biochemistry and Molecular Biology, Institute for Comparative Genomics, Dalhousie University, Halifax, NS, Canada.

Received: 26 October 2022 Accepted: 10 March 2023
Published online: 04 May 2023

References

1. Flegontova O, Flegontov P, Malviya S, Audic S, Wincker P, de Vargas C, et al. Extreme diversity of diplonemid eukaryotes in the ocean. Curr Biol. 2016;26(22):3060–5.
2. Obiol A, Giner CR, Sánchez P, Duarte CM, Acinas SG, Massana R. A metagenomic assessment of microbial eukaryotic diversity in the global ocean. Mol Ecol Resour. 2020;20(3):718–31.
3. Flegontova O, Flegontov P, Londoño PAC, Walczowski W, Šantić D, Edgcomb VP, et al. Environmental determinants of the distribution of planktonic diplonemids and kinetoplastids in the oceans. Environ Microbiol. 2020;22(9):4014–31.
4. de Vargas C, Audic S, Henry N, Decelle J, Mahe F, Logares R, et al. Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. Science (New York, NY). 2015;348(6237):1261605.
5. Flegontova O, Flegontov P, Jachníková N, Lukeš J, Horák A. Water masses shape pico-nano eukaryotic communities of the Weddell Sea. Commun Biol. 2023;6(1):64.
6. Schoenle A, Hohlfeld M, Hermanns K, Mahé F, de Vargas C, Nitsche F, et al. High and specific diversity of protists in the deep-sea basins dominated by diplonemids, kinetoplastids, ciliates and foraminiferans. Commun Biol. 2021;4(1):501.

7. Mukherjee I, Salcher MM, Andrei A, Kavagutti VS, Shabarova T, Grujčić V, et al. A freshwater radiation of diplonemids. Environ Microbiol. 2020;22(11):4658–68.
8. Tashyreva D, Prokopchuk G, Votýpka J, Yabuki A, Horák A, Lukeš J. Life cycle, ultrastructure, and phylogeny of new diplonemids and their endosymbiotic bacteria. mBio. 2018;9(2):e02447-17.
9. Elbrächter M, Schnepf E, Balzer I. Hemistasia phaeocysticola (Scherffel) comb. nov., redescription of a free-living, marine, phagotrophic kinetoplastid flagellate. Arch Protistenkd. 1996;147(2):125–36.
10. Roy J, Faktorová D, Lukeš J, Burger G. Unusual mitochondrial genome structures throughout the Euglenozoa. Protist. 2007;158(3):385–96.
11. Prokopchuk G, Korytář T, Juricová V, Majstorović J, Horák A, Šimek K, et al. Trophic flexibility of marine diplonemids - switching from osmotrophy to bacterivory. ISME J. 2022;16:1409–19.
12 Pilátová J, Tashyreva D, Týč J, Vancová M, Bokhari SNH, Skoupý R, et al. Massive accumulation of strontium and barium in diplonemid protists. mBio. 2023;14(1):e0327922.
13. Lara E, Moreira D, Vereshchaka A, Lopez-Garcia P. Pan-oceanic distribution of new highly diverse clades of deep-sea diplonemids. Environ Microbiol. 2009;11(1):47–55.
14. Okamoto N, Gawryluk RMR, Del Campo J, Strassert JFH, Lukeš J, Richards TA, et al. A revised taxonomy of diplonemids including the Eupelagonemidae n. fam. and a type species, *Eupelagonema oceanica* n. gen. & sp. J Eukaryot Microbiol. 2019;66(3):519–24.
15. Kostygov AY, Karnkowska A, Votýpka J, Tashyreva D, Maciszewski K, Yurchenko V, et al. Euglenozoa: taxonomy, diversity and ecology, symbioses and viruses. Open Biol. 2021;11(3):200407.
16. Tashyreva D, Prokopchuk G, Yabuki A, Kaur B, Faktorová D, Votýpka J, et al. Phylogeny and morphology of new diplonemids from Japan. Protist. 2018;169(2):158–79.
17. Prokopchuk G, Tashyreva D, Yabuki A, Horák A, Masařová P, Lukeš J. Morphological, ultrastructural, motility and evolutionary characterization of two new Hemistasiidae species. Protist. 2019;170(3):259–82.
18. Tashyreva D, Simpson AGB, Prokopchuk G, Škodová-Sveráková I, Butenko A, Hammond M, et al. Diplonemids - a review on "new" flagellates on the oceanic block. Protist. 2022;173(2):125868.
19. Gawryluk RMR, Del Campo J, Okamoto N, Strassert JFH, Lukeš J, Richards TA, et al. Morphological identification and single-cell genomics of marine diplonemids. Curr Biol. 2016;26(22):3053–9.
20. López-García P, Vereshchaka A, Moreira D. Eukaryotic diversity associated with carbonates and fluid-seawater interface in Lost City hydrothermal field. Environ Microbiol. 2007;9(2):546–54.
21. Derelle R, Torruella G, Klimeš V, Brinkmann H, Kim E, Vlček Č, et al. Bacterial proteins pinpoint a single eukaryotic root. Proc Natl Acad Sci U S A. 2015;112(7):E693–9.
22. Lukeš J, Wheeler R, Jirsová D, David V, Archibald JM. Massive mitochondrial DNA content in diplonemid and kinetoplastid protists. IUBMB Life. 2018;70(12):1267–74.
23. Kiethega GN, Yan Y, Turcotte M, Burger G. RNA-level unscrambling of fragmented genes in *Diplonema* mitochondria. RNA Biol. 2013;10(2):301–13.
24. Moreira S, Valach M, Aoulad-Aissa M, Otto C, Burger G. Novel modes of RNA editing in mitochondria. Nucleic Acids Res. 2016;44(10):4907–19.
25. Lukeš J, Kaur B, Speijer D. RNA editing in mitochondria and plastids: weird and widespread. Trends Genet. 2021;37(2):99–102.
26. Kaur B, Záhonová K, Valach M, Faktorová D, Prokopchuk G, Burger G, et al. Gene fragmentation and RNA editing without borders: eccentric mitochondrial genomes of diplonemids. Nucleic Acids Res. 2020;48(5):2694–708.
27. Valach M, Moreira S, Hoffmann S, Stadler PF, Burger G. Keeping it complicated: mitochondrial genome plasticity across diplonemids. Sci Rep. 2017;7(1):14166.
28. Kaur B, Valach M, Peña-Diaz P, Moreira S, Keeling PJ, Burger G, et al. Transformation of *Diplonema papillatum*, the type species of the highly diverse and abundant marine microeukaryotes Diplonemida (Euglenozoa). Environ Microbiol. 2018;20(3):1030–40.
29. Faktorová D, Kaur B, Valach M, Graf L, Benz C, Burger G, et al. Targeted integration by homologous recombination enables *in situ* tagging and replacement of genes in the marine microeukaryote *Diplonema papillatum* Environ Microbiol. 2020;22:3660–70.
30. von der Heyden S, Chao EE, Vickerman K, Cavalier-Smith T. Ribosomal RNA phylogeny of bodonid and diplonemid flagellates and the evolution of Euglenozoa. J Eukaryot Microbiol. 2004;51(4):402–16.
31. Busse I, Preisfeld A. Phylogenetic position of *Rhynchopus* sp. and *Diplonema ambulator* as indicated by analyses of euglenozoan small subunit ribosomal DNA. Gene. 2002;284(1–2):83–91.
32. Záhonová K, Lax G, Sinha SD, Leonard G, Richards TA, Lukeš J, et al. Single-cell genomics unveils a canonical origin of the diverse mitochondrial genomes of euglenozoans. BMC Biol. 2021;19(1):103.
33. Wideman JG, Lax G, Leonard G, Milner DS, Rodríguez-Martínez R, Simpson AGB, et al. A single-cell genome reveals diplonemid-like ancestry of kinetoplastid mitochondrial gene structure. Philos Trans R Soc Lond B Biol Sci. 2019;374(1786):20190100.
34. Butenko A, Opperdoes FR, Flegontova O, Horák A, Hampl V, Keeling P, et al. Evolution of metabolic capabilities and molecular features of diplonemids, kinetoplastids, and euglenids. BMC Biol. 2020;18(1):23.
35. 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 diplonemids and kinetoplastids. Proc Biol Sci. 1830;2016(283):20160520.
36. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2.
37. Preußer C, Jaé N, Bindereif A. mRNA splicing in trypanosomes. Int J Med Microbiol. 2012;302(4–5):221–4.
38. Tessier LH, Keller M, Chan RL, Fournier R, Weil JH, Imbault P. Short leader sequences may be transferred from small RNAs to pre-mature mRNAs by trans-splicing in *Euglena* EMBO J. 1991;10(9):2621–5.
39. Salzberg SL. Next-generation genome annotation: we still struggle to get it right. Genome Biol. 2019;20(1):92.
40. Kim E, Magen A, Ast G. Different levels of alternative splicing among eukaryotes. Nucleic Acids Res. 2006;35(1):125–31.
41. Yeoh LM, Goodman CD, Mollard V, McHugh E, Lee VV, Sturm A, et al. Alternative splicing is required for stage differentiation in malaria parasites. Genome Biol. 2019;20(1):151.
42 Frey KA-O, Pucker BA-O. Animal, fungi, and plant genome sequences harbor different non-canonical splice sites. Cells. 2020;9(2):458.
43. 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(1):49.
44. Kolev NG, Ullu E, Tschudi C. The emerging role of RNA-binding proteins in the life cycle of *Trypanosoma brucei* Cell Microbiol. 2014;16(4):482–9.
45. Gray MW, Burger G, Derelle R, Klimeš V, Léger MM, Sarrasin M, et al. The draft nuclear genome sequence and predicted mitochondrial proteome of *Andalucia godoyi*, a protist with the most gene-rich and bacteria-like mitochondrial genome. BMC Biol. 2020;18(1):22.
46. 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(2):161–72.
47. 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(5):631–42.
48. Ebenezer TE, Zoltner M, Burrell A, Nenarokova A, Novák Vanclová AMG, Prasad B, et al. Transcriptome, proteome and draft genome of *Euglena gracilis* BMC Biol. 2019;17(1):11.
49. Kojima KK. Structural and sequence diversity of eukaryotic transposable elements. Genes Genet Syst. 2020;94(6):233–52.
50. Vlcek C, Marande W, Teijeiro S, Lukeš J, Burger G. Systematically fragmented genes in a multipartite mitochondrial genome. Nucleic Acids Res. 2011;39(3):979–88.
51. Calabrese FM, Balacco DL, Preste R, Diroma MA, Forino R, Ventura M, et al. NumtS colonization in mammalian genomes. Sci Rep. 2017;7(1):16357.
52. Ko YJ, Kim S. Analysis of nuclear mitochondrial DNA segments of nine plant species: size, distribution, and insertion loci. Genom Inform. 2016;14(3):90–5.
53. Michalovova M, Vyskot B, Kejnovsky E. Analysis of plastid and mitochondrial DNA insertions in the nucleus (NUPTs and NUMTs) of six plant species: size, relative age and chromosomal localization. Heredity. 2013;111(4):314–20.
54. Kolev NG, Franklin JB, Carmi S, Shi H, Michaeli S, Tschudi C. The transcriptome of the human pathogen *Trypanosoma brucei* at single-nucleotide resolution. PLoS Pathog. 2010;6(9):e1001090.
55 Sturm NR, Maslov DA, Grisard EC, Campbell DA. Diplonema spp possess spliced leader RNA genes similar to the Kinetoplastida. J Eukaryot Microbiol. 2001;48(3):325–31.

56. Vanhamme L, Pays E. Control of gene expression in trypanosomes. Microbiol Rev. 1995;59(2):223–40.
57. Borst P, Sabatini R. Base J: discovery, biosynthesis, and possible functions. Annu Rev Microbiol. 2008;62:235–51.
58. van Luenen HG, Farris C, Jan S, Genest PA, Tripathi P, Velds A, et al. Glucosylated hydroxymethyluracil, DNA base J, prevents transcriptional readthrough in *Leishmania* Cell. 2012;150(5):909–21.
59. Schmitz RJ, Lewis ZA, Goll MG. DNA methylation: shared and divergent features across eukaryotes. Trends Genet. 2019;35(11):818–27.
60. Ketting RF. The many faces of RNAi. Dev Cell. 2011;20(2):148–61.
61. Gutbrod MJ, Martienssen RA. Conserved chromosomal functions of RNA interference. Nat Rev Genet. 2020;21(5):311–31.
62. 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.
63. Matveyev AV, Alves JM, Serrano MG, Lee V, Lara AM, Barton WA, et al. The evolutionary loss of RNAi key determinants in kinetoplastids as a multiple sporadic phenomenon. J Mol Evol. 2017;84(2–3):104–15.
64. Matzov D, Taoka M, Nobe Y, Yamauchi Y, Halfon Y, Asis N, et al. Cryo-EM structure of the highly atypical cytoplasmic ribosome of *Euglena gracilis* Nucleic Acids Res. 2020;48(20):11750–61.
65. Hałakuc P, Karnkowska A, Milanowski R. Typical structure of rRNA coding genes in diplonemids points to two independent origins of the bizarre rDNA structures of euglenozoans. BMC Ecol Evol. 2022;22(1):59.
66. Ehrlich R, Davyt M, López I, Chalar C, Marín M. On the track of the missing tRNA genes: a source of non-canonical functions? Front Mol Biosci. 2021;8.
67. Guy MP, Phizicky EM. Two-subunit enzymes involved in eukaryotic post-transcriptional tRNA modification. RNA Biol. 2014;11(12):1608–18.
68. Rubio MA, Pastar I, Gaston KW, Ragone FL, Janzen CJ, Cross GA, et al. An adenosine-to-inosine tRNA-editing enzyme that can perform C-to-U deamination of DNA. Proc Natl Acad Sci U S A. 2007;104(19):7821–6.
69. Drouin G, Tsang C. 5S rRNA gene arrangements in protists: a case of nonadaptive evolution. J Mol Evol. 2012;74(5–6):342–51.
70. Jean-Joseph B, Flisser A, Martinez A, Metzenberg S. The U5/U6 snRNA genomic repeat of *Taenia solium* J Parasitol. 2003;89(2):329–35.
71. Makiuchi T, Annoura T, Hashimoto M, Hashimoto T, Aoki T, Nara T. Compartmentalization of a glycolytic enzyme in *Diplonema*, a non-kinetoplastid euglenozoan. Protist. 2011;162(3):482–9.
72. Škodová-Sveráková I, Prokopchuk G, Peña-Diaz P, Záhonová K, Moos M, Horváth A, et al. Unique dynamics of paramylon storage in the marine euglenozoan *Diplonema papillatum* Protist. 2020;171(2).
73. Valach M, Léveillé-Kunst A, Gray MW, Burger G. Respiratory chain Complex I of unparalleled divergence in diplonemids. J Biol Chem. 2018;293(41):16043–56.
74. Škodová-Sveráková I, Záhonová K, Bučková B, Füssy Z, Yurchenko V, Lukeš J. Catalase and ascorbate peroxidase in euglenozoan protists. Pathogens (Basel, Switzerland). 2020;9(4):317.
75. Škodová-Sveráková I, Záhonová K, Juricová V, Danchenko M, Moos M, Baráth P, et al. Highly flexible metabolism of the marine euglenozoan protist *Diplonema papillatum* BMC Biol. 2021;19(1):251.
76. Le Costaouëc T, Unamunzaga C, Mantecon L, Helbert W. New structural insights into the cell-wall polysaccharide of the diatom *Phaeodactylum tricornutum* Algal Res. 2017;26:172–9.
77. Michel G, Tonon T, Scornet D, Cock JM, Kloareg B. Central and storage carbon metabolism of the brown alga *Ectocarpus siliculosus*: insights into the origin and evolution of storage carbohydrates in eukaryotes. New Phytol. 2010;188(1):67–81.
78. Becker S, Tebben J, Coffinet S, Wiltshire K, Iversen MH, HarderT, et al. Laminarin is a major molecule in the marine carbon cycle. Proc Natl Acad Sci U S A. 2020;117(12):6599–607.
79. Passow U. Transparent exopolymer particles (TEP) in aquatic environments. Progress Oceanogr. 2002;55(3):287–333.
80 Roy J, Faktorová D, Benada O, Lukeš J, Burger G. Description of *Rhynchopus euleeides* n. sp. (Diplonemea), a free-living marine euglenozoan. J Eukaryot Microbiol. 2007;54(2):137–45.
81. Ralton JE, Sernee MF, McConville MJ. Evolution and function of carbohydrate reserve biosynthesis in parasitic protists. Trends Parasitol. 2021;37(11):988–1001.
82 Porter D. *Isonema papillatum* sp. n., a new colorless marine flagellate: a light- and electronmicroscopic study. J Protozool. 1973;20(3):351–6.
83. Fan X, Qiu H, Han W, Wang Y, Xu D, Zhang X, et al. Phytoplankton pangenome reveals extensive prokaryotic horizontal gene transfer of diverse functions. Sci Adv. 2020;6(18):0111.
84. Husnik F, McCutcheon JP. Functional horizontal gene transfer from bacteria to eukaryotes. Nat Rev Microbiol. 2018;16(2):67–79.
85. Eme L, Gentekaki E, Curtis B, Archibald JM, Roger AJ. Lateral gene transfer in the adaptation of the anaerobic parasite *Blastocystis* to the gut. Curr Biol. 2017;27(6):807–20.
86. Alsmark C, Foster PG, Sicheritz-Ponten T, Nakjang S, Martin Embley T, Hirt RP. Patterns of prokaryotic lateral gene transfers affecting parasitic microbial eukaryotes. Genome Biol. 2013;14(2):R19.
87. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20(1):238.
88. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32(1):268–74.
89. Csurös M. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics. 2010;26(15):1910–2.
90. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The protein kinase complement of the human genome. Science (New York, NY). 2002;298(5600):1912–34.
91. Lehti-Shiu MD, Shiu SH. Diversity, classification and function of the plant protein kinase superfamily. Philos Trans R Soc Lond B Biol Sci. 2012;367(1602):2619–39.
92. Wawrik B, Bronk DA, Baer SE, Chi L, Sun M, Cooper JT, et al. Bacterial utilization of creatine in seawater. Aquat Microb Ecol. 2017;80(2):153–65.
93. Larsen J, Patterson DJ. Some flagellates (Protista) from tropical marine sediments. J Nat Hist. 1990;24:801–937.
94. Gerbracht JV, Harding T, Simpson AGB, Roger AJ, Hess S. Comparative transcriptomics reveals the molecular toolkit used by an algivorous protist for cell wall perforation. Curr Biol. 2022;32(15):3374-84.e5.
95. Ibarbalz FM, Henry N, Brandão MC, Martini S, Busseni G, Byrne H, et al. Global trends in marine plankton diversity across kingdoms of life. Cell. 2019;179(5):1084-97.e21.
96. Kopf A, Bicak M, Kottmann R, Schnetzer J, Kostadinov I, Lehmann K, et al. The ocean sampling day consortium. GigaScience. 2015;4:27.
97. Käse L, Kraberg AC, Metfies K, Neuhaus S, Sprong PAA, Fuchs BM, et al. Rapid succession drives spring community dynamics of small protists at Helgoland Roads, North Sea. J Plankton Res. 2020;42(3):305–19.
98. Massana R, Gobet A, Audic S, Bass D, Bittner L, Boutte C, et al. Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing. Environ Microbiol. 2015;17(10):4035–49.
99. Ramond P, Sourisseau M, Simon N, Romac S, Schmitt S, Rigaut-Jalabert F, et al. Coupling between taxonomic and functional diversity in protistan coastal communities. Environ Microbiol. 2019;21(2):730–49.
100. Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, et al. Eleven grand challenges in single-cell data science. Genome Biol. 2020;21(1):31.
101. Rodriguez-Ezpeleta N, Teijeiro S, Forget L, Burger G, Lang BF. Construction of cDNA libraries: focus on protists and fungi. Methods Mol Biol. 2009;533:33–47.
102. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, et al. A whole-genome assembly of *Drosophila* Science (New York, NY). 2000;287(5461):2196–204.
103. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27(5):722–36.
104. Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. HMMER web server: 2018 update. Nucleic Acids Res. 2018;46(W1):W200–4.
105. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. Rfam: an RNA family database. Nucleic Acids Res. 2003;31(1):439–41.
106 Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013;29(22):2933–5.

107. Valach M, Moreira S, Petitjean C, Benz C, Butenko A, Flegontova O, Nenarokova A Prokopchuk G, Batsone T, Lapebie P, Limogo L, Sarrasin M, Stretenowich P, Tripathi P, Nara T, Henrissat B, Lang BF, Gray MW, Williams TA, Lukes J and Burger G. https://identifiers.org/biosample:SAMN30986590 (2023).
108. Valach M, Moreira S, Petitjean C, Benz C, Butenko A, Flegontova O, Nenarokova A, Prokopchuk G, Batsone T, Lapebie P, Limogo L, Sarrasin M, Stretenowich P, Tripathi P, Nara T, Henrissat B, Lang BF, Gray MW, Williams TA, Lukes J and Burger G. https://identifiers.org/bioproject:PRJNA883718 (2023).

Publisher's Note

Supporting Information for

Recent expansion of metabolic versatility in *Diplonema papillatum*, the model species of a highly speciose group of marine eukaryotes

Matus Valach[1*], Sandrine Moreira[1], Celine Petitjean[2], Corinna Benz[3], Anzhelika Butenko[3,4,5], Olga Flegontova[3,5], Anna Nenarokova[2,3], Galina Prokopchuk[3,4], Tom Batstone[2,6], Pascal Lapébie[7], Lionnel Lemogo[1,8], Matt Sarrasin[1], Paul Stretenowich[1,9], Pragya Tripathi[3,4], Euki Yazaki[10], Takeshi Nara[11], Bernard Henrissat[7,12,13], B. Franz Lang[1], Michael W. Gray[14], Tom A. Williams[2], Julius Lukeš[3,4] and Gertraud Burger[1*]

[1] Department of Biochemistry, Robert-Cedergren Centre for Bioinformatics and Genomics, Université de Montréal, Montreal, Quebec, Canada.
[2] School of Biological Sciences, University of Bristol, Bristol, UK.
[3] Institute of Parasitology, Biology Centre, Czech Academy of Sciences, České Budějovice, Czech Republic.
[4] Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic.
[5] Faculty of Science, University of Ostrava, Ostrava, Czech Republic.
[6] Current address: High Performance Computing Centre, Bristol, UK.
[7] Architecture et Fonction des Macromolécules Biologiques (AFMB), CNRS, Aix Marseille Université, Marseille, France.
[8] Current address: Environment Climate Change Canada, Dorval, Quebec, Canada.
[9] Current address: Canadian Centre for Computational Genomics; McGill Genome Centre, McGill University, Montreal, Quebec, Canada.
[10] RIKEN Interdisciplinary Theoretical and Mathematical Sciences Program (iTHEMS), Hirosawa, Wako, Saitama, Japan.
[11] Laboratory of Molecular Parasitology, Graduate School of Life Science and Technology, Iryo Sosei University, Iwaki City, Fukushima, Japan.
[12] Current address: DTU Bioengineering, Technical University of Denmark, Lyngby, Denmark;
[13] Department of Biological Sciences, King Abdulaziz University, Jeddah, Saudi Arabia.
[14] Department of Biochemistry and Molecular Biology, Institute for Comparative Genomics, Dalhousie University, Halifax, Nova Scotia, Canada.
* Corresponding authors.

Matus Valach
Email: matus.a.valach@gmail.com
Gertraud Burger
Email: gertraud.burger@umontreal.ca

This PDF file includes:

Supporting Information Text
Supplementary Figures
Supplementary Tables
Supporting Information References

Supporting Information Text

Contents

1. Physical structure and size of the *D. papillatum* nuclear genome

INTRODUCTION

Long repetitive regions such as centromeres or stretches rich in modified nucleotides have been refractory to complete genome assembly in many organisms. Consequently, we do not know the exact size of the corresponding genomes nor the number and topology of chromosomes. Still, these genomic features can be investigated by various alternative approaches, which we have employed to characterize the nuclear genome of *D. papillatum*.

Flow cytometry

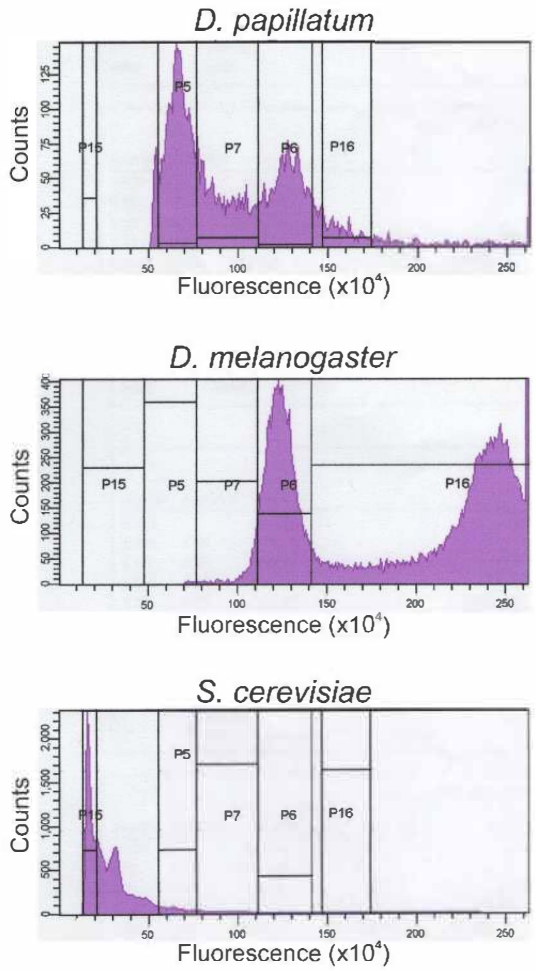
One of the most common experimental methods to estimate the size of a nuclear genome is flow cytometry, which involves the isolation of intact nuclei, DNA staining with a fluorescence dye, and the measurement of the fluorescence signal intensity of individual nuclei. The signal is then compared to that from nuclei of organisms whose genome size is known. Flow cytometry measures the DNA *amount* in cell nuclei, whereby typically the DNA content in both the G1 and G2 phases can be determined. For inferring the genome size, the ploidy level of the organism must be known. If the species is haploid, i.e., mitotic proliferation takes place in the haploid stage, then the genome size corresponds to the DNA amount in the G1 phase. If, however, the species is diploid or of higher ploidy, the genome size would be half etc., of the DNA amount in G1.

To measure the relative nuclear DNA content of *D. papillatum* by flow cytometry, we used a diploid *Saccharomyces cerevisiae* strain (24 Mbp) and a diploid *Drosophila melanogaster* cell line (360 Mbp) as references. *D. papillatum* and the organismal material used as a reference have a single nucleus per cell. The DNA content was measured by the fluorescence intensity after staining the nuclei with the fluorophore 4′,6-diamidino-2-phenylindole (DAPI). **Supplementary Figure S1** shows the distribution of the DNA amount across the nuclei examined. All three samples display a prominent peak of G1-phase cells and a minor population in G2, since the cultures were asynchronous. The DNA content of *D. papillatum* is smaller than that of the fly but considerably larger than that of yeast.

However, extrapolating the exact DNA content of *D. papillatum* nuclei is not feasible in this case due to several limitations. One issue is that DAPI binds preferentially to A+T-rich DNA and that the A+T content of the three nuclear genomes varies between yeast 52% (yeast), *Diplonema* (55%) and fly (57%). Still, the advantage of DAPI is that, in contrast to other dyes, DNA binding is not influenced by chromatin structure, which fluctuates drastically between the three organisms. The yeast genome contains nearly no heterochromatin, whereas that of the fly includes 30%, and that of *D. papillatum* is estimated at 50% considering the large portion of repetitive sequence and mobile elements. The second issue is the difference in cell-wall composition and rigidity that necessitated specifically adapted cell-rupture and nuclei-isolation procedures, which in turn produced nuclei of various purity and yield. The elevated fluorescence background that we observed in *D. papillatum* was most likely due to the contamination by mitochondrial DNA, which is present in an extraordinarily high proportion in this species (LUKEŠ *et al.* 2018).

In conclusion, the flow-cytometry determination of the DNA content in the *D. papillatum* nucleus shown here is in rough agreement with the haploid genome size inferred from the DNA sequence, i.e. 260 Mbp estimated from the k-mer distribution in reads and 280 Mbp, the size of the nuclear genome assembly (see main text).

Yet, to infer the actual genome size of an organism, the ploidy level must be known. As detailed in the **Supplementary Information**: Section 3. The ploidy level of *Diplonema papillatum*, the *D. papillatum* nuclear genome displays extremely low heterozygosity as revealed by the unimodal k-mer distribution and the low number of single-nucleotide variants. This indicates that the *D. papillatum* nuclear genome is either haploid or alternatively, autodiploid (or even tetraploid, etc.) originating from a recent genome duplication without allelic divergence. However, autodiploidy of *Diplonema* is highly unlikely because the nuclear DNA content would then be >500 Mbp, which contradicts the flow cytometry experiments shown here. Therefore, we conclude that the *D. papillatum* nuclear genome is haploid with a (haploid) genome size between 100 and 300 Mbp. Considering the size obtained by k-mer counts in sequencing reads, we consider 260 Mbp to be the most accurate estimate of the *D. papillatum* nuclear genome.



Supplementary Figure S1. Histograms of fluorescence intensity corresponding to nuclear DNA content. The left and right peaks represent nuclei in the cell-cycle phases G1 and G2, respectively. Nuclei were stained with DAPI, and fluorescence was measured with a FACScalibur instrument. Counts, number of nuclei. P5 to P16, nuclei populations selected by gating.

Chromosomes and nuclear organization

D. papillatum cells have a single nucleus (MARANDE *et al.* 2005), but the number of chromosomes and their condensation state are unknown. We addressed these questions through pulsed-field gel electrophoresis and electron microscopy experiments.

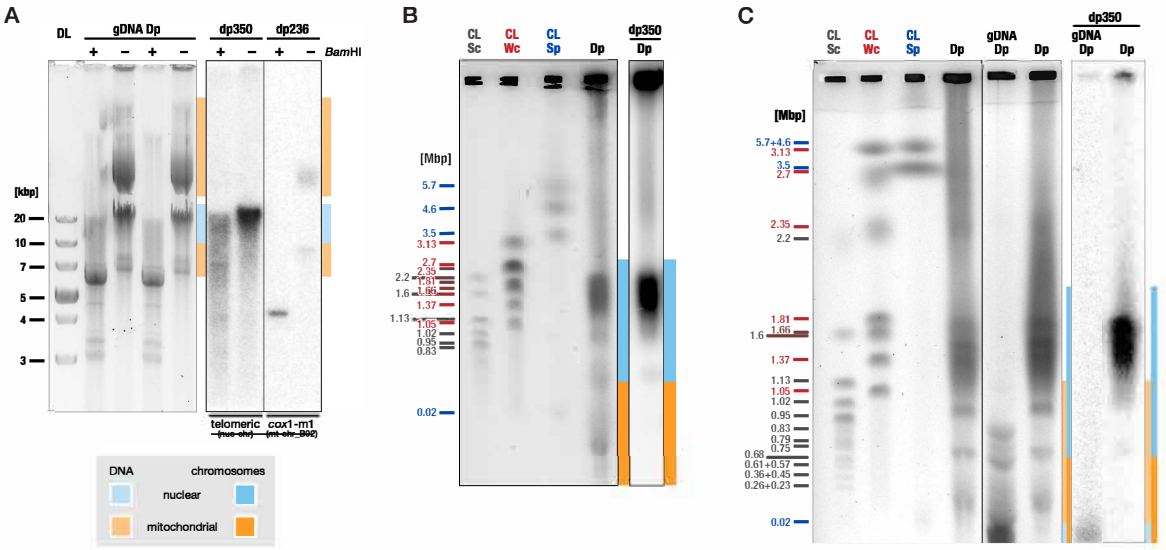
Number and size of chromosomes

Pulsed-field gel electrophoresis (PFGE) separated readily intact nuclear chromosomes from the 6-7 kbp circles of mitochondrial DNA (Marande et al. 2005; Moreira et al. 2016). After migration, the nuclear DNA was visualized by Southern hybridization using a probe that contains telomeric repeats. Three signals were observed: a smear spanning from 1.1 to 1.8 Mbp with only a slight indication of a banding pattern and, in addition, two thin bands at ~500 kbp and ~1 Mbp (**Supplementary Figure S2**). Most of the nuclear DNA in the 1.1–1.8 Mbp zone most likely corresponds to a large number of unresolved chromosomes, whereas the thin bands appear to represent single chromosomes. At an estimated size of ~260 Mbp for the entire genome (which is haploid, see main text), the *D. papillatum* nucleus is predicted to contain approximately 180 chromosomes.

Compared to other Euglenozoa, the *Diplonema* nuclear genome appears much more fragmented than that of its relatives. Trypanosomes and leishmanias (nuclear genome size ~18–33 Mbp) have 11–36 chromosomes ranging in size from 0.5 to 6 Mbp. In addition, trypanosomes also contain up to a hundred smaller, non-Mendelian chromosomes (Maslov et al. 2019). Similarly, the *Euglena gracilis* nuclear genome (size ~500 Mbp (Ebenezer et al. 2019)) has an estimated chromosome number of 42 (Dooijes et al. 2000).

DNA arrangement in the nucleus

The arrangement of nuclear DNA was assessed by thin-section transmission electron microscopy and classical and expansion fluorescence microscopy. The nucleus of *Diplonema* cells occupied a volume of, on average, 14.35 μm^3 and was roughly spherical (0.895 sphericity). In all three nuclei analyzed in detail, the chromosomes formed a network structure (0.104 sphericity) that took up ~18% of the nucleus (**Supplementary Figure S3A-D, S4**). Despite the size variation of the nuclei examined, the proportions of sphericity and volume of nuclei and reticulated chromatin were constant. Occasionally, we observed small (<0.117 μm^3) chromatin granules detached from the chromosomal network that might represent analogs of Cajal bodies or nuclear speckles. When cells were stained with DAPI or other DNA-binding dyes and inspected by fluorescence microscopy, the network appeared as a pattern of bright clusters in a spherical arrangement characteristic of diplonemids (Lukeš et al. 2018) (**Supplementary Figure S3E, S4A,E**). This indicates that certain chromatin regions accumulated more dye and thus were likely more condensed. Inside the nucleus of the *D. papillatum* cells resided a clearly discernable nucleolus that was spherical and slightly granular and filled ~8% of the nucleus (**Supplementary Figure S3**).

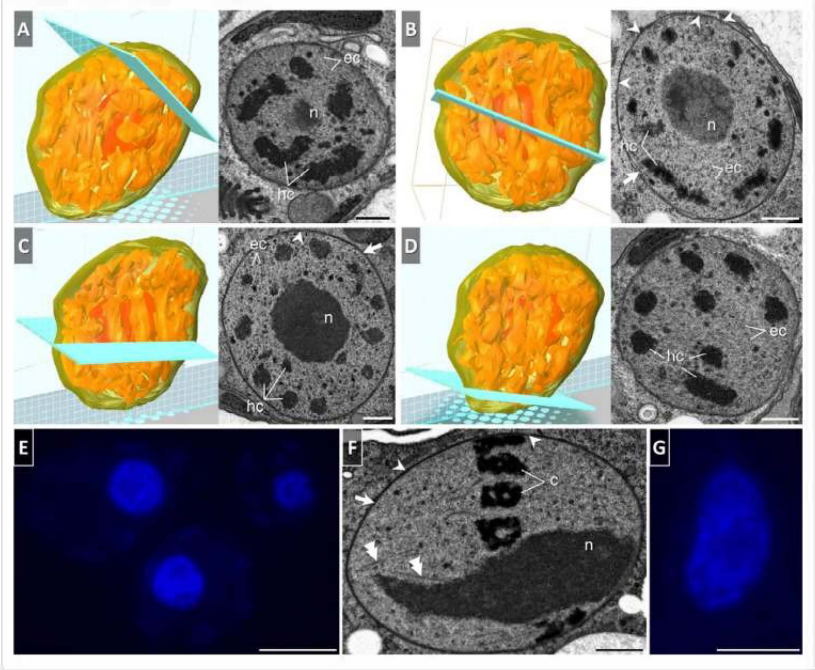


Supplementary Figure S2. Electrophoretic separation of *D. papillatum* nuclear chromosomes. (**A**) Total genomic DNA from *D. papillatum* cells was separated by conventional agarose gel electrophoresis after (+) or prior to (−) digestion with BamHI, blotted, and hybridized to a nuclear telomeric probe (dp350) or a probe targeting the mitochondrial gene piece cox1-m1 (dp236; located on the circular chromosome B02). DL, DNA ladder (GeneRuler 1kb Plus; Thermo Fisher). For oligonucleotide-probe sequences, see **Supplementary Table S2**. (**B**, **C**) DNA from in-gel-lysed cells (Dp) or total genomic DNA (gDNA Dp) was separated by PFGE using two different programs (#1 in B, #2 in C; for details, see **Supplementary Table S2**). After Southern blotting, membranes were hybridized to the telomeric probe dp350. CL, chromosomal ladders (Bio-Rad): Sc, *Saccharomyces cerevisiae*; Wc, *Wickerhamomyces canadensis* (*Hansenula wingei*); Sp, *Schizosaccharomyces pombe*. Key, highlighted zones in which most of the DNA staining-signal originates from nuclear DNA (lighter and darker orange) or mitochondrial DNA (lighter and darker blue).

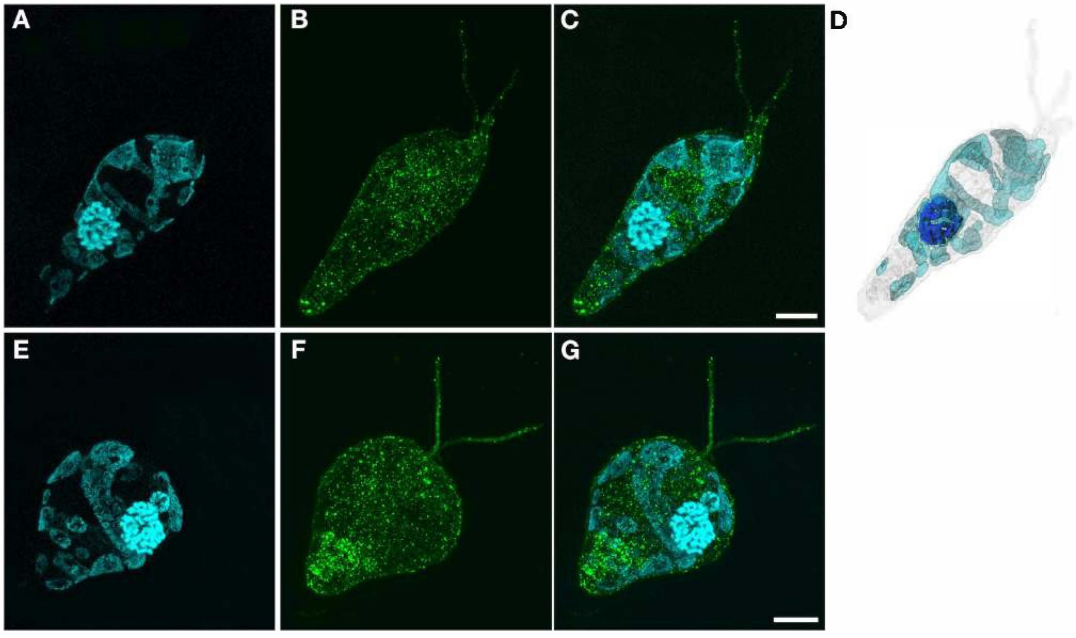
The study presented here is the first high-resolution examination of the chromatin organization in a diplonemid and, to the best of our knowledge, in a euglenozoan. The first morphological studies of *D. papillatum* already reported that the cells bear a relatively **large nucleus** (~3-4 µm in diameter) of a somewhat ellipsoidal shape and medial location (PORTER 1973). Similar observations were made in other diplonemid species (TASHYREVA *et al.* 2018a; TASHYREVA *et al.* 2018b; PROKOPCHUK *et al.* 2019). A large nucleus is a hallmark of Euglenozoa in general (ANGELER *et al.* 1999; BREGLIA *et al.* 2010; LUKEŠ *et al.* 2021). Another particular feature of the diplonemid nucleus is that the shape and size do not change with growth conditions and stage (e.g., starvation or exponential vs. stationary phase), while the opposite has been reported for trypanosomatids (VICKERMAN AND PRESTON 1970; SCHENKMAN *et al.* 2011).

The **nucleolus** of *D. papillatum* is a conspicuous structure positioned centrally within the nucleus, a trait shared with other Euglenozoa. Size-wise, the nucleolus of *D. papillatum* is comparable with that of kinetoplastids and euglenids examined so far: it occupied a volume slightly larger than in *T. brucei* (ERSFELD 2011), but smaller than in *Euglena gracilis* (O'DONNELL 1965).

The most noticeable nuclear substructure is the permanently reticulated **heterochromatin** (**Supplementary Figures S3A–E, S4**), probably representing permanently condensed subdomains of the chromosomal network. Hemistasiids (PROKOPCHUK *et al.* 2019) and symbiotids (YUBUKI *et al.* 2009) display similar patches and agglomerated masses of heterochromatin in the nucleoplasm, but the corresponding studies did not reconstruct the nuclear ultrastructure in three dimensions and therefore did not inform about the degree of reticulation.



Supplementary Figure S3. Nuclear organization of *D. papillatum*. (**A-D**) TEM images showing cross sections through different parts of the nucleus. The planes of sections are marked in cyan on a 3D model of the entire nucleus. Yellow – nuclear membrane; orange, hc – heterochromatin; red, n – nucleolus; ec – euchromatin; arrows – perinuclear space; arrowheads – nuclear pores. (**E**) DAPI-stained nucleus. (**F-G**) TEM and DAPI-stained nucleus during division; c – chromosomes. Spindle fibers are seen (double arrowheads). Scale bar: 500 nm (**A-D**, F), 5 µm (**F**), 2.5 µm (**G**).



Supplementary Figure S4. Nuclear and mitochondrial organization of *D. papillatum*. (**A-C, E-G**) Expansion microscopy (ExM) images of cells stained with DAPI (**A, E**) and anti-tubulin antibodies (**B, F**), and their merged display (**C, G**). ExM enables more even staining of DNA in the nucleus and mitochondria (compare with Supplementary Figure S3E–F). (**D**) 3D model of a *D. papillatum* cell based on the Z-stack series of the cell shown in **A–C**. Light gray – cell surface tubulin; cyan – reticulated mitochondrion; blue – reticulated nuclear heterochromatin. Scale bars: 10 µm.

In *D. papillatum* the **chromatin distribution** is uniform throughout the nuclear space. Such a pattern is common across diplonemids (ROY *et al.* 2007; TASHYREVA *et al.* 2018a; TASHYREVA *et al.* 2018b; PROKOPCHUK *et al.* 2019) and *E. gracilis* (O'DONNELL 1965). However, in most kinetoplastids, symbiotids, and certain euglenids, the chromatin distribution pattern is very different, with condensed heterochromatin occupying the periphery of the nucleus close to the envelope (ELIAS *et al.* 2002; BREGLIA *et al.* 2010).

D. papillatum, and Discoba in general, undergo closed **mitosis** (PATTERSON 1999), whereby the nuclear membrane and the nucleolus structure–which elongates and divides into two daughter nucleoli–are preserved. Further, during cell division, diplonemids form a distinct metaphase plate (**Supplementary Figure S3F**) (PORTER 1973; TRIEMER 1992), while kinetoplastids and certain euglenids assemble chromosomes in a loose equatorial plane (TRIEMER AND FARMER 1991). Metaphase plates occur only in a few other protist groups, but are a typical feature of mitotic division in animals and plants.

Lastly, we show here that expansion microscopy (ExM), a powerful technique allowing visualization of sub-cellular ultrastructures through physical (or, more precisely, mechano-chemical) rather than optical magnification (WASSIE *et al.* 2019), can be successfully applied to diplonemids. ExM readily exposed the reticulated nature of *Diplonema*'s mitochondrion, initially described almost two decades ago (MARANDE *et al.* 2005). In addition, ExM corroborated the reticulated structure of the nuclear heterochromatin (**Supplementary Figure S4D**), which otherwise requires the more resource-demanding TEM.

METHODS

Flow cytometry

D. papillatum was cultured as described earlier (VALACH *et al.* 2014). From ~3×10^6 cells, a nuclei-enriched fraction was recuperated after sucrose gradient centrifugation from the 60% bottom layer and treated with RNase. The diploid *S. cerevisiae* strain BY 4743 (diploid genome size 24 Mbp) was kindly provided by Dr. S. Michnick. Yeast was cultured in a liquid medium containing 0.5 % yeast extract plus 3% glucose and digested with Zymolase to obtain spheroplasts (KISELEVA *et al.* 2007). The *D. melanogaster* male cell line S2 (diploid genome size 360 Mbp) from an asynchronous culture was kindly provided by Dr. V. Archambault. Nuclei from all three eukaryotes were extracted with the Biosciences BD cycleTEST DNA Reagent Kit. We essentially followed the manufacturer's instructions, but for yeast, we ruptured spheroplasts by repeatedly pipetting the suspension up and down through a yellow tip. In addition to the propidium iodine dye included in the kit, the nuclei were stained overnight with DAPI at a final concentration of 1 µg/mL. Fluorescence was measured with a FACScalibur instrument at an FSC setting of 500 V. Approximately 32,000 (yeast), 42,000 (fly) and 100,000 events (*Diplonema*) were collected. Signals from contaminating cytosolic material were excluded based on the scatter signal. The measurements were performed by the flow-cytometry facility of the Institute for Research in Immunology (IRIC) at the Université de Montréal.

Pulsed-field gel electrophoresis and Southern blot hybridization

Separation of chromosomes by pulsed-field gel electrophoresis (PFGE) was performed using a CHEF DR-II Chiller System (Bio-Rad). Briefly, whole cells from an exponential-phase culture were embedded in a 0.7% (w/v) low-gelling temperature agarose solution (50 mM EDTA pH8.0), incubated overnight at 50 °C in a lysis solution (0.45 M EDTA pH8.0, 1% N-lauroylsarcosine, 1 mg/mL proteinase K), and then stored at 5 °C (in 0.45 M EDTA pH8.0, 0.1% N-lauroylsarcosine). DNA (from ~10^8 cells per lane) was then separated in a 1% (w/v) agarose gel in 0.5× TBE buffer at 10 °C and two alternative separation programs (**Supplementary Table S1**). DNA was then blotted onto a nylon membrane (Zeta-Probe, Bio-Rad) *via* capillary alkaline transfer (Brown 2001) and fixed for 60 min at 80 °C. As a probe, we used the oligonucleotide dp350 (**Supplementary Table S2**) that targets telomeric repeats. Hybridization in ULTRAhyb-oligo buffer (Ambion) with radioactively labeled oligo-deoxynucleotides, washing, and detection were performed as detailed previously (Valach et al. 2014).

Electron microscopy and fluorescence microscopy

For fluorescence and electron microscopy, cell growth took place in an artificial sea salt solution (Sigma; 36 g/L) enriched with 1% (v/v) heat-inactivated fetal bovine serum (FBS) and 0.1% (w/v) tryptone (TASHYREVA *et al.* 2018b). Cells were collected at the exponential growth phase by centrifugation at 3,000×g for 15 min and were frozen with a Leica EM PACT2 high-pressure freezer (Leica Microsystems). Successive ultra-thin (100 nm) serial sections were prepared as described elsewhere (YURCHENKO *et al.* 2014). Observations were performed using a JEOL 7401-F microscope at an accelerating voltage of 6 kV. High-resolution micrographs of nuclei were aligned to build 3D reconstructions using the Amira software (Thermo Fisher). Sphericity of nucleus and chromatin was calculated based on their volume and surface area (nucleus ~32

µm^2, chromatin ~87 µm^2). For fluorescence microscopy, cells were harvested from cultures grown for 72 h, fixed for 30 min with 4% paraformaldehyde in artificial seawater, washed in PBS, and allowed to adhere onto poly-L-lysine coated slides. The samples were subsequently mounted in ProLong Gold antifade reagent (Life Technology) containing 4',6-diamidino-2-phenylindole (DAPI) and examined by an Olympus BX53 fluorescence microscope.

Expansion microscopy

Cells were grown as for the other microscopy applications (see above). Typically 10^6 cells were pelleted by centrifugation for 5 min at 800×g, followed by a wash with seawater. For fixation, the cells were re-suspended in seawater and transferred onto a coverslip. The fixation solution containing 4% formaldehyde and 4% acrylamide in seawater was added, and the cells were incubated overnight at room temperature, then washed in seawater. For gelation, the monomer solution containing 19% sodium acrylate, 10% acrylamide, and 0.1% N, N'-methylene bisacrylamide in PBS was mixed with N, N, N0, N0-tetramethylethylenediamine and ammonium persulfate (at a final concentration of 0.5% each). The mixture was quickly transferred onto parafilm in a wet chamber. Coverslips with the cells facing down were placed on the drop of the monomer solution and incubated for 5 min. Subsequently, the wet chamber was transferred to 37 °C and incubated for 30 min to allow gel polymerization. The specimens were then transferred to a well of a 12-well plate containing 1 mL of the denaturation buffer (50 mM Tris-HCl pH9.0, 200 mM NaCl, 200 mM SDS), detached from the parafilm, and then moved to a microcentrifuge tube with additional denaturation buffer and incubated for 1 h at 95 °C. The denatured gels were transferred to Petri dishes and expanded by three 20-min incubations with 15 mL of water. An approximately 10 × 10-mm piece of the gel was cut per staining, performed in the dark while gently rocking. Primary (mouse anti-tubulin) and secondary (anti-mouse) antibodies were diluted in ExM blocking buffer (2% bovine serum albumin in PBS) at 1:50 and 1:1,000 ratios, respectively. The gels were incubated overnight at room temperature and washed 3× for 20 min with water (first with primary, then with secondary antibodies). Selected specimens were also incubated for 30 min at room temperature with 10 µg/mL of 40 ,6-diamidino-2-phenylindole (DAPI) in PBS. Finally, the gels were washed 5× for 15 min with 4 mL of water. A stained piece of the gel was transferred to the center of a glass-bottom dish coated with poly-L-lysine. Specimens were imaged using a Leica TCS SP8 confocal microscope with an HC PL apochromatic 150x/NA 1.40 oil immersion objective. Excitation was performed with a 405 nm diode laser (50 mW) in the case of DAPI and Hoechst 33342 and a 488 nm solid-state laser (20 mW) in the case of Alexa Fluor 488; emissions were detected using hybrid detectors (HyD). Z-stacks were acquired with a step size of 100 nm (without averaging). The pixel size and dwell time were between 52 and 97 nm and 400 ns, respectively. The size of the pinhole was adjusted based on the signal strength, but was typically around 0.4 Airy unit to improve resolution.

Supplementary Table S1. Parameters of PFGE separation programs.

Program	Step	Switch duration	Angle	Electric field strength	Run time
#1	1	1,200 s	94°	2.0 V/cm	30 h
	2	1,500 – 12,000 s	106°	1.5 V/cm	90 h
	3	120 – 480 s	120°	2.5 V/cm	22 h
#2	1	120 – 500 s	100°	3.5 V/cm	10 h
	2	500 s	106°	3.0 V/cm	40 h

Supplementary Table S2. Oligonucleotides used in this study.

Oligonucleotide	Sequence (5′ → 3′)	Target
dp350	CAAACCCGCAAACCCGCAAACCCGCA	telomeric repeat
dp236	ACACGCACCCTATGAGCTTAGCATTGGTAGT	cox1-m1

ACKNOWLEDGEMENTS

We thank Shona Teijeiro (UdeM) for performing the flow-cytometry experiments, Dr. Eva Doleželová (Institute of Parasitology, Biology Centre, Czech Academy of Sciences, Ceske Budejovice, Czech Republic) for critical discussion of the results, Dr. Stephen Michnick, Dr. Vincent Archambault, and Louise Cournoyer for providing cell lines and strains that were used as references in the flow cytometry experiments, and Monique Vasseur (all from UdeM) for assistance in nuclei staining.

AUTHOR CONTRIBUTIONS

Conceptualization, Data curation, Formal analysis, Investigation, Visualization, Writing, original draft – G.P., P.T., G.B., M.V.; **Writing, review & editing** – all co-authors.

REFERENCES

Angeler, D. G., A. N. Muellner and M. Schagerl, 1999 Comparative ultrastructure of the cytoskeleton and nucleus of *Distigma* (Euglenozoa). Eur J Protistol 35**:** 309-318.

Breglia, S. A., N. Yubuki, M. Hoppenrath and B. S. Leander, 2010 Ultrastructure and molecular phylogenetic position of a novel euglenozoan with extrusive episymbiotic bacteria: *Bihospites bacati* n. gen. et sp. (Symbiontida). BMC Microbiol 10**:** 145.

Elias, M. C., M. Faria, R. A. Mortara, M. C. Motta, W. de Souza *et al.*, 2002 Chromosome localization changes in the *Trypanosoma cruzi* nucleus. Eukaryot Cell 1**:** 944-953.

Ersfeld, K., 2011 Nuclear architecture, genome and chromatin organisation in *Trypanosoma bruce*i. Res Microbiol 162**:** 626-636.

Kiseleva, E., T. D. Allen, S. A. Rutherford, S. Murray, K. Morozova *et al.*, 2007 A protocol for isolation and visualization of yeast nuclei by scanning electron microscopy (SEM). Nat Protoc 2**:** 1943-1953.

Lukeš, J., B. Kaur and D. Speijer, 2021 RNA editing in mitochondria and plastids: weird and widespread. Trends Genet 37**:** 99-102.

Lukeš, J., R. Wheeler, D. Jirsová, V. David and J. M. Archibald, 2018 Massive mitochondrial DNA content in diplonemid and kinetoplastid protists. IUBMB Life 70**:** 1267-1274.

Marande, W., J. Lukeš and G. Burger, 2005 Unique mitochondrial genome structure in diplonemids, the sister group of kinetoplastids. Eukaryot Cell 4**:** 1137-1146.

O'Donnell, E. H. J., 1965 Nucleolus and Chromosomes in *Euglena gracilis*. Cytologia 30**:** 118-154.

Patterson, D. J., 1999 The diversity of eukaryotes. Am Nat 154**:** S96-s124.

Porter, D., 1973 *Isonema papillatum* sp. n., a new colorless marine flagellate: a light- and electronmicroscopic study. J Protozool 20**:** 351-356.

Prokopchuk, G., D. Tashyreva, A. Yabuki, A. Horák, P. Masařová *et al.*, 2019 Morphological, ultrastructural, motility and evolutionary characterization of two new Hemistasiidae species. Protist 170**:** 259-282.

Roy, J., D. Faktorová, J. Lukeš and G. Burger, 2007 Unusual mitochondrial genome structures throughout the Euglenozoa. Protist 158**:** 385-396.

Schenkman, S., S. Pascoalino Bdos and S. C. Nardelli, 2011 Nuclear structure of *Trypanosoma cruzi*. Adv Parasitol 75**:** 251-283.

Tashyreva, D., G. Prokopchuk, J. Votýpka, A. Yabuki, A. Horák *et al.*, 2018a Life cycle, ultrastructure, and phylogeny of new diplonemids and their endosymbiotic bacteria. MBio 9**:** e02447-02417.

Tashyreva, D., G. Prokopchuk, A. Yabuki, B. Kaur, D. Faktorová *et al.*, 2018b Phylogeny and morphology of new diplonemids from Japan. Protist 169**:** 158-179.

Triemer, R. E., 1992 Ultrastructure of mitosis in *Diplonema ambulato*r Larsen and Patterson (Euglenozoa). Eur J Protistol 28**:** 398-404.

Triemer, R. E., and M. A. Farmer, 1991 An ultrastructural comparison of the mitotic apparatus, feeding apparatus, flagellar apparatus and cytoskeleton in euglenoids and kinetoplastids. Protoplasma 164**:** 91-104.

Valach, M., S. Moreira, G. N. Kiethega and G. Burger, 2014 Trans-splicing and RNA editing of LSU rRNA in *Diplonema* mitochondria. Nucleic Acids Res 42**:** 2660-2672.

Vickerman, K., and T. M. Preston, 1970 Spindle microtubules in the dividing nuclei of trypanosomes. J Cell Sci 6**:** 365-383.

Wassie, A. T., Y. Zhao and E. S. Boyden, 2019 Expansion microscopy: principles and uses in biological research. Nat Methods 16**:** 33-41.

Yubuki, N., V. P. Edgcomb, J. M. Bernhard and B. S. Leander, 2009 Ultrastructure and molecular phylogeny of *Calkinsia aureus*: cellular identity of a novel clade of deep-sea euglenozoans with epibiotic bacteria. BMC Microbiol 9**:** 16.

Yurchenko, V., J. Votýpka, M. Tesarová, H. Klepetková, N. Kraeva *et al.*, 2014 Ultrastructure and molecular phylogeny of four new species of monoxenous trypanosomatids from flies (Diptera: Brachycera) with redefinition of the genus *Wallaceina*. Folia Parasitol (Praha) 61**:** 97-112.

2. Assembly and annotation of the nuclear genome and transcriptome of *Diplonema papillatum*

INTRODUCTION

Despite the recent development of powerful algorithms, the *de novo* assembly of genomes can be difficult. The main challenge encountered when assembling the *D. papillatum* nuclear genome was the large proportion of long repeat regions. In contrast, the assembly of transcripts was straightforward, because these repeat regions are not transcribed (at least in the conditions under which the protist was cultured). Structural annotation was also complicated by repeats, because RNA-Seq reads that otherwise assist in the selection of the most probable gene structure may be misleading if they can align with multiple locations in the genome. To assess the correctness of gene models predicted by the automated pipeline, the three largest contigs were manually curated by an expert. Finally, predicted protein-coding genes were assigned a molecular function by annotation transfer from model organisms to the examined species based on sequence similarity. Due to the high divergence of *Diplonema* sequences and the absence of closely related model organisms, we are left without any functional information for more than half of the gene models.

RESULTS AND DISCUSSION

Genome assembly and polishing

The overall approach involved five steps: (i) generation of short and long reads; (ii) correction of long reads, which have a high error rate, with short reads; (iii) separate assembly of the two read types; (iv) merging of contigs from the two assemblies; and (v) detection and splitting of misassembled contigs.

More specifically, we generated 462 million pairs of short reads (Illumina) and ~725,000 long reads (PacBio) totalling 126.4 Gbp raw data (**Supplementary Table S3**). Short and long reads were assembled separately with the Celera Assembler (MYERS *et al.* 2000) and Canu (KOREN *et al.* 2017). The 'primary merged assembly' was compiled by complementing the PacBio contigs with those contigs from the Celera assembly that are not included in the former, and the resulting contig set was deduplicated to yield the 'primary merged deduplicated (PMD) assembly' (**Supplementary Figure S5**). We noted a low rate of contig removal by the latter step, which is indicative for a genome of very low heterozygosity (i.e., haploid or autodiploid; see **Supplementary Information**: Section 3. The ploidy level of *Diplonema papillatum*).

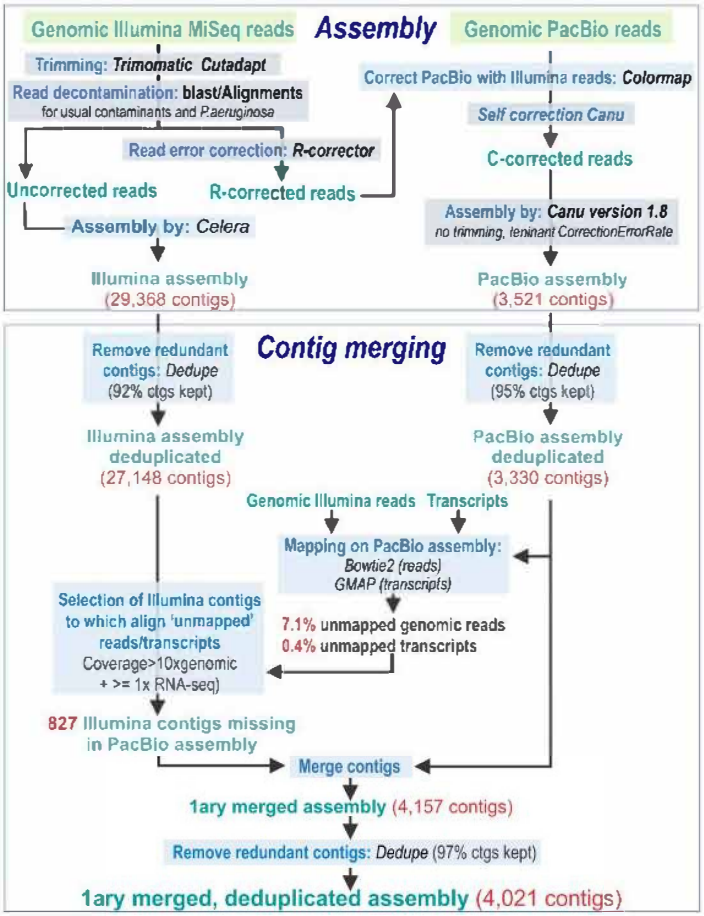
The PMD assembly was further processed to split contigs that arose by incorrectly assembled reads. Such erroneous joining is due to a low percentage of chimeric reads, i.e., reads originating from unrelated DNA fragments that were ligated together during library construction. Erroneous joining sites were detected by low read coverage after aligning short and long reads with the contigs of the PMD assembly. Unless high-confidence transcripts spanned such a trough in coverage of both short and long reads, contigs were split at these positions and contig portions smaller than 100 bp removed (**Supplementary Figure S6**). Note that a more conventional scaffolding approach, which generally produces a more contiguous assembly, was not implemented, because such assemblies contained an even higher number of obviously chimeric contigs than the PMD assembly.

The final product, the *D. papillatum* nuclear genome assembly version 1.0 described here, comprises 6,181 contigs ≥200 bp, has a length of 280,293,864 bp and an N50 value of 190,080 bp (**Supplementary Table S4**).

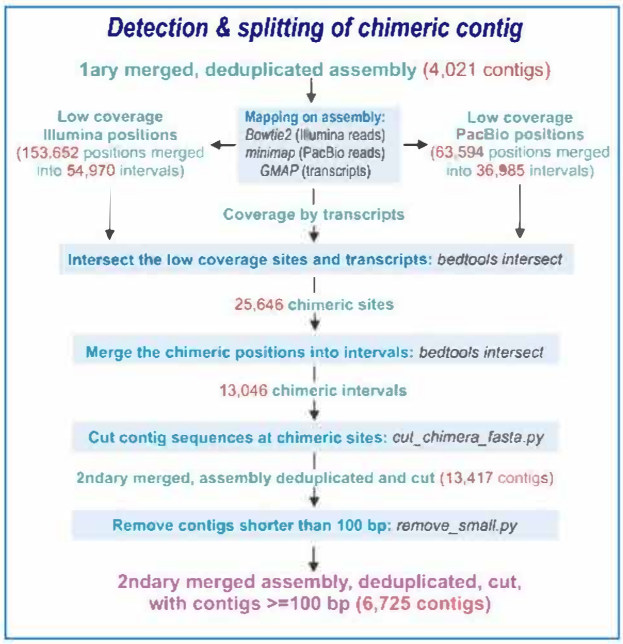
Supplementary Table S3. Libraries and reads used in this study.

Library name	Material	Prep kit (approx. insert size)	Techno-logy[a]	Service provider	SRA ID	Nr. of raw reads[b]	Read Length
Dp-nucDNA	Total DNA	Illumina TruSeq DNA (2–8 kbp)	HiSeq MP	Genome Quebec	/[c]	17,772,001	150 nt
Dp_301113	Total DNA	Illumina TruSeq DNA (0.5 kbp)	MiSeq PE	Genome Quebec	SRR21741423	5,740,630	250 nt
Dp_301113-2nd	Total DNA	Illumina TruSeq DNA (0.5 kbp)	MiSeq PE	Genome Quebec	SRR21741422	20,139,574	250 nt
Dp_HiSeq_Dp_gDNA	Nuclear DNA	Eurofins (3 kbp)	HiSeq PE	Eurofins	SRR21741421	13,305,159	100 nt
Dp_MiSeq_Dp_gDNA	Nuclear DNA	Eurofins (8 kbp)	MiSeq PE	Eurofins	SRR21741420	18,367,804	250 nt
PacBio-20kbp	Size-selected DNA (>20 kbp)	BluePippin+SMRTbell (PacBio Large Insert)	Sequel	Genome Quebec	SRR21741419	149,527	2,500 nt[d]
Dp_PacBio	Size-selected DNA (>20 kbp)	BluePippin+SMRTbell (PacBio Large Insert)	Sequel II	Takara Bio	SRR21741418	575,764	8,500 nt[d]
PA	Poly(A) RNA	Epicentre ScriptSeq RNA (0.2 kbp)	HiSeq PE SS	Macrogen Korea	SRR21741417	30,618,141	100 nt
DPA2	Poly(A) RNA	Illumina TruSeq RNA (0.25 kbp)	HiSeq PE SS	Macrogen Korea	SRR21741416	261,280,954	125 nt
Dp_RNASeq_Dpa	Poly(A) RNA	Illumina TruSeq RNA (0.25 kbp)	MiSeq PE SS	Novogene	SRR21741415	67,496,825	150 nt

[a] MP, mate-paired (long inserts); PE, paired-end (short inserts); SS, strand-specific.
[b] Read pairs in the case of Illumina reads; single reads in the case of PacBio reads.
[c] Not used in the assembly.
[d] Average read length (up to ~45,000 nt).



Supplementary Figure S5. Procedure of nuclear genome assembly and contig merging. The procedure consists of the '*Assembly*' phase in which short (Illumina MiSeq) and long (PacBio) reads are assembled separately. In the subsequent '*Contig merging*' phase, genomics reads and transcripts (the latter from RNA-Seq assembly) were identified that do not align to the PacBio assembly. Then, the contigs in the Illumina assembly were selected to which these sequences could be mapped. These Illumina contigs were merged with the PacBio contigs and then deduplicated. Subsequent polishing steps are depicted in Supplementary Figure S6.



Supplementary Figure S6. Procedure of assembly polishing. Potentially mis-assembled (chimeric) contigs are recognized by a sudden drop in read coverage. The screen for low-coverage (LC) genomic positions is performed separately for short and long reads. Adjacent LC positions are merged into intervals. Then, shared LC intervals from Illumina and PacBio contigs are determined and compared with the locations where transcripts align to the genome. Only LC intervals that are not spanned by transcripts are further processed. In the next step, adjacent LC intervals are again combined, and the contigs are then split at the interval borders. Contig pieces shorter than 100 bp are eliminated to yield the final assembly.

Transcriptome assembly

Transcript sequences were obtained by *de novo* assembly of about 645 million reads from the two strand-specific poly-A RNA libraries PA and DPA2 (see **Supplementary Table S3**), using the Trinity software. The inferred transcriptome contains nearly 200,000 contigs (transcripts) with a cumulative length of ~16 Mb (**Supplementary Table S4**). Transcript sequences assisted in the structural genome annotation, and a subset of the sequences referred to as 'high-confidence transcriptome' (mRNAs) were employed for detecting mis-joined contigs (see section Genome assembly). The nearly 22,000 high-confidence mRNAs were selected based on the presence of a Spliced-Leader (SL) sequence at their 5′ end and a minimum overall read coverage of 100.

As a measure of genome completeness complementary to the BUSCO benchmarking (see below), we determined the rate of mapping transcriptome-derived reads to the genome sequence. Out of 174,212,093 RNA-Seq reads >40 nt (after contaminant filtering, as well as quality and adaptor trimming), 96.8% mapped onto the genome. Of the remaining reads, 2,608,091 (representing 1.5%) mapped exclusively onto the transcriptome assembly (usually onto transcripts split across multiple genomic contigs), while 2,930,982 reads (representing 1.7%) could not be mapped onto either assembly. Many of the unmapped reads contained homopolymer tracts (e.g., poly-A) and low-complexity repeats. These observations suggest that the *D. papillatum* genome assembly version 1.0 contains essentially all actively transcribed regions.

Supplementary Table S4. Statistics of genome and transcriptome assemblies.

Assembly	Total length	Nr. of contigs	Contig size range	N50	Average A+T content (range)
Genome Dp_v1.0	280,293,864 bp	6,181	200 - 1,009,103 bp	190,080	44.8% (20.0%-75.6%)
Transcriptome (*de novo*)	161,272,633 nt	194,546	182 - 43,150 nt	/	43.8% (16.4%-73.3%)
High-confidence transcriptome[a]	46,055,449 nt	21,747	298 - 43,118 nt	/	40.9% (22.6%-68.6%)

[a]Minimum read coverage per contig 100x; SL sequence at 5'end (see text)

Structural genome annotation

Structural genome annotation, also referred to as gene-model prediction, denotes the identification of genome regions that have the potential to code for proteins or specify structural (non-(protein-)coding) RNAs. Our automated annotation pipeline constructed about ~37,000 gene models, predominantly protein-coding gene models (**Supplementary Table S5, S6**). Most *D. papillatum* protein-gene model structures are simple, lacking introns and isoforms. Furthermore, most models are

complete with standard start/stop codons. Only a small minority appear to be incomplete either at the 5′ (~2.5%) or the 3′ (~2.5%) end, and an even smaller proportion (~1.5%) lack both start and stop codons.

To further assess the completeness of protein-coding gene models, their respective conceptually translated sequences were benchmarked against the so-called 'universal single-copy ortholog' set of proteins (BUSCO (SEPPEY *et al.* 2019)) using v10 of the Eukaryota OrthoDB (built 2019-11-20 (KRIVENTSEVA *et al.* 2019) (**Supplementary Table S7, S8**). The results indicated that 84% of models are complete with ~17% in duplicate, whereas only 4% were reported as fragmented, and 11% (29) as apparently missing. As for the models reported as duplicated, around 1/3 are exact copies and thus might be due to either recent genomic duplications or assembly errors, whereas non-identical duplicates most likely represent paralogs.

A closer inspection of the orthologs for which BUSCO reported missing matches revealed that four are also absent in other diplonemids and one is missing in the sister group, *Kinetoplastida*, whereas 16 are in fact present in *D. papillatum* but are divergent to the degree that falls short of BUSCO's cut-off criteria. Eight BUSCO Hidden Markov Models (HMMs) have inconclusive hits that would need further verification. In summary, BUSCO completeness is above 90% when excluding models known to be missing from Euglenozoa. Since detecting the remaining ≤10% of BUSCO orthologs appears to be hampered by extreme sequence divergence, potential assembly errors, or incorrect gene modelling, we consider the *D. papillatum* genome assembly version 1.0 as quasi complete.

Structural annotation also included the prediction of spliced-leader trans-splicing sites (for details see below), which were pinpointed by aligning RNA-Seq reads to the genome assembly and determining the positions where soft-clipped portions of the reads consist of the spliced leader sequence.

Supplementary Table S5. Summary of structural genome annotation.

Prediction type	Initial, fully automated annotation	Annotation partially expert-curated
Protein-coding genes	35,978	37,054
Non-coding RNAs	7,362	573
rRNA[a]	5,111[b]	151
tRNAs[c]	375	211
Small nucleolar (sno) RNAs[a]	9	9
Spliceosomal (sn) RNAs[a]	202	202
MicroRNAs	1,250	0
Other RNAs[a,e]	417	0
Total	43,340	37,627

[a] Identified by Rfam HMM
[b] Includes rRNAs fragments
[c] Identified by tRNAScan-SE
[d] includes pseudo tRNAs, and those with undetermined anticodon
[e] HDV ribozymes, rsrG Hfq binding RNA, etc.

Supplementary Table S6. Summary of protein-coding gene models.

Feature	Status	Count (percent)	Average length	Median length	Average coding length	Median coding length
Transcripts	Complete	36,763 (94%)	4,044 bp	2,424 bp	1,699 bp	1,152 bp
	Incomplete	2,354 (6%)	4,381 bp	2,070 bp	2,126 bp	1,206 bp
	Single exon	21,485 (55%)	1,823 bp	1,333 bp	1,318 bp	927 bp
	All	39,117	4,065 bp	2,409 bp	1,724 bp	1,155 bp
Exons	Initial	15,259	832 bp	516 bp	/	/
	Internal	23,016	616 bp	284 bp	/	/
	Terminal	15,113	593 bp	279 bp	/	/
	Single	21,384	1,319 bp	927 bp	/	/
	All	74,764	857 bp	472 bp	/	/
UTRs[a]	3′ UTR	21,896	1020 bp	777 bp	/	/
	5′ UTR	20,251	144 bp	66 bp	/	/
Introns		36,604	1,528 bp	671 bp	/	/

[a] Statistics were calculated based on annotated UTRs, i.e., genes that lack UTR annotations were ignored.

Supplementary Table S7. BUSCO report.

BUSCO report for assembly version 1.0	Nr. of models reported			
	Complete	duplicate	fragmented	missing
	215/255 (84.3%)	42 (16.5%)	11 (4.3%)	29 (11.4%)

Supplementary Table S8. BUSCO benchmarking of assembly completeness.

Busco category	Total (percentage)
Complete	215 (84.3%)
Complete and single-copy	173 (67.8%)
Complete and duplicated	42 (16.5%)
Fragmented	11 (4.3%)
Missing	29 (11.4%) [a]
Total	255 (100%)

[a] Manual curation showed that 16 of these are present in *D. papillatum* but divergent, and that 4 are absent from all 11 diplonemids, for which RNA-Seq data are available.

Quality assessment of the automated structural annotation

Models of genes predicted to encode a protein delineate exons, introns, and 5′ and 3′ UnTranslated Regions (UTRs). Here we assess the accuracy of predicted gene models *via* inspection by an expert. The structural annotation of the automated pipeline was validated by human curation of the three longest contigs (tig00022654_12, tig00022740_1, and tig00022679). Together these contigs represent ~1% of the total assembly and contain 319 predicted protein-coding genes (**Supplementary Table S9**). As detailed in the following, inaccuracies in automated gene-model prediction include omissions, false positives, and incorrect UTR termini and exon-intron boundaries. The underlying causes for these shortcomings will be discussed.

5′ UTRs. The nuclear gene expression of diplonemids (and kinetoplastids) relies on the addition of a short sequence called spliced leader (SL) upstream of a primary transcript containing a coding sequence (CDS) to generate the 5′ end of the mRNA (STURM *et al.* 2001; CLAYTON 2019). SL addition takes place by a trans-splicing reaction that is catalyzed by the nuclear spliceosome. Several inspected gene models were lacking a 5′ UTR annotation, which was generally due to the concealed Spliced Leader (SL) Trans-Splicing (SLTS) acceptor site upstream of the predicted ORF. Visual inspection revealed spurious, short matches of the 3′ moiety of the SL to other locations upstream of the gene, thus counteracting the diagnostic soft-clipping at the SLTS site. The RNA-Seq read-mapping algorithm considers such matches valid when the interval is flanked by the GT and AG dinucleotides to model an intron. This mapping artifact can misplace the inferred SLTS-site up to a dozen kbp upstream or impede identification of the splice site and the 5′ UTR. In all validated cases, the matches of the SL to upstream regions were short (≤6 nt), making alternative splicing unlikely.

3′ UTRs. The structural annotation algorithm was trained to place a gene's end, i.e., the 3′ terminus of its 3′ UTR, at the position where a transcript aligning with the genomic region ended. However, we observed a sharp drop in RNA-Seq read coverage downstream of many CoDing Sequences (CDSs) coinciding with low complexity or highly repetitive regions, followed further downstream by a coverage similar to the CDS. In most such instances, the region immediately downstream of the annotated CDS was predicted to contain a short ORF without a SLTS acceptor site, strongly suggesting that these ORFs were spurious. We reasoned that the fluctuating RNA-Seq read coverage indicated a 3′ UTR extending beyond the pipeline-assigned end and that the premature positioning of the gene termini was actually due to artificially shortened transcripts. Indeed, during transcriptome assembly by the Trinity software (GRABHERR *et al.* 2011), transcript sequences are not further elongated in two instances: (i) when the number of reads is high with no clear path along the assembly graph e.g., in highly repetitive regions; or (ii) when the number of reads is low, e.g., in a homopolymeric stretch that is not efficiently sequenced. The gene model DIPPA_17048 is a typical example. The region covered by the transcript sequence terminates with a succession of ~50 bp-long T, G and C homopolymeric runs forming an ~6.7 kbp-long 3′ UTR. However, downstream of the predicted end of the gene, a reasonable coverage of RNA-Seq read extends for an additional ~2.5 kbp. Thus the 3′ UTR is considered to have a total length of ~9.2 kbp, which is 2.5 kbp longer than predicted by the annotation pipeline.

Supplementary Table S9. Expert validation of gene-models from the three largest contigs.

			Contig ID		
		Total	tig00022654_12	tig00022740_1	tig00022679
Gene models	All [a]	334	131	108	95
	Before curation	319	123	104	92
	UTRs, introns corrected [b]	125	47	38	40
	New [c]	15	8	4	3
	False positive [d]	68	15	32	21
	After curation, unsupported [e]	34	10	18	6
	After curation, supported	232	107	57	68
Corrections	5′ UTR [f]	51	18	9	24
	3′ UTR [g]	101	38	32	31
	Introns	23	7	9	7
Gene models with spliced leader	Automated detection [h]	181	90	52	39
	Addit. Expert detection [i]	27	11 (10+1)	4 (1+3)	12 (3+9)
	No evidence [j]	18	6 (3+3)	1 (0+1)	11 (5+6)
SLTS sites	Genes with multiple sites [k]	30	14	11	5

[a] The sum of all processed gene models combining those prior to curation and the newly added ones.
[b] UTR boundaries and/or introns were corrected.
[c] Unrecognized by the automated annotation pipeline, but evident from transcriptome data.
[d] Models that became obsolete, because e.g., located within a newly recognized intron.
[e] Models that are typically a part of a transposable or other repetitive element and without detectable SLTS site or RNA-Seq read coverage.
[f] Generally, the UTR has been shortened (see text).
[g] In most cases, the UTR has been considerably extended (see text).
[h] Recognized by the soft-clipped spliced-leader (SL) sequence of ≥11 bp in RNA-Seq reads; only counting a single, primary site per gene model.
[i] Failure to detect the site by the automated approach was due to (i) misaligned reads, or (ii) soft-clipped SL segments of ≤10 bp.
[j] Soft-clipped SL sequences are absent, while the gene has RNA-Seq read coverage. The failure to detect the SL may be due to repeats or low-complexity regions upstream of the CDS impeding correct alignment of RNA-Seq reads. In addition, the gene's 5′ UTR region may have a low read coverage.
[k] Exclusively sites reported by the automated procedure were counted.

Repercussions of repetitive genome regions. Repeats, and tandem repeats in particular, represented a major source of difficulty for the annotation algorithm caused by the incorrect alignment of transcripts and RNA-Seq reads to the genome assembly. This problem affected all gene features, but most frequently led to the annotation of spurious introns and to missed genes. For example, two genes in a tandem arrangement were often annotated as a single gene. We also detected multiple cases where tandemly repeated portions of a CDS flanked by different UTRs or including different intron sequences were annotated as a single gene by combining the 5′ moiety of the upstream CDS and 3′ moiety of the downstream CDS. DIPPA_17054 represented such a case. Curation consisted in splitting off its 3′ portion, from which the new gene model DIPPA_70071 was generated. In all such instances the accuracy of the genome assembly in the corresponding region was verified by inspecting the mapped PacBio reads.

Missing gene models. Rarely, protein-coding genes were not recognized by the annotation pipeline despite unequivocal evidence from transcript (RNA-Seq) data, and the absence of other issues such as complex repeats. Missing models tended to cluster within contigs; the largest cluster comprised 11 consecutive unidentified gene models in contig tig00023309_11. The reason for this problem remains unclear. Nevertheless, most of these omissions could be rectified by a post-processing procedure that used transcripts carrying a SL sequence at their 5′ ends as a lead.

Reconstructed transcripts absent from the genome. A total of 18,480 reconstructed transcripts were found to contain ORFs preceded by a splice-leader sequence. Surprisingly, a considerable portion (~20%) of those transcripts were not identified in the genome assembly. Genome mis-assemblies may partially explain this problem since subsequences of some missing transcripts were identified across two or more contigs. For downstream analyses, we established a comprehensive sequence collection by combining into a non-redundant set all the missing transcripts with the manually curated set and all remaining protein-coding genes automatically generated by the pipeline.

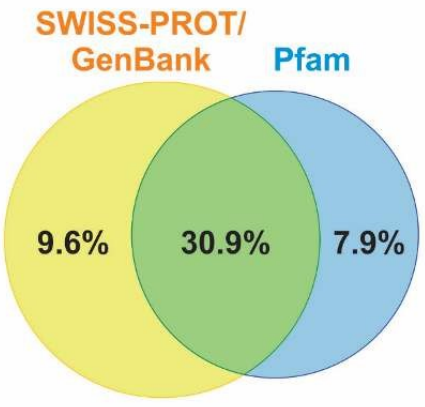
Genes split by genomic sequence gaps. An issue encountered in contigs other than the three longest were genes artificially split within their intron by a sequence gap in the scaffold. This gap led the annotation pipeline either to infer an incorrect CDS terminus, or to omit the annotation altogether. In all investigated cases, the problem was due to highly repetitive sequence elements in the corresponding introns.

Alternatively spliced introns. Only a few genes with alternatively spliced introns were identified in the validated contigs, at least under the growth conditions used to prepare the RNA-Seq libraries. Therefore, a random sample of intron-containing genes models were examined (DIPPA_00001 to DIPPA_03296). Among these, 11 cases had convincing support by RNA-Seq data for alternative splicing. Several genes with confirmed alternative splicing combine more than one splicing type. For example, the expression of gene DIPPA_00069 (tig00013476_1) involves both exon skipping and intron retention, while that of gene DIPPA_03285 (tig00022668_1) implicates in addition alternative splice-site selection.

Conclusion. The performance of the structural annotation pipeline used in this study ranks among the best available tools. Still, among the set of expert-inspected gene models, about ¼ had incorrectly placed UTRs or (more rarely) exon/intron boundaries, another ¼ were false positives, and ~4% were not retrieved. As discussed above, these shortcomings are mostly due to the high proportion of long sequence repeats in the *Diplonema* nuclear genome. While the automated annotation provides a fair picture of the *Diplonema* nuclear genome, thorough expert validation is warranted should seemingly unusual or unique genes and gene features be encountered.

Functional genome annotation

Protein-coding genes. As shown in Supplementary Figure S7, about 51% of all gene models were assigned functional information by either sequence similarity to the SWISS-PROT database and the downloaded GenBank accessions, or probabilistic HMM searches against Pfam v31.0, or both. The products of the gene models without SWISS-PROT information were labelled as 'hypothetical proteins'. Most models with assigned functional information have significant hits by both similarity to protein sequence databases and to protein domains. As there are currently few publicly available sequences from Discoba species, and those that are available come mostly from derived kinetoplastids (e.g., *Trypanosoma*), neither taxonomically broad nor Discoba-specific profile HMMs can be built for all nucleus-encoded proteins. This implies that functional information transfer must be made through less sensitive searches, notably searches for sequence similarity and conserved protein domains.



Supplementary Figure S7. **Summary of gene models assigned a function.** Functional information was assigned to a total of 19,078 protein-coding gene models. SWISS-PROT/GenBank (yellow plus green intersection) represents the portion with explicit information about the gene product. Pfam (blue plus green intersection) represents the portion that includes a conserved protein domain.

Non-coding (structural) RNAs. Transfer RNA genes were searched with tRNAscan-SE, and rRNAs, snoRNAs and microRNAs with Hmmerscan using profile HMMs from Rfam (GRIFFITHS-JONES *et al.* 2003). The employed version of the annotation procedure did not report if a gene is incomplete, which explains the large number of rRNA gene models initially reported, which was corrected secondarily (see **Supplementary Table S5**).

METHODS

Genome assembly

Read cleaning. For the genome assembly, we first cleaned Illumina and PacBio reads. We used cutadapt v1.2.1 (http://journal.embnet.org/index.php/embnetjournal/article/view/200) on MiSEQ reads to clip adaptors (-e 0.1, -O 10 -m 20) and then trimmed low-quality sequences at the ends of reads (-q 20 -m 20). Reads with more than three non-defined bases ('N') in a row were discarded. PacBio reads were self-corrected and quality trimmed with Canu (KOREN *et al.* 2017) and then corrected with CoLoRMap (HAGHSHENAS *et al.* 2016) (default options) by mapping onto these long reads a set of Illumina reads equivalent to 50X of the genome coverage. Using the CoLoRMap program suite, we applied a scrupulous two-step correction, first using the shortest-path algorithm and second the One-End-Anchor based algorithm. The quality of libraries was evaluated along the cleaning process with Prinseq v.0.20.4 (SCHMIEDER AND EDWARDS 2011) and FastQC (EDWARDS, 2010, https://www.bibsonomy.org/bibtex/2b6052877491828ab53d3449be9b293b3/ozborn).
Read decontamination. Illumina and PacBio reads were screened for potential contaminations with sequences from other organisms and organelles. For that, we generated a crude assembly. Illumina reads were assembled with the Celera package (DENISOV *et al.* 2008) executed by the script runCA release 8.3.rc2 (http://wgs-assembler.sourceforge.net/wiki/index.php/RunCA), and long reads with Canu. We downloaded 16S and 18S rRNA sequences from the SILVA database (QUAST *et al.* 2013) (file SILVA_128_SSURef_Nr99_tax_silva.fasta) and performed a BLAST search against the provisional assemblies, shaft contigs and remaining non-assembled reads. We selected all hits with an E-value larger than 1.0e-2, a length above 300 bp, and a similarity >80%. Only four species remained after the filtering: *Pseudomonas fluorescens*, an uncultured *Bacillus* sp., *Candidatus amoebophilus*, and *Streptococcus agalactiae*. To the list of possible contaminants were further added the phage PhiX, which is used as a spike-in during sequencing, *E. coli* and *Homo sapiens*. The complete genome sequence of all contamination candidates was downloaded from NCBI. *Diplonema* reads from all genomic libraries were mapped onto the 'contaminant' genome sequences with Bowtie2 v.2.3.4 (LANGMEAD AND SALZBERG 2012) in local mode (--local), and aligning reads (plus their mates) were discarded. As a final decontamination step, we removed reads originating from the mitochondrial genome. Because previous analyses suggested that some mitochondrial sequences may be present in the nuclear genome (NUMTs), we discarded read pairs where both mates aligned with mitochondrial sequences (GenBank acc. Nos. EU123536-8 and HQ28819-33) after mapping with Bowtie2. Among unpaired reads, we removed those whose sequence identity with mtDNA was above 95%.
Genome assembly. Canu was used with parameters specifying the expected error rate and not allowing the trimming of read ends (errorRate=0.035, corMinCoverage=0). For Celera we used parameters (defined in the specification file) to fit the memory and threads available on our server and to set error correction thresholds (cnsErrorRate=0.10, ovlErrorRate=0.10). Duplicated contigs were removed with dedupe2.sh (https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/dedupe-guide/, http://sourceforge.net/projects/bbmap/), setting the parameters as follows: maxedits=10 (allows up to 10 substitutions or indels), minidentity=90 (absorbs contained sequences with at least 90% identity), findoverlap=t (finds overlaps between contigs) and cluster=t (groups overlapping contigs into clusters); the other parameters were default. The command used was:

```
$ dedupe2.sh in=genome.fasta out=genome_dedup.fasta threads=4 absorbrc=t absorbmatch=t absorbcontainment=t
overwrite=true maxedits=10 minidentity=90 findoverlap=t cluster=t outd=genome_del.fasta
```

As expected, the long-read (Canu) assembly was more contiguous, while the short-read assembly (Celera) was more complete as assessed by the number of transcripts mapped to the contigs.

Merging of the short- and long-read assemblies. The two assemblies were merged by adding to the long-read assembly those contigs that were exclusively present in the short-read assembly. To identify contigs missing in the long-read assembly, we mapped short reads and selected transcripts against the long-read assembly. The mapping of the Illumina reads was performed using Bowtie2 with specific parameters (--end-to-end –no-unal –un-conc –no-mixed) to extract the unmapped reads.

1. Build genome index and map reads:

```
$ bowtie2-build -f PacBioAssembly_dedup.fasta PacBioAssembly_dedup
$ bowtie2 --end-to-end --no-unal -p 40 --un-conc MiSeq.unmapped --no-mixed -x PacBioAssembly_dedup -1
MiSeqReads_1.fastq.gz -2 MiSeqReads_2.fastq.gz -U MiSeqReads_S.fastq.gz 1> MiSeqReads.mapped.sam 2>
```

```
MiSeqReads.mapped.log | sambamba view -S -f bam /dev/stdin | sambamba sort /dev/stdin -o
MiSeqReads.mapped.bam ; sambamba index MiSeqReads.mapped.bam
```

Transcripts were mapped to the long-read assembly, and unmapped transcripts collected.

1. Build genome index and map transcripts with the split-read aligner gmap version 2017-11-15 (WU AND WATANABE 2005):

```
$ gmap_build -D . -d PacBioAssembly_dedup PacBioAssembly_dedup.fasta > gmapbuild.out 2>&1
$ gmap --format gff3_gene --nofails -t 12 --min-identity=0.90 --min-intronlength=20 --gff3-add-
separators=0 -D . -d PacBioAssembly_dedup/ transcripts.fasta > ranscrpits.mapped.gff3 2>
transcripts.mapped.log
```

2. Extract the complete transcripts list:

```
$ grep ">" transcripts.fasta | cut -d " " -f 1 | cut -d "ﾝ -f2 | sort > transcripts.list
```

3. Extract the list of mapped transcripts:

```
$ awk 'NR>2 {if($3=="gene") {split($9, a, ";"); split(a[1], b,"="); split(b[2], c,"."); print c[1]}}'
transcripts.mapped.gff3 | sort | uniq > transcripts.mapped.list
```

4. Extract the list of unmapped transcripts:

```
$ awk 'NR==FNR{a[$0]=1;next}!a[$0]' transcripts.mapped.list
 transcripts.list > transcripts.unmapped.list
```

5. Retrieve transcripts that do not align with long-read contigs:

```
$ xargs samtools faidx file.fasta < transcripts.unmapped.list > transcripts.unmapped.fasta
```

To identify the short-read contigs that were absent from the long-read assembly, we mapped the Illumina reads that did not align with the long-read assembly, against the short-read assembly.

1. Build genome index and map reads with Bowtie2 in a way that labels non-primary-mapping reads. Then extract a list of mapped reads sorted by the number of occurrences per contig:

```
$ bowtie2-build -f MiSeqAssembly_dedup.fasta MiSeqAssembly_dedup
$ bowtie2 -p 40 -x MiSeqAssembly_dedup -1 MiSeq.unmapped.1.fastq.gz -2 MiSeq.unmapped.2.fastq.gz --no-unal
-k 2 1> MiSeqReads.unmapped.mapped.sam 2>Dp_Mi_M-B_ctg_dedup-unmapped.log
```

From the above short reads, we removed the non-primary mapped Illumina reads and generated a list of reads aligning to the short-read (but not the long-read) assembly and then sorted them by the number of occurrences per contig.

1. Remove non-primary mapping reads:

```
$ samtools view -Sh -F 256 MiSeqReads.unmapped.mapped | grep -v "XS:i:" >
MiSeqReads.unmapped.mapped.no_mult.sam
```

2. Generate the sorted list of Illumina reads:

```
$ grep -v '^ @'MiSeqReads.unmapped.mapped.no_mult.sam | cut -f 3 | sort | uniq -c | sort -rn >
nbReadsMappedToMiSeqCtgs.list
```

Similarly, we mapped the transcripts not aligning with the long-read assembly onto the short-read assembly. For that step, we used gmap. Subsequently, we generated a list of transcripts that aligned with the short-read (but not the long-read) assembly, sorted by the number of occurrences per contig.

1. Build genome index and map transcripts:

```
$ gmap_build -D . -d MiSeqAssembly_dedup  MiSeqAssembly_dedup.fasta > gmapbuild.out 2>&1
$ gmap --format gff3_gene --nofails -t --min-identity=0.90 --min-intronlength=20 --gff3-add-separators=0 -
D . -d MiSeqAssembly_dedup/ transcripts.unmapped.fasta > transcripts.unmapped.mapped.gff3 2>
transcripts.unmapped.mapped.log
```

2. Generate the sorted list of transcripts:

```
$ grep 'gene' file.gmap.gff3 | grep -v '^#' | cut -f 1 | sort | uniq -c | sort -rn >
nbTranscriptsMappedToMiSeqCtgs.list
```

Based on the above outputs, we then selected short-read contigs to be added to the long-read assembly by applying two criteria: coverage by either (i) one or more transcripts, or (ii) more than 10× coverage of Illumina reads on a given contig. This was performed by the following seven steps.

1. Combine the lists short reads aligning to the short-read (but not the long-read) assembly:

```
$ awk 'NR==FNR { split($0, s, " "); a[s[2]]=s[1]"\t"s[2]; next } {split($0, ss, " "); if(ss[2] in a)
{print ss[1]"\t"ss[2]"\t"a[ss[2]]} else {print ss[1]"\t"ss[2]} }' nbTranscriptsMappedToMiSeqCtgs.list
nbReadsMappedToMiSeqCtgs.list > nbReads+TranscriptsMappedToMiSeqCtgs.list
```

2. Count the length of each contig in the short-read assembly with the in-house script ‘cn’:

```
$ cn MiSeqAssembly_dedup.fasta > MiSeqAssembly_dedup.cn
```

3. Add counts to the file generated in step 1 and arrange data in four columns: NameOfCtg, LengthOfCtg, NbReads, NbTranscripts

```
$ awk 'NR==FNR { split($9, s, "="); a[$1]=s[2]; next } {if($2 in a) {print $2"\t"a[$2]"\t"$1"\t"$3} }'
MiSeqAssembly_dedup.cn nbReads+TranscriptsMappedToCtgs.list >
ctgLengt+nbReadsMappedToCtgs+nbTranscripts.list
```

4. Add in the 1st column the coverage calculated by NbReads*200/(LengthCtg+1) (Average of 200 bp per Illumina read):

```
$ awk '{print ($3*200)/($2+1)"\t"$0 >> "ctgLengt+nbReadsMappedToCtgs+nbTranscripts+cov.list"}'
ctgLengt+nbReadsMappedToCtgs+nbTranscripts.list
```

5. Parse the file generated in 4. To retain only contigs with at least one transcript (unmapped_transcript) mapped or 10x short-read coverage:

```
$ awk '$1>10 || $5 > 0 ' ctgLengt+nbReadsMappedToCtgs+nbTranscripts+cov.list >>
ctgLengt+nbReadsMappedToCtgs+nbTranscripts+cov_covgt10_transcgt0.list
```

6. Create the list of short-read contigs absent from the long-read assembly:

```
$ awk '{print $2}' ctgLengt+nbReadsMappedToCtgs+nbTranscripts+cov_covgt10_transcgt0.list >
miseq_ctg_to_extract.list
```

7. Extract the contig sequences:

```
$ xargs samtools faidx MiSeqAssembly_dedup.fasta < miseq_ctg_to_extract.list >
MiSeqAssembly_dedup_missingctg.fasta
```

Contigs from the short-read assembly that were absent from the long-read assembly were combined with those from the long-read assembly and, after merging, the assembly was deduplicated.

1. Concatenate short-read contigs missing in long-read assembly:

```
$ cat MiSeqAssembly_dedup_missingctg.fasta PacBioAssembly_dedup.fasta > merged_genome_assembly.fasta
```

2. Deduplicate assembly:

```
$ dedupe2.sh in=merged_genome_assembly.fasta out=merged_genome_assembly_dedup.fasta threads=40 absorbrc=t
absorbmatch=t absorbcontainment=t overwrite=true maxedits=10 minidentity=90 findoverlap=t cluster=t
outd=merged_genome_assembly_del.fasta
```

Correction of mis-junctions. We realized that both assemblies contained incorrectly assembled contigs (referred to as chimeric contigs, or briefly "chimera"). To detect mis-junctions, we screened contigs for read and transcript coverage. This involved mapping of reads against the merged assembly followed by computing the read coverage for each position.

1. Build genome index and map Illumina reads to the assembly with Bowtie2 using options to consider relevant reads only, and sort bam file with sambamba v.0.6.7 (TARASOV *et al.* 2015) :

```
$ bowtie2-build -f merged_genome_assembly_dedup.fasta merged_genome_assembly_dedup
$ bowtie2 --local --no-unal -p 40 --no-mixed --no-discordant -x merged_genome_assembly_dedup.fasta -1
MiSeqReads_1.fastq.gz -2 MiSeqReads_2.fastq.gz -U MiSeqReads_S.fastq.gz 2> MiSeqReads.mapped.log |
sambamba view -S -f bam /dev/stdin | sambamba sort /dev/stdin -o MiSeqReads.mapped.sorted.bam ; sambamba
index MiSeqReads.mapped.sorted.bam
```

2. Calculate the coverage by position with bedtools genomecov (BEDTools v.2-2.25.0, (QUINLAN AND HALL 2010)):

```
$ bedtools genomecov -bga -ibam MiSeqReads.mapped.sorted.bam > MiSeqReads.mapped.sorted.genomecov.bedgraph
```

3. Extract genome positions where the read coverage is below 3:

```
$ awk '{if($4<3) print $0}' MiSeqReads.mapped.sorted.genomecov.bedgraph >
MiSeqReads.mapped.sorted.genomecov.lowcov.bedgraph
```

A similar procedure was performed with long reads, using the long-read mapper minimap2 (LI 2018) with recommended options for this type of reads.

1. Build genome index and map PacBio reads to assembly with minimap2 and sort output:

```
$ minimap2 -t 40 -d merged_genome_assembly_dedup.mmi merged_genome_assembly_dedup.fasta
$ minimap2 -ax map-pb -t 40 merged_genome_assembly_dedup.mmi PacBioReads.fasta 2> PacBioReads.mapped.log |
sambamba view -S -f bam /dev/stdin | sambamba sort /dev/stdin -o PacBioReads.mapped.sorted.bam ; sambamba
index PacBioReads.mapped.sorted.bam
```

2. Keep the reads with a mapping quality >0 and a length >1, and remove unmapped reads:

```
$ sambamba view -f bam -F "not unmapped and sequence_length > 1 mapping_quality > 0" -o
PacBioReads.mapped.sorted.filtered.bam PacBioReads.mapped.sorted.bam
```

3. Extract genome positions where the read coverage is below 2:

```
$ bedtools genomecov -bga -ibam PacBioReads.mapped.sorted.filtered.bam >
PacBioReads.mapped.sorted.filtered.genomecov.bedgraph
$ awk '{if($4<2) print $0}' PacBioReads.mapped.sorted.filtered.genomecov.bedgraph >
PacBioReads.mapped.sorted.filtered.genomecov.lowcov.bedgraph
```

A similar procedure was performed with transcripts.

1. Build genome index and map transcript sequences to assembly with gmap and sort output:

```
$ gmap_build -D . -d merged_genome_assembly_dedup  merged_genome_assembly_dedup.fasta > gmapbuild.out 2>&1
$ gmap --format gff3_gene --nofails -t --min-identity=0.90 --min-intronlength=20 --gff3-add-separators=0 -
D . -d merged_genome_assembly_dedup/ transcripts.fasta > transcripts.mapped.gff3 2> transcripts.mapped.log
$ awk 'NR>3' transcripts.mapped.gff3 | sort -k 1,1 -k 4,4n > transcripts.mapped.sorted.gff3
```

Next, we examined which low-coverage positions were spanned by transcripts or long reads, and would therefore not be considered as mis-junctions anymore.

1. Intersect short-read low-coverage positions with intervals spanned by at least one transcript and report *absence* of overlap:

```
$ bedtools intersect -v -a MiSeqReads.mapped.sorted.genomecov.lowcov.bedgraph -b
transcripts.mapped.sorted.gff3 > MiSeq_lowcov-Transcripts.intersect.bedgraph
```

2. Intersect non-overlapping positions from step 1 with intervals of low (<2) long-read coverage and report overlap (i.e., report that none of the datasets provides evidence for contiguity of the contig):

```
$ bedtools intersect -a MiSeq_lowcov-Transcripts.intersect.bedgraph -b
PacBioReads.mapped.sorted.filtered.genomecov.lowcov.bedgraph > MiSeq_lowcov-Transcripts-
PacBio_lowcov.intersect.bedgraph
```

3. Sort positions and merge them to create intervals:

```
$ sort -k1,1 -k2,2n MiSeq_lowcov-Transcripts-PacBio_lowcov.intersect.bedgraph > MiSeq_lowcov-Transcripts-
PacBio_lowcov.intersect.sorted.bedgraph
$ bedtools merge -i MiSeq_lowcov-Transcripts-PacBio_lowcov.intersect.sorted.bedgraph > MiSeq_lowcov-
Transcripts-PacBio_lowcov.intersect.sorted.merged.bed
```

Finally, we cut the merged genome at the mis-junction sites and removed the resulting contig 'pieces' that are shorter than 100 bp.

1. Cut contigs with the in-house script 'cut_chimera_fasta.py':

```
$ cut_chimera_fasta.py -b MiSeq_lowcov-Transcripts-PacBio_lowcov.intersect.sorted.merged.bed -f
merged_genome_assembly_dedup.fasta -o merged_genome_assembly_dedup_cut.fasta
```

2. Remove all contigs shorter than 100 bp with the in-house script 'remove_small.pl':

```
$ remove_small.pl 100 merged_genome_assembly_dedup_cut.fasta  >
merged_genome_assembly_dedup_cut_longer100.fasta
```

Transcriptome assembly

RNA-seq reads (Illumina HiSEQ) from the two stranded poly-RNA libraries PA and PA2 were collectively assembled *de novo* with Trinity v.2.6.6 (https://github.com/trinityrnaseq/trinityrnaseq; (GRABHERR *et al.* 2011)) using default parameters. We also generated a genome-guided transcriptome assembly, again with default parameters. A set of high-confidence transcripts was generated to test whether genomic contig regions with low-read coverage were due to mis-joining or were spurious (see Section 1.5. Correction of mis-joining). These were selected based on two criteria: (i) presence of a Spliced-Leader (SL) sequence at the 5′ end, and (ii) an overall read coverage of >100. To quantify gene expression, the corresponding transcript sequence was first inferred for each protein ID from the 'submission' proteome (see below). Next, the poly-A RNA-Seq reads mentioned above were mapped on individual transcripts using bowtie2 v2.4.5 (LANGMEAD AND SALZBERG 2012) in the end-to-end mode, assuming the maximum insert length of 600 bp and only allowing alignment concordant with strand-specific expression. The SAM file was converted to a BAM file with sambamba v0.7.1 (TARASOV *et al.* 2015). Transcript levels were then calculated using salmon v1.9.0 (PATRO *et al.* 2017) in the alignment mode (default parameters, library type 'ISF').

Structural genome annotation

The structural genome annotation pipeline employed here was implemented similarly as described earlier (GRAY *et al.* 2020).

Structural annotation of protein-coding genes. Compared to the earlier pipeline, we used a modified RNA-Seq read-mapping step here. Briefly, the genome assembly was first masked for simple repeats using RepeatScout v1.0.5 (PRICE *et al.* 2005) and RepeatMasker v4.0.9 [unpublished: https://www.repeatmasker.org/].

```
$ build_lmer_table -sequence Dp_PB-MI_190104_dedup_cut_l100.fasta -freq lmer_table.txt
$ RepeatScout -minthres 150 -sequence Dp_PB-MI_190104_dedup_cut_l100.fasta -output repscout.fasta -freq
lmer_table.txt
RepeatMasker -xsmall -gff -s -pa 40 -lib repscout.fasta Dp_PB-MI_190104_dedup_cut_l100.fasta
```

RNA-Seq reads were then mapped to the genome assembly using STAR v2.6.1b (DOBIN *et al.* 2013) to retrieve the positions of SL sequences, inferred within soft-clipped regions, using an in-house script.

```
$ STAR --runThreadN 40 --runMode genomeGenerate --genomeDir STAR-index --genomeFastaFiles Dp_PB-
MI_190104_dedup_cut_l100.fasta --genomeSAindexNbases 13
$ STAR --runThreadN 40 --genomeDir build-index --alignEndsType Local --readFilesIn PA+DPA2_1.fastq.gz
PA+DPA2_2.fastq.gz --outSAMtype BAM SortedByCoordinate --outSJfilterIntronMaxVsReadN 100 300 500 --
alignIntronMin 19 --alignIntronMax 20000 --outFileNamePrefix STAR_ --outSAMattributes All --
outSAMattrIHstart 0 --outSAMstrandField intronMotif --limitBAMsortRAM 27643756136 --readFilesCommand zcat
```

Reads were then depleted of the SL sequence (5′–AACCAACGATTTAAAAGCTACAGTTTCTGTACTTTATTG), using trimmomatic v0.30 (BOLGER *et al.* 2014):

```
$ java -classpath trimmomatic-0.30.jar org.usadellab.trimmomatic.TrimmomaticPE -threads 24 -phred33
PA+DPA2_1.fastq.gz PA+DPA2_2.fastq.gz PA+DPA2_1.filtered.fastq PA+DPA2_2_unpaired.fastq
PA+DPA2_2.filtered.fastq PA+DPA2_2_unpaired.fastq ILLUMINACLIP:splicedleader.fasta:3:30:9:10 MINLEN:50
```

and then remapped to the genome assembly using STAR

```
$ STAR --runThreadN 40 --genomeDir build-index --alignEndsType Local --readFilesIn PA+DPA2_1.fastq.gz
PA+DPA2_2.fastq.gz --outSAMtype BAM SortedByCoordinate --outSJfilterIntronMaxVsReadN 100 300 500 --
alignIntronMin 19 --alignIntronMax 20000 --outFileNamePrefix STAR_ --outSAMattributes All --
outSAMattrIHstart 0 --outSAMstrandField intronMotif --limitBAMsortRAM 27643756136 --readFilesCommand zcat
```

The SL-trimmed reads, both unmapped and mapped, were assembled *de novo* and guided by the genome assembly, respectively, using Trinity v2.6.6 (GRABHERR *et al.* 2011).

```
$ Trinity --seqType fq --max_memory 150G --left PA+DPA2_2.filtered.fastq --right PA+DPA2_1.filtered.fastq
--CPU 40 --output trinity-denovo --full_cleanup --SS_lib_type RF
$ Trinity --genome_guided_max_intron 20000 --max_memory 250G --CPU 40 --genome_guided_bam
STAR_Aligned.sortedByCoord.out.bam --output trinity-gg --full_cleanup --SS_lib_type RF
```

The resulting transcriptome assemblies were combined into a single file, trinity-comprehensive.fasta, and aligned to the genome assembly using PASA v2.3.3 (HAAS *et al.* 2003).

```
$ Launch_PASA_pipeline.pl -c alignAssembly.config -C -r -R -g Dp_PB-MI_190104_dedup_cut_l100.fasta -t
trinity-comprehensive.fasta.clean -T -u trinity-comprehensive.fasta --ALIGNERS gmap,blat --CPU 40 --TDN
tdn.accs -I 20000 --stringent_alignment_overlap 30.0 -transcribed_is_aligned_orient
```

The alignments were then combined into a single, comprehensive assembly:

```
build_comprehensive_transcriptome.dbi -c alignAssembly.config -t Dp_PB-
MI_190104_dedup_cut_l100.sqlite.assemblies.fasta --min_per_ID 95 --min_per_aligned 95
```

Protein sequence accessions GCA_000002725.2, GCA_000002845.2, GCA_000002875.2, GCA_000209065.1, GCA_000227375.1, GCA_000691245.1, GCA_001457755.2, GCA_001680005.1, GCA_002087225.1, GCA_002157705.1, GCA_003719475.1, GCA_003719485.1, GCA_900002335.1, GCA_900005765.1, GCA_900005855.1, GCA_900090025.2, GCA_900090045.1, GCA_900097015.1, GCA_900240055.1, along with other available sequences from *Discoba,* were aligned to the genome using Spaln v2.2.2 (GOTOH 2008).

```
$ spaln -C1 -O12 -Q5 -yL20 -yX -t40 -dDp_PB-MI_190104_dedup_cut_l100 all_protein_data.faa
```

The *ab initio* predictors employed were Genemark v4.33 with intron intervals as hints derived from RNA-Seq read mapping (LOMSADZE *et al.* 2014),

```
$ gmes_petap.pl --soft 1000 --ET=introns.gff --et_score=3 --cores=40 --sequence=Dp_PB-
MI_190104_dedup_cut_l100.fasta
```

CodingQuarry v2.0 was run with transcript alignments as hints (TESTA *et al.* 2015):

```
$ CodingQuarry -p 40 -f Dp_PB-MI_190104_dedup_cut_l100.fasta -t pasa_transcripts.gff3
```

Augustus v3.3.2 (STANKE *et al.* 2006) was employed with spliced-leader-sequence genomic positions as transcription-start-site (tss) hints, along with protein sequence alignments, RNA-Seq read coverage, and transcript alignments (as described at https://bioinf.uni-greifswald.de/bioinf/wiki/pmwiki.php?n=Augustus.Augustus), and Snap (KORF 2004) trained on Augustus models with a score of 1 (as per the instructions on https://github.com/KorfLab/SNAP). Finally, the PASA assembly, Spaln alignments, as well as Augustus, Snap, and Codingquarry gene models, were combined into a single consensus with Evidencemodeler v1.1.1 (HAAS *et al.* 2008) following the instructions at https://evidencemodeler.github.io/.

The comprehensive, non-redundant transcript and protein sequence collection was generated by first combining the manually curated gene models with the remaining automated models. Models manually identified as split were conceptually translated and reassembled. Those combined sequences were clustered with the assembled transcripts containing ORFs using CD-HIT (default parameters of v4.8.1) (FU *et al.* 2012). Transcript sequences with no corresponding matches were combined with the curated and automated model sequences and the nucleotide sequences were then conceptually translated. This comprehensive sequence collection was used for downstream analyses.

Structural annotation of non-coding RNA genes. Modeling of tRNAs was performed using tRNAscan-SE v1.3.1 (LOWE AND EDDY 1997):

```
$ tRNAscan-SE --brief --codons --output tRNAscan-SE.out Dp_PB-MI_190104_dedup_cut_l100.fasta
```

Infernal v1.1.1 (NAWROCKI AND EDDY 2013) was used to identify other classes of non-coding RNAs using the covariance models from Rfam v12.1 (GRIFFITHS-JONES *et al.* 2003):

```
$ cmscan --rfam --fmt 2 -E 1e-3 --nohmmonly --oskip --clanin Rfam.clanin --cpu 40 --tblout rfam.out -noali
Rfam.cm Dp_PB-MI_190104_dedup_cut_l100.fasta
```

Functional annotation of protein-coding gene models

Coding sequences from gene models were extracted from the output of Evidencemodeler and conceptually translated. The translated sequences were then searched against UniProt/SWISS-PROTKB (downloaded March 23, 2018; (BOUTET *et al.* 2007)), as well as the GenBank accessions and other *Discoba* sequences used for structural annotation, to identify the single best hit below the maximum threshold E-value of 1.0e-7 using Blastp v2.2.31+.

```
$ makeblastdb -dbtype prot -in all_protein_data.faa
$ blastp -db all_protein_data.faa -num_threads 40 -outfmt '6 qseqid sseqid stitle pident length qlen slen
evalue bitscore' -max_target_seqs 5 -evalue 1e-7 -query Dp_PB-MI_190104_dedup_cut_l100.faa -out Dp_PB-
MI_190104_dedup_cut_l100.faa.blastp_all_protein_data
```

Product names were transferred to *D. papillatum* gene models by taking the single best hit against the SWISS-PROT database, i.e., lowest E-value below a global cutoff of 1e-7. Precedence was given to hits of GenBank accessions if the E-value was lower than a competing hit to the SWISS-PROT database, otherwise the product name was automatically transferred in the absence of a SWISS-PROT hit provided the E-value was below the global cutoff. All remaining models without hits below the threshold were assigned 'hypothetical protein' as their 'product'. Hmmer v3.1b1 was also used to search for conserved domains described in Pfam v31.0 using the model-specific noise threshold as E-value cutoff. Blastp and Hmmer search hits were included in the 9th column of the gff3 file (https://www.ncbi.nlm.nih.gov/genbank/genomes_gff/) as 'product' and 'inference' attributes, respectively, as per the NCBI eukaryotic genome annotation guidelines (https://www.ncbi.nlm.nih.gov/genbank/eukaryotic_genome_submission_annotation/).

Genome viewing and annotation curation

For viewing contigs, read coverage, gene models and functional assignments, and editing structural and functional annotations, we used the web-based Apollo software v.2.6.2 (https://github.com/GMOD/Apollo/releases/latest; (DUNN *et al.* 2019). Apollo allows real-time, collaborative and simultaneous genome annotation editing, whereby models can later be exported in gff3 format. Manually curated models were exported from Apollo in gff3 format and incorporated into the automated annotation.

AUTHOR CONTRIBUTIONS

Conceptualization – S.M., C.P., M.S., G.B.; **Data curation** – S.M., M.S.; **Formal analysis, Investigation, Writing, original draft** – S.M., C.P., M.S., P.S., G.B., M.V.; **Visualization** – P.S., G.B.; **Writing, review & editing** – all co-authors.

REFERENCES

Bolger, A. M., M. Lohse and B. Usadel, 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114-2120.

Boutet, E., D. Lieberherr, M. Tognolli, M. Schneider and A. Bairoch, 2007 UniProtKB/Swiss-Prot. Methods Mol Biol 406: 89-112.

Clayton, C., 2019 Regulation of gene expression in trypanosomatids: living with polycistronic transcription. Open Biol 9: 190072.

Denisov, G., B. Walenz, A. L. Halpern, J. Miller, N. Axelrod *et al.*, 2008 Consensus generation and variant detection by Celera Assembler. Bioinformatics 24: 1035-1040.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski *et al.*, 2013 STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29: 15-21.

Dunn, N. A., D. R. Unni, C. Diesh, M. Munoz-Torres, N. L. Harris *et al.*, 2019 Apollo: democratizing genome annotation. PLoS Comput Biol 15: e1006790.

Fu, L., B. Niu, Z. Zhu, S. Wu and W. Li, 2012 CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28: 3150-3152.

Gotoh, O., 2008 A space-efficient and accurate method for mapping and aligning cDNA sequences onto genomic sequence. Nucleic Acids Res 36: 2630-2638.

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson *et al.*, 2011 Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29: 644-652.

Gray, M. W., G. Burger, R. Derelle, V. Klimeš, M. M. Léger *et al.*, 2020 The draft nuclear genome sequence and predicted mitochondrial proteome of *Andalucia godoyi*, a protist with the most gene-rich and bacteria-like mitochondrial genome. BMC Biol 18: 22.

Griffiths-Jones, S., A. Bateman, M. Marshall, A. Khanna and S. R. Eddy, 2003 Rfam: an RNA family database. Nucleic Acids Res 31: 439-441.

Haas, B. J., A. L. Delcher, S. M. Mount, J. R. Wortman, R. K. Smith, Jr. *et al.*, 2003 Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res 31: 5654-5666.

Haas, B. J., S. L. Salzberg, W. Zhu, M. Pertea, J. E. Allen *et al.*, 2008 Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol 9: R7.

Haghshenas, E., F. Hach, S. C. Sahinalp and C. Chauve, 2016 CoLoRMap: Correcting Long Reads by Mapping short reads. Bioinformatics 32: i545-i551.
Koren, S., B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman *et al.*, 2017 Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res 27: 722-736.
Korf, I., 2004 Gene finding in novel genomes. BMC Bioinformatics 5: 59.
Kriventseva, E. V., D. Kuznetsov, F. Tegenfeldt, M. Manni, R. Dias *et al.*, 2019 OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Res 47: D807-d811.
Langmead, B., and S. L. Salzberg, 2012 Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357-359.
Li, H., 2018 Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34: 3094-3100.
Lomsadze, A., P. D. Burns and M. Borodovsky, 2014 Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acids Res 42: e119.
Lowe, T. M., and S. R. Eddy, 1997 tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25: 955-964.
Myers, E. W., G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo *et al.*, 2000 A whole-genome assembly of *Drosophila*. Science 287: 2196-2204.
Nawrocki, E. P., and S. R. Eddy, 2013 Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29: 2933-2935.
Patro, R., G. Duggal, M. I. Love, R. A. Irizarry and C. Kingsford, 2017 Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods 14: 417-419.
Price, A. L., N. C. Jones and P. A. Pevzner, 2005 De novo identification of repeat families in large genomes. Bioinformatics 21 Suppl 1: i351-358.
Quast, C., E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer *et al.*, 2013 The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41: D590-596.
Quinlan, A. R., and I. M. Hall, 2010 BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841-842.
Schmieder, R., and R. Edwards, 2011 Quality control and preprocessing of metagenomic datasets. Bioinformatics 27: 863-864.
Seppey, M., M. Manni and E. M. Zdobnov, 2019 BUSCO: assessing genome assembly and annotation completeness. Methods Mol Biol 1962: 227-245.
Stanke, M., O. Schöffmann, B. Morgenstern and S. Waack, 2006 Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics 7: 62.
Sturm, N. R., D. A. Maslov, E. C. Grisard and D. A. Campbell, 2001 *Diplonema* spp. possess spliced leader RNA genes similar to the Kinetoplastida. J Eukaryot Microbiol 48: 325-331.
Tarasov, A., A. J. Vilella, E. Cuppen, I. J. Nijman and P. Prins, 2015 Sambamba: fast processing of NGS alignment formats. Bioinformatics 31: 2032-2034.
Testa, A. C., J. K. Hane, S. R. Ellwood and R. P. Oliver, 2015 CodingQuarry: highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts. BMC Genomics 16: 170.
Wu, T. D., and C. K. Watanabe, 2005 GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21: 1859-1875.

3. The ploidy level of *Diplonema papillatum*

INTRODUCTION

Most eukaryotes alternate between a diploid and a haploid phase. Such an alternation is a consequence of sexual reproduction, because gamete fusion leads to a diploid phase and meiosis to a haploid phase (MABLE AND OTTO 1998). However, the duration of these phases varies considerably across eukaryotes. In most metazoans and land plants, the diploid stage of the life cycle is the predominant state, i.e., mitotic cell divisions occur exclusively in the diploid phase. Other types of alternation are observed in fungi and diverse protist groups. In the haplontic cycle, mitotic cell division occurs entirely in the haploid form (e.g., in the fission yeast *Schizosaccharomyces*), whereas in the haploid-diploid cycle, mitosis takes place in both phases, often forming morphologically different organisms (e.g., in jellyfish).

Diplonema papillatum has never been observed to reproduce sexually. Here we attempt to infer the ploidy status of the strain propagated in the laboratory over many years, *via* the heterozygosity of its nuclear genome.

RESULTS AND DISCUSSION

Extremely low heterozygosity of the *Diplonema* nuclear genome

Heterozygosity of a genome is typically determined by analyzing either single-nucleotide polymorphisms (SNPs) in reads aligned to the reference genome —which requires an assembly— or k-mers occurring in sequencing reads —which does not require a genome assembly.

The approach using **SNPs** (also called single-nucleotide variants (SNVs)) relies on accurate read mapping to the reference genome. However, ~50% of the 280.4 Mbp *Diplonema* nuclear genome assembly consists of repeats up to >20 kbp long, some of which occur in nearly 5,000 copies. These repeats cause erroneous mapping of Illumina reads to the assembly, and thus generate spurious polymorphic sites. Therefore, for this analysis we only used read pairs that congruently and uniquely align with the *Diplonema* genome assembly and we excluded all variants that fall in repeat regions. After removing sites with low mapping quality and other biases (strand bias, etc.), only 557 SNPs remained in the 142 Mbp-long genome portion outside repeats. As commonly observed, nucleotide transitions (327) are more frequent than transversions (230) in this final set of variants (**Supplementary Table S10**). The resulting heterozygous SNP rate of the *Diplonema* nuclear genome is extremely low. With only 4 e-06 per 1 kbp, it is merely 0.2% of that of a diploid eukaryote, e.g., human (SACHIDANANDAM *et al.* 2001). In addition, the profile of allele frequencies within the *Diplonema* genome (**Supplementary Figure S8**) shows no peaks, in contrast to profiles from diploid and polyploid organisms.

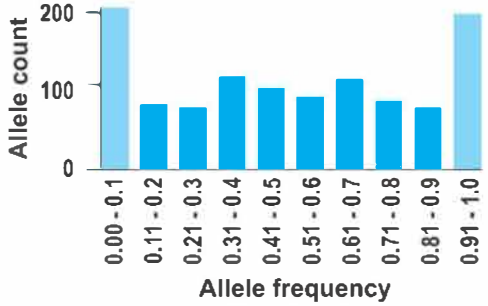
The second approach to evaluating genome heterozygosity counts **k-mers** directly in reads. In theory, k-mer profiles of homozygous (haploid or higher ploidies) genomes are Poisson distributions centred at the average read coverage, but several biases cause distortions. Repeats in the genome over-amplify certain k-mers, thus adding a drifting-off shoulder at higher coverage, whereas sequencing errors generate numerous low-frequency k-mers, adding a high peak toward zero coverage. For analysing the k-mer distribution, we used nuclear Illumina reads (57,742,700 reads totalling 13,139,885,848 bp; mitochondrial reads were removed; see Methods). The resulting k-mer profile (**Supplementary Figure S9**) resembles a Poisson distribution typical for homozygotic genomes. The only slight deviation from a true Poisson distribution is most likely due to sequencing errors and repeats.

K-mer frequencies of reads also allow one to estimate the genome size, notably based on the total number of k-mers and the peak position of the distribution. For the k-mer length range of 17, 19, 21, and 23, the average inferred genome length is 259,725,615 bp.

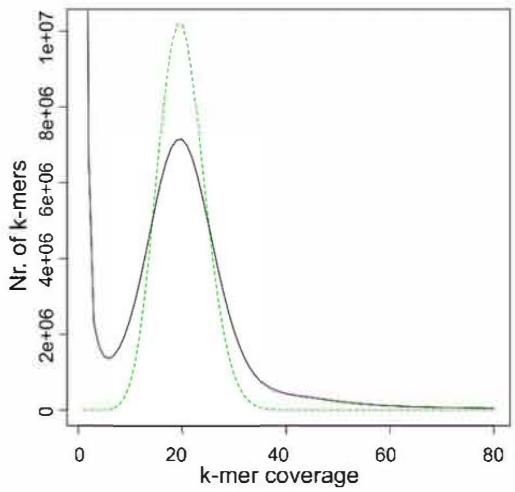
Supplementary Table S10. Single nucleotide variants[a].

Reference allele	Alternative Allele				Sum of reference alleles	Cumulative substitutions
	A	C	G	T		
A	/	41	67	24	132	**A↔G: 172**
C	32	/	25	94	151	A↔C: 73
G	105	21	/	25	151	A↔T: 57
T	33	84	36	/	153	C↔G: 46
Sum of alternative alleles	170	146	128	143		**C↔T: 178**
						T↔G: 66
Total transitions						350 (59%)
Total transversions						242 (41%)

[a]Significant variants after filtering; see Methods



Supplementary Figure S8. Allele frequency outside repeat regions of the *D. papillatum* nuclear genome. After filtering sites for mapping quality, strand and location biases, a total of 557 SNPs are retained in the 142,402,538 Mbp nuclear genome regions outside repeats. Bars represent bins for allele frequencies. The left-most and right-most bins most likely include sequencing and base-calling errors. The absence of pronounced peaks in the frequency range of 0.2 to 0.8 and the low heterozygosity level corroborate that the genome is haploid.



Supplementary Figure S9. K-mer profile of reads from the *D. papillatum* nuclear genome. X-axis, k-mer coverage, i.e., number of 'individuals' per k-mer species; Y-axis, number of distinct k-mer species. Black line, the 17-mer profile of *Diplonema* nuclear MI-Seq reads. The shape differs only marginally from those for 19-mers, 21-mers and 23-mers. Green dotted line, Poisson distribution centered at the mean coverage. The reads' k-mer distribution lacks a shoulder left to the main peak, otherwise indicative of heterozygosity. The high number of unique k-mers and k-mers with a very low coverage is most probably due to reads with sequencing errors.

Diplonema is most likely haploid

The lack of heterozygosity is inherent in haploid genomes but also in diploid lines originating from recent spontaneous chromosome duplication (autodiploidization leading to whole genome duplication). Lack of heterozygosity can also occur when haploid cells in a clonal population mate and when homolog chromosomes fail to separate during meiosis, thus giving rise to diploid gametes.

With computational methods, it is impossible to distinguish homozygous haploid from homozygous 2N or >2N genomes. However, gene replacement experiments are evidence for a haploid *Diplonema* nuclear genome (KAUR *et al.* 2018). In diploid organisms, gene replacement leads nearly always to two gene versions, the original one on one chromosome, and the replaced one on the other, which can be tested by PCR amplification and Southern hybridization of the corresponding genomic region. In multiple knock-in transformation experiments in *Diplonema* with an engineered gene version, only a single allele was ever detected, strongly suggesting haploidy.

It should be noted that the statement of haploidy refers to the *D. papillatum* culture used in the laboratory. Here, the exclusive form of reproduction appears to be mitosis, which is probably the predominant reproduction mode in *Diplonema's* natural habitat. While as of now, **sexual reproduction** or a diploid stage have not been observed in this protist, the gene repertoire (see main text) implies that this organism has the potential to form diploid zygotes that undergo meiosis.

METHODS

Mapping of reads onto the nuclear genome sequence. We used all genomic MiSeq reads of *Diplonema* generated by us and removed mitochondrial reads by eliminating those that aligned with sequences of complete mitochondrial chromosomes (GenBank acc. nos. EU123536.1; EU12353637.1; HQ288823.1, HQ288824.1; JQ302962.1, JQ314396.1) using Bowtie2 version 2.3.4.3 (command line arguments --local --un) (LANGMEAD AND SALZBERG 2012). The ~58 million 'cleaned' reads (~13 Gbp) were mapped to the reference assembly using Bowtie2 (--no-discordant --no-mixed --no-unal) yielding 98.7% alignment rate. The resulting sam output file was converted with Samtools v1.8 subcommand view (LI *et al.* 2009) to the bam format, and further processed with the Picard tools v.1.139, using subcommands Sortsam, MarkDuplicates, AddGroup, and IndexBam (https://broadinstitute.github.io/picard/), all in default mode.

Determination of single nucleotide variants. Sequence variants across reads per alignment column were called with Freebayes v1.2.0-2-g29c4002 (--ploidy 4 to allow for higher ploidy) (GARRISON E AND MARTH G 2012). To remove from the resulting vcf file variants that fall into genomic regions containing repeats, we used the corresponding gff file generated by our genome annotation pipeline (Sarrasin et al., unpublished). This pipeline calls Repeatmasker v.4.0.6 (http://repeatmasker.org) developed by A.F.A. Smit, R. Hubley & P. Green, and uses RepBase24.06, release 06-20-2019 (JURKA *et al.* 2005). The relative complement between repeat intervals and variant positions in the vcf file were obtained by executing BEDTools v2-2.28 subcommand intersect (-a -v) (QUINLAN 2014). Then, indels and low quality variants were removed with Vcftools v0.1.12b (--minQ 30 –remove --indels --recode --recode-INFO-all) (http://vcftools.sourceforge.net/docs.html; (DANECEK *et al.* 2011)), and variant sites were further filtered by read placement score (RPP >20) and strand bias (SRP <20) using an in-house script. The number of single-nucleotide variants (with frequencies between >0 and <1) recovered at each step are listed in **Supplementary Table S11**. The vcf file was converted to an allele-frequency table with the Genome Analysis ToolKit (GATK; (MCKENNA *et al.* 2010)) subcommand VariantToTable, by extracting two fields, the reference allele observation count (RO) and the alternate allele observation count (AO). The table was then imported into Excel to generate the chart.

Supplementary Table S11. Number of single-nucleotide variants at the diverse filtering steps.

Type of variants	Count
Initial variant count across the entire genome and including indels and substitutions	2,469,945
Nr. of variants outside repeats	871,013
Nr. of variants after indel removal and quality filtering (minQ 30)	16,640
Nr. of variants after read-placement and strand-bias filtering (RPP>20; SAP, SRP<20)	587

Determination of k-mer distribution and estimation of genome size. We 'decontaminated' Illumina reads from those containing mitochondrial sequences by mapping reads against the nuclear genome assembly. A total of 58 million 'decontaminated' reads remained summing up to 13 Gbp. The k-mer distribution of these genomic reads was calculated with the k-mer counter Jellyfish (MARÇAIS AND KINGSFORD 2011) using the command jellyfish count with the option -C for

counting k-mers from both strands and -m 17 to -m 23 for k-mer sizes from 17 to 23. The histogram of k-mer occurrences were generated with jellyfish histo and default parameters following the tutorial described at http://koke.asrc.kanazawa-u.ac.jp/HOWTO/kmer-genomesize.html. Histograms were plotted with R, including the comparison with a Poisson distribution and the estimation of the genome size inferred from the total number of k-mers (area under the curve) divided by the mean k-mer coverage. The size estimates based on the k-mer counts of lengths 17, 19, 21, and 23 are listed in **Supplementary Table S12**.

Supplementary Table S12. Genome-size estimation based on K-mer distribution.

K-mer length	17	19	21	23	**Average**
Size estimate	250,341,083 bp	258,503,293 bp	267,598,590 bp	262,459,497 bp	**259,725,615 bp**

Determination of allele frequencies. From the vcf file containing only single-nucleotide variants of high quality, we extracted the allele frequencies (AD) and coverage (DP) with the GATK subcommand VariantsToTable (-GF AD –GF DP). Then the numbers were imported into Excel to calculate the quotient AD/DP for each allele and generate a frequency graph.

AUTHOR CONTRIBUTIONS

Conceptualization, Data curation, Formal analysis, Investigation, Visualization, Writing, original draft – G.B., L.L.; **Writing, review & editing** – all co-authors.

REFERENCES

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks *et al.*, 2011 The variant call format and VCFtools. Bioinformatics 27**:** 2156-2158.

Garrison E, and Marth G, 2012 Haplotype-based variant detection from short-read sequencing, pp., edited by arXiv.

Jurka, J., V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany *et al.*, 2005 Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110**:** 462-467.

Kaur, B., M. Valach, P. Peña-Diaz, S. Moreira, P. J. Keeling *et al.*, 2018 Transformation of *Diplonema papillatum*, the type species of the highly diverse and abundant marine microeukaryotes Diplonemida (Euglenozoa). Environ Microbiol 20**:** 1030-1040.

Langmead, B., and S. L. Salzberg, 2012 Fast gapped-read alignment with Bowtie 2. Nat Methods 9**:** 357-359.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics 25**:** 2078-2079.

Mable, B. K., and S. P. Otto, 1998 The evolution of life cycles with haploid and diploid phases. BioEssays 20**:** 453-462.

Marçais, G., and C. Kingsford, 2011 A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27**:** 764-770.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis *et al.*, 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20**:** 1297-1303.

Quinlan, A. R., 2014 BEDTools: the swiss-army tool for genome feature analysis. Curr Protoc Bioinformatics 47**:** 11.12.11-34.

Sachidanandam, R., D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein *et al.*, 2001 A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409**:** 928-933.

4. Intron splicing and structural RNAs

INTRODUCTION

Nuclear protein-coding genes of eukaryotes often contain intervening sequences, which are removed from pre-mRNAs by RNA splicing. The large majority of these **spliceosomal introns** have a GT dinucleotide at their 5′ end and an AG at their 3′ end, and are referred to as the 'major intron type'. Introns with AT-AC splice-site combinations (the 'minor intron type') occur occasionally in embryophyte plants and most metazoan groups, but have also been reported in oomycetes, mycetozoans, and a few basally-branching fungal lineages. Among all introns in plant and animal nuclear genomes, less than 0.5% have AT-AC boundaries, while those of baker's and fission yeast are exclusively of the major type (TURUNEN *et al.* 2013).

GT-AG and AT-AC introns are processed by the major and minor **spliceosome**, respectively, that are composed of five small nuclear RNAs (U RNAs) and several proteins (PATEL AND STEITZ 2003; TURUNEN *et al.* 2013). The most clear-cut diagnostic feature for distinguishing the two spliceosome types is their RNA-subunit composition, with U1, U2, U4, U5, and U6 RNAs defining the major spliceosome, whereas U11, U12, U4atac, U5, and U6atac RNAs are hallmarks of the minor spliceosome. Thus, the two spliceosome types share U5, whereas U1 and U11, U2 and U12, U4 and U4 atac, and U6 and U6atac are functionally analogous. They share most features of their two and three-dimensional structure but differ significantly in sequence (TARN AND STEITZ 1996b; TARN AND STEITZ 1996a). Yet, in some eukaryotes, intron splice sites deviate from the conserved GT-AG or AT-AC motifs (e.g., AT-AG), and in others, the number of 'minor introns' has increased significantly in both absolute numbers and relative proportion (LARUE *et al.* 2021). Therefore, the notion of 'major' and 'minor' introns or spliceosomes has been abandoned in favour of U2-type and U12-type. Note also that based on intron boundaries, one cannot infer with certainty whether the intron is spliced by a U12 or a U2-spliceosome (SHARP AND BURGE 1997).

Both U2- and U12-type introns and their cognate spliceosomes are believed to derive from a common eukaryotic ancestor (RUSSELL *et al.* 2006). This view is corroborated by the recent publication of the draft nuclear genome sequence from *Andalucia godoyi*, a slowly-evolving member of the early-branching Discoba supergroup (GRAY *et al.* 2020), whose nuclear genome encodes exclusively U RNA counterparts of the U2-type spliceosome. The same is true for the rapidly evolving discoban group, kinetoplastids, although their U RNAs are highly derived (GÜNZL 2010).

The spectrum of U-RNA variants across eukaryotes is likely much larger than currently appreciated. Although U RNAs are best identified using covariance models (CM) that capture the conserved, distinct sequence features and higher-order structure of U RNAs, recognition of homologs from poorly studied eukaryotic groups can still be challenging because available CMs (e.g., those compiled in the RFAM database (KALVARI *et al.* 2021)) are built with a taxonomically biased set of sequences.

The spliceosome not only catalyzes *cis*-splicing, i.e., intron excision and exon joining within the same pre-mRNA molecule, but also ***trans*-splicing** by which separate transcripts are joined. More specifically, in some metazoan groups and in euglenozoans including *Diplonema*, a short (~15-50 nt) leader sequence termed spliced-leader (SL) RNA is added to the 5′ end of pre-mRNAs. SL-*trans*-splicing in euglenozoans provides a 5′ cap structure to mRNAs and resolves long polycistronic transcripts into single-gene RNAs (HASTINGS 2005). Among euglenozoans, the spliceosome of trypanosomes, and in particular *Trypanosoma brucei*, has been studied in great detail. It belongs to the U2-type and is responsible for both the predominant SL-*trans*-splicing, as well as for *cis*-splicing of the rare introns (the *T. brucei* nuclear genome contains only two introns, one in a gene encoding a poly(A)-polymerase and the other in an RNA-helicase gene) (MICHAELI 2011).

RESULTS AND DISCUSSION

Intron types and spliceosome components in *D. papillatum*

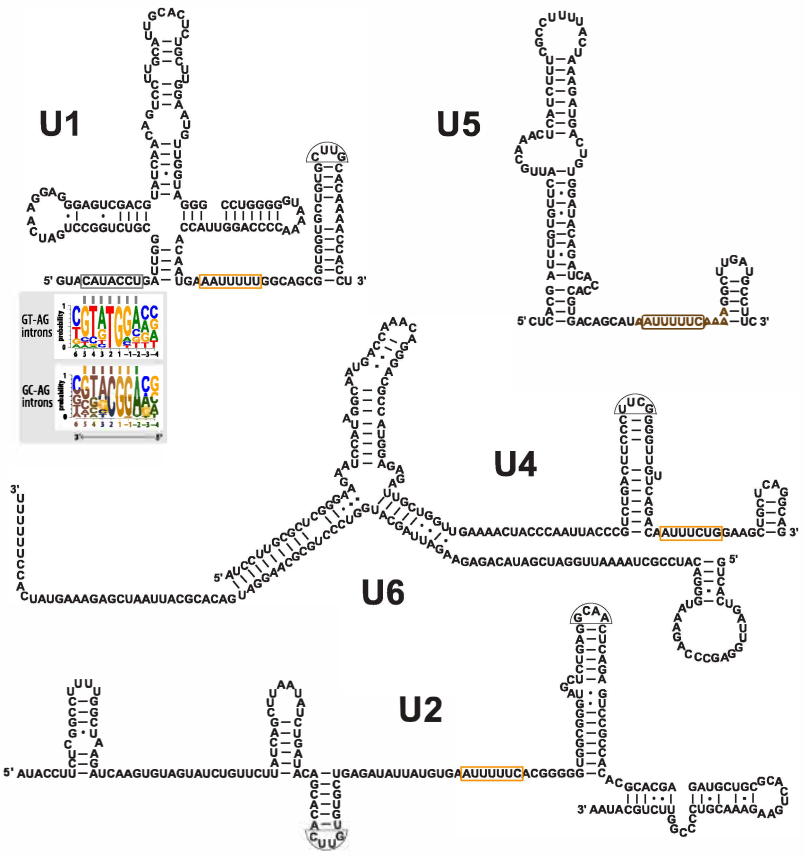
The large majority of **introns** in the *D. papillatum* nuclear genome assembly are canonical, bearing GT at their 5′-end and AG at the 3′-end (U2-type). In addition, non-canonical introns with GC-AG splice-site combinations were also detected. As in the case of numerous animals, fungi, and plants (FREY AND PUCKER 2020), GC-AG introns from *Diplonema* are most likely spliced by the same major U2-type spliceosome as GT-AG introns.

The lack of U12-type spliceosomal introns (AT-AC and variants thereof) in *Diplonema* is consistent with the lack of the U4atac, U6atac, U11, and U12 **RNAs** among the set of spliceosomal RNAs detected. To guarantee a highly sensitive identification of U RNAs, we used the corresponding RFAM CMs (KALVARI *et al.* 2021), and constructed new CM versions representing the largest range possible for eukaryotes (for details, see Methods). Compared to the RFAM models, our

improved CMs yielded significantly higher scores in identifying U RNAs of *Diplonema* and across eukaryotes as a whole, which strongly suggests the absence of a U12-type spliceosome in *D. papillatum*.

To identify potential structural variations of the five *Diplonema* spliceosomal RNAs, we designed two-dimensional (2D) diagrams following the characteristics of the human and yeast counterparts for which atomic structures are available ((YAN *et al.* 2015; ZHANG *et al.* 2017) (**Supplementary Figure S10**). Contrary to the situation in kinetoplastids (GÜNZL 2010), all five U-RNAs from *Diplonema* closely resemble their homologs from human and yeast U2-type spliceosomes regarding conserved sequence motifs and 2D structure. This lends further support to the view that the U2-type spliceosomal machinery has an ancient evolutionary origin and generally underwent only minor changes except in certain extremely fast-evolving species such as *Trypanosoma*.

Conservation of **spliceosomal proteins** is far more difficult to assess and interpret. According to the most recent data in the UniProt database (UNIPROTKB 2021), more than 200 spliceosomal proteins have been identified in human, but only 67 in yeast. When searching the complete set of *Diplonema* proteins for homologs of the yeast spliceosomal proteins using the highly sensitive Hidden Markov model (HMM) search algorithm (EDDY 2011), not more than a half returned hits below a confidence E-value threshold of 1.0 e-5. Among these, it was often difficult to distinguish between true orthologs and other members of the corresponding protein family, or even to detect the mere presence of conserved protein domains. This leaves U-RNA sequences and structures as the only clear-cut criterion for diagnosing the presence of U2- versus U12-type spliceosomes in organisms that are phylogenetically distant from human and yeast. In summary, from the intron-boundary sequences and U RNA features, we conclude that *Diplonema* possesses a single kind of spliceosome, the U2-type spliceosome.

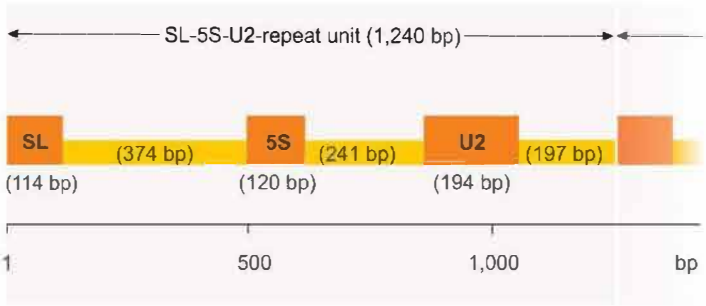


Supplementary Figure S10. 2D structure diagrams of the five *Diplonema* spliceosomal RNAs. The 2D structures were drawn following the conserved sequence and helical structure conservation in the alignments of respective CMs, and in a style found in many publications including (ZHAO *et al.* 2018). Orange rectangles highlight putative Sm protein-binding sites. The grey rectangle indicates the 5′ splice-site recognition motif in U1 RNA; the potential base-pairing interactions between the motif and 5′ splice sites of GT-AG and GC-AG introns are shown in the grey box.

Spliced-leader RNA genes

The gene encoding the SL RNA in *D. papillatum* was determined already two decades ago (STURM *et al.* 2001). It is 114 bp-long and includes in its 5′-portion the 39-bp non-coding exon (outron) that will be covalently linked to the 5′- terminus of pre-mRNAs *via trans*-splicing. We found as many as 109 SL-RNA gene copies (at ≥95% sequence identity) in the *Diplonema* nuclear genome assembly. All these genes are part of a tandem repeat, which also contains the 5S rRNA and U2 snRNA genes and occurs in multiple contigs (**Supplementary Figure S11**) (see also the Supplementary Information on repeats). This particular gene organization appears to be shared by all euglenozoans (KELLER *et al.* 1992; SANTANA *et al.* 2001; STURM *et al.* 2001). Our nuclear genome assembly confirms the earlier conjecture that all SL-RNA genes of *Diplonema* occur in SL-5S-U2 repeat clusters. However, we note that hundreds of fragments of the SL RNA gene—a vast majority representing partial 39 bp-long exon portion—are spread across the genome. These do not appear to be functional because the SL sequences capping mRNAs are identical in *Diplonema*, unlike in some dinoflagellates (ALACID *et al.* 2022). Furthermore, because virtually all SL gene fragments occur in intergenic regions and introns of protein-coding genes, we hypothesize that they arose as a collateral damage of retrotransposon mobility, i.e., reverse transcription of mRNAs.

When inspecting mRNA sequences in RNA-Seq data, we realized that the 5′-most A (indicated in lower case in the sequence: 5′–aACCAACGATTTAAAAGCTACAGTTTCTGTACTTTATTG–3′) of the SL sequence was consistently missing. Apparently, the 5′-A fails to be synthesized during the reverse-transcription step of RNA-Seq library construction, because of nucleotide modifications that diplonemids most probably share with kinetoplastids: studies in trypanosomes have revealed a distinctive cap-4 structure at the 5′-terminus of their SL RNA that arises from the methylation of seven sites within its first four nucleotides (AACU). More specifically, the cap consists of m^{7}guanosine -ppp- N^6,N^6,2′-*O*-trimethyladenosine -p- 2′-*O*-methyladenosine -p- 2′-*O*-methylcytosine -p- N^3,2′-*O*-methyluridine, of which the m^{7}guanosine is added by a guanylyltransferase, and the other nucleotides are chemically modified *in situ* (BANGS *et al.* 1992).



Supplementary Figure S11. Repeat unit including the spliced-leader RNA gene of *D. papillatum*. Genes are represented as orange rectangles. Transcribed spacers are shown as dark yellow bars. The sizes of the elements are shown in parentheses. The SL-5S-U2-repeat unit comprises the genes specifying the spliced-leader RNA (SL), the 5S rRNA (5S), and the U2 spliceosomal RNA.

Spliced-leader *trans*-splicing

We identified spliced-leader *trans*-splicing (SLTS) sites by an automated procedure that searched for SL-sequence matches in soft-clipped segments of RNA-Seq reads (see Materials and Methods). Of the 23,720 SLTS sites detected in the genome assembly, about half are located exactly at the 5′ end of gene models, which is expected since SLTS sites were used in the annotation procedure as one of the hints supporting protein-coding gene models. The vast majority of the remaining sites occurred in close vicinity of gene starts, which indicates that many genes possess multiple SLTS sites. The site closest to the 5′ end of a gene model is typically the highest scoring one and thus represents the primary SLTS site.

According to this automated search, only ~41% of *Diplonema* gene models possess at least one associated upstream SLTS site. This suggested initially that only a fraction of the *Diplonema* protein-coding genes require the attachment of an SL, as for example, in *Drosophila* (LASDA AND BLUMENTHAL 2011). Yet, this observation contrasts with the situation in kinetoplastids, in which an SL is *trans*-spliced to each protein-coding gene transcript. Therefore, we re-examined the RNA-Seq read alignment to the three longest and expert-validated contigs of the *D. papillatum* assembly (see Supplementary Information on *D. papillatum* genome assembly and annotation). Upon visual inspection, we found that 90% of the gene models had at least one upstream SLTS site (with 14% of genes having multiple sites). The principal reason for SLTS sites not being recognized by the automated procedure was the short length (≤10 bp) of the soft-clipped SL portion in RNA-Seq reads. About half of the remaining gene models with a seemingly absent SLTS site were preceded by repeats or low-complexity regions that complicated read mapping. The other half had an RNA-Seq read coverage that was low throughout or decreased progressively towards the 5′ end of the gene model. Based on these observations, we conclude that all cytosolic mRNAs of *Diplonema* are decorated with an SL.

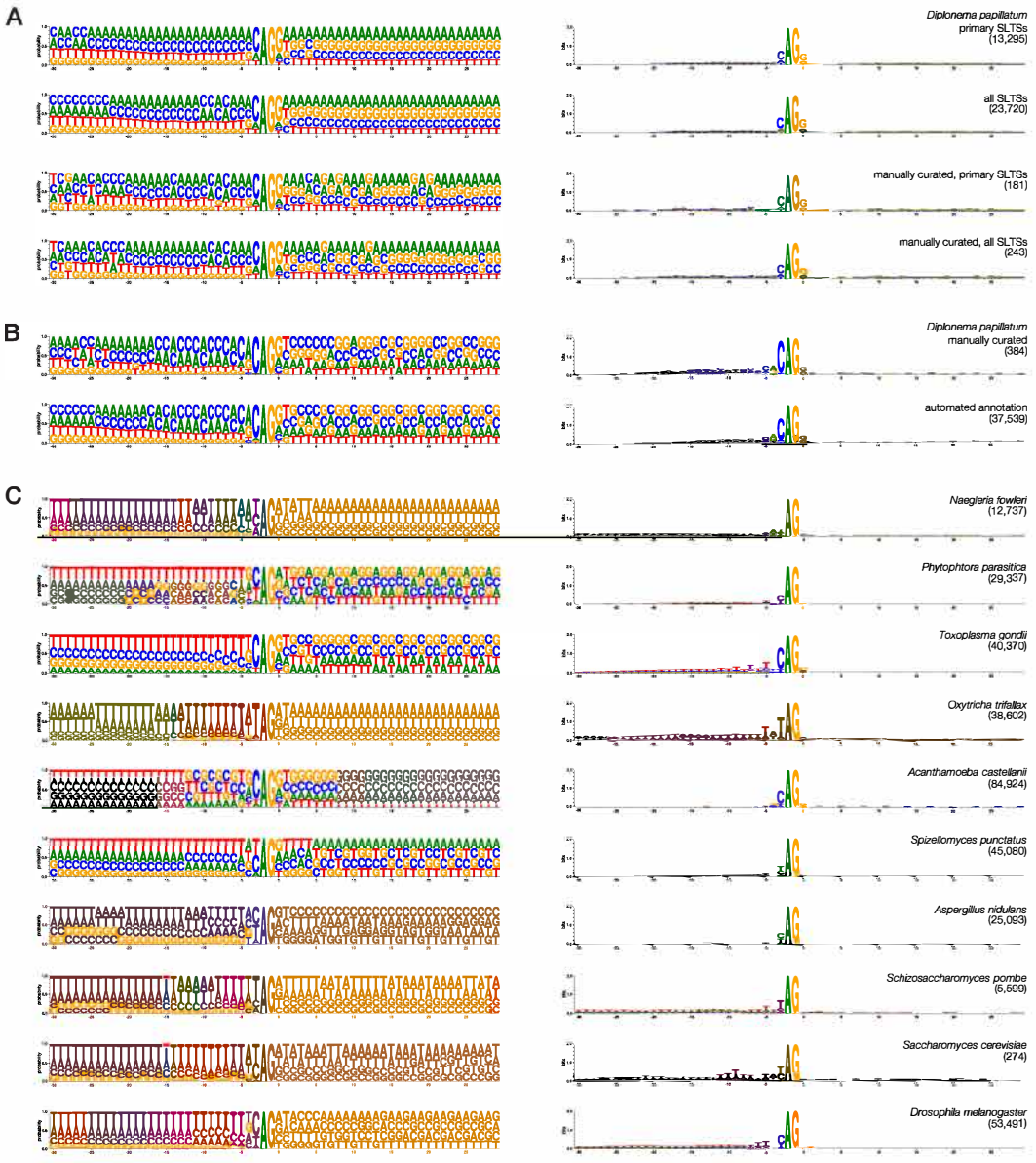
Sequence context of acceptor and donor splice sites in *cis*- and *trans*-spliced introns

In kinetoplastids, a large body of experimental data document that *trans*-splicing of the SL sequence requires only a few SLTS-specific factors, but otherwise, the set of spliceosomal proteins is the same as for splicing (the few) *cis* introns (reviewed in (MICHAELI 2011)). Further, the AG-splice acceptor motif in kinetoplastids has a slight preference for an upstream C residue (DILLON *et al.* 2015). In *Diplonema*, we observed a similar trend at the corresponding position and a bias toward a G residue immediately downstream of the AG acceptor (**Supplementary Figure S12A**). We did not note significant differences in sequence composition between primary and non-primary sites or between automated and expert-curated datasets within a window of 30 bp up and downstream of the SLTS sites.

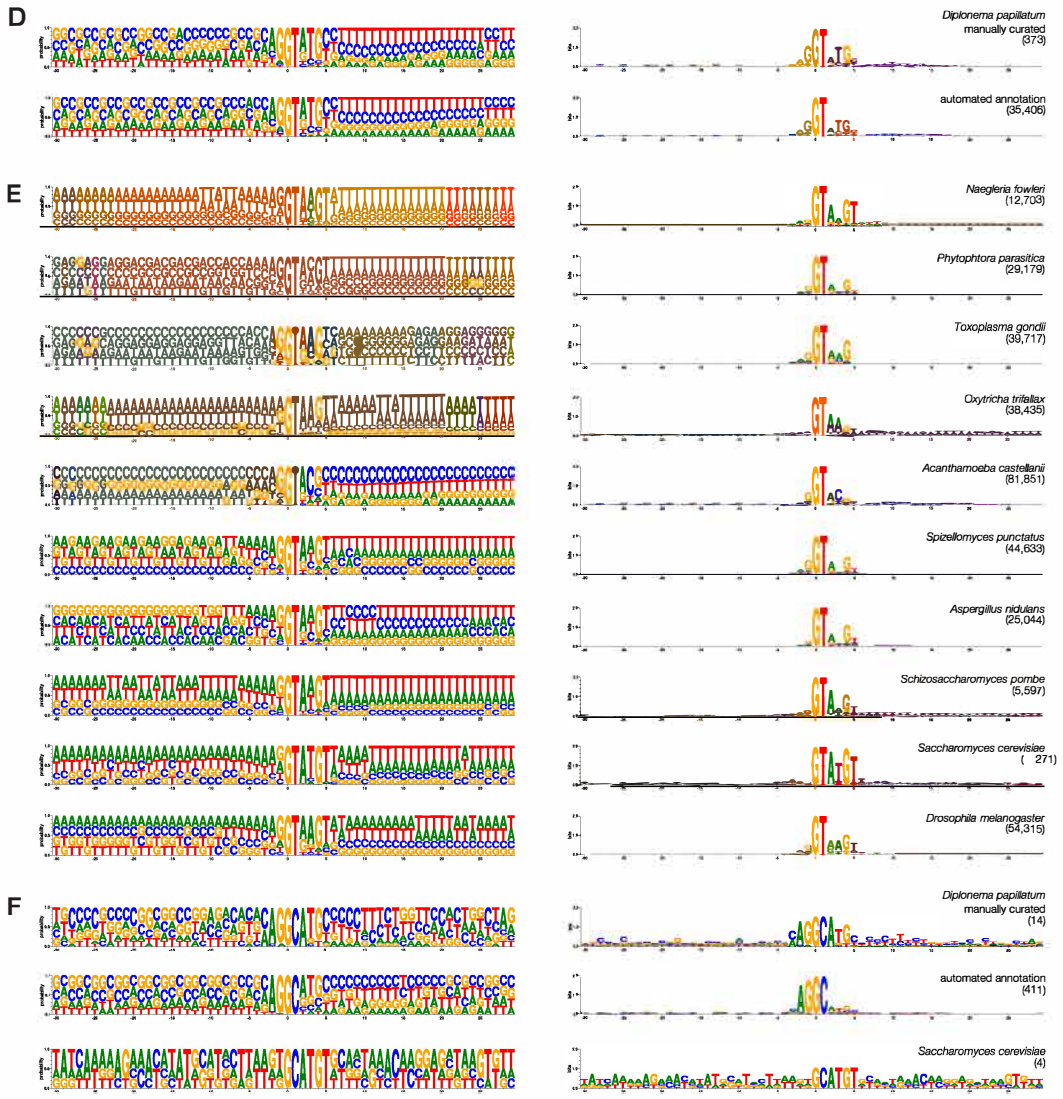
In *D. papillatum,* the upstream-sequence context of the AG-splice-acceptor from *trans*-spliced introns is very similar to that of *cis*-spliced introns, only that the latter have an even more pronounced preference for the upstream C (**Supplementary Figure S12B**). In contrast, the downstream sequences differ, which is expected because they fall into functionally different regions, 5′ UTR and CDS, which in *Diplonema* are slightly more A+G- and C+G-rich, respectively (**Supplementary Figure S12A,B**).

The acceptor site motif of *cis*-introns in *Diplonema* conforms to the 5′–YAG–3′ motif largely conserved across eukaryotes (**Supplementary Figure S12C**). In contrast, the GT-donor site motif of the *Diplonema* GT-type *cis*-introns is slightly more divergent, albeit still within the range of variation observed in other organisms (**Supplementary Figure S12D,E**). The only atypical aspect in *D. papillatum* is a higher than usual frequency of A and G at two positions just upstream of the donor site motif. Interestingly, in almost all of the few GC-AG-type *cis*-introns, the GT-donor site is preceded by an AG (**Supplementary Figure S12F**).

The 5′ splice-site recognition motif of the *Diplonema* U1 RNA is 5′–**CAUACCU**–3′ (see box in **Supplementary Figure S10**), which is reverse complementary to the motif 5′–**aGGTaTG**–3′ at the 5′ boundary of GT-AG introns (lower case indicates less conserved positions). In GC-AG introns of *D. papillatum*, the AG-splice acceptor context is 5′–**AGGCatg**–3′. Thus, the **A-C** mismatch between the **A**4 of the U1-RNA motif and splice acceptor region of GC-AG introns is compensated for by base pairing between **U**7 of the spliceosomal RNA and a strongly conserved **A** in the adjacent exon of this intron type.



Supplementary Figure S12. Sequence logos of splicing acceptor and donor motifs in *Diplonema* and other eukaryotes. Sequence logos are displayed as probabilities or bits (left and right column, respectively). The number of analyzed unique sequences is indicated in parentheses. **A**, Spliced-leader *trans*-splicing (SLTS) acceptor sites in *D. papillatum*. **B**, *Cis*-intron acceptor sites in *D. papillatum*. **C**, *Cis*-intron acceptor sites in selected eukaryotes (see Methods).



Supplementary Figure S12, continued. Sequence logos of splicing acceptor and donor motifs in *Diplonema* and other eukaryotes. Sequence logos are displayed as probabilities or bits (left and right column, respectively). The number of analyzed unique sequences is indicated in parentheses. **D**, GT-type *cis*-intron donor sites in *D. papillatum*. **E**, GT-type *cis*-intron donor sites in selected eukaryotes. **F**, GC-type *cis*-intron donor sites in *D. papillatum* and *Saccharomyces cerevisiae*.

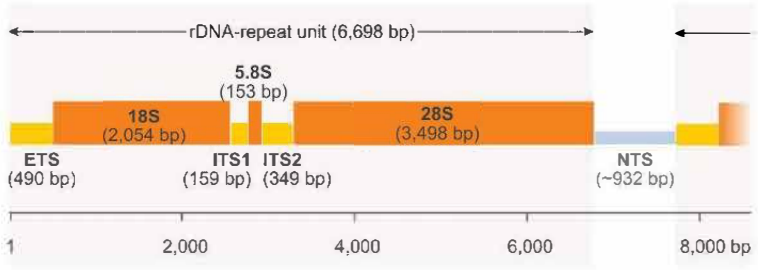
STRUCTURAL RNAS INVOLVED IN PROTEIN SYNTHESIS

Ribosomal RNA genes and rDNA clusters

In eukaryotes, the RNA components of the cytosolic ribosomes are encoded by the following genes: *rns* or 18S associated with the small subunit (SSU), and *rnl* or 25S–28S, *rrn5.8* or 5.8S, and *rrn5* or 5S associated with the large subunit (LSU). The genes *rns*, *rrn5.8*, and *rnl* are arranged in tandem, forming the classical rDNA repeat from which all three components are co-transcribed. In the *D. papillatum* genome assembly, an rDNA repeat unit plus the adjacent spacer has a total length of

7,628 bp (Supplementary Figure S13). The assembly includes a total of 21 such rDNA units (at ≥90% sequence identity) with contigs containing up to two adjacent ribosomal DNA clusters situated in most cases at contig boundaries. This arrangement indicates that in reality, the clusters are more extended but were not assembled due to multiple conflicting solutions.

As mentioned above and similar to the situation in other eukaryotes, the gene of the cytosolic 5S rRNA component is not part of the *rns-rrn5.8-rnl* rDNA repeat unit but is rather located in a separate repeat unit.



Supplementary Figure S13. The ribosomal DNA-repeat unit of *D. papillatum*. Genes are represented as orange rectangles. Transcribed spacers are shown as dark yellow bars, and non-transcribed spacer regions as light-blue bars. The sizes of the elements is shown in parentheses. The rDNA-repeat unit consists of the external transcribed spacer (ETS), two internal transcribed spacers (ITS1, ITS2), and the genes encoding the small subunit rRNA (18S), the 5.8S rRNA (5.8S), and the large subunit rRNA (28S). The non-transcribed spacer (NTS) separates the tandemly arranged rDNA repeat units. Note that in other euglenozoans (kinetoplastids, euglenids), post-transcriptional excision of additional ITSs generates more fragmented cytosolic LSU rRNAs (for details, see for example (MATZOV *et al.* 2020)).

Transfer RNA genes and codon usage

A stringent tRNA-SE scan retrieved 211 tRNA genes in the *D. papillatum* genome assembly. This gene set allows the recognition of all standard amino acids. Several tRNA genes occur in multiple identical copies with numbers varying from two (trnA(UGC)) to 10 (trnK(CUU)). Further, up to five tRNA genes have the same anticodon but a distinct sequence (e.g., trnR(ACG)). The set of tRNA genes represents 47 out of 64 possible anticodons. Isoacceptor tRNAs missing from the *Diplonema* gene complement coincide with those that have been reported absent from many species either across all domains of life or throughout eukaryotes (EHRLICH *et al.* 2021) (Supplementary Table S13). The only exception is the apparent absence of trnL(UAA). It could either be missing in the *D. papillatum* genome assembly, or be a diplonemid or *D. papillatum*-specific gene loss. Since TTA codons do exist in essential nucleus-encoded and mitochondrion-encoded protein-coding genes of *Diplonema*, these codons might be decoded by tRNAs with anticodons other than UAA and carrying nucleotide modifications (JACKMAN AND ALFONZO 2013). Transfer RNA^{Leu}(UAA) is most likely functionally substituted by one of the tRNA^{Leu}(CAA) —for which two identical gene copies exist— provided C34, the wobble position, has been post-transcriptionally deaminated to U. In *Trypanosoma*, an adenosine deaminase complex acting on tRNAs (ADAT2/3) is responsible not only for A-to-I, but also for a C-to-U modification in trnT(AGU) (RUBIO *et al.* 2006; RUBIO *et al.* 2017). The *D. papillatum* genome encodes homologs of both deaminase subunits (DIPPA_16350: ADAT2, and DIPPA_33161: ADAT3), which would allow *Diplonema* to compensate for the missing anticodon by tRNA modification. Even if the process were not efficient, this limitation might not be detrimental since UUA is the most rarely used codon across all *Diplonema* nuclear genes (Supplementary Table S14). Not only UUA, but A+U-rich codons in general are under-represented, with an A+T content of *Diplonema*'s nuclear coding sequences amounting to only 38.72%. The bias towards G+C-rich codons is especially prominent at the third position (74.74%).

Supplementary Table S13. Missing genes for specific isoacceptor tRNAs in the *D. papillatum* genome assembly[a]

Anticodon	AAA	ACA	ACC	ACU	AUA	AUC	AUG	AUU	GAC	GAG	GAU	GCG	GGA	GGC	GGG	GGU	UAA
Amino acid	Phe	Cys	Gly	Ser	Tyr	Asp	His	Asn	Val	Leu	Ile	Arg	Ser	Ala	Pro	Thr	Leu
Taxa[b]	A,B,E	A,B,E	A,B,E	A,B,E	A,B,E	A,B,E	A,B,E	A,B,E	E	E	E	E,B	E	E	E	E	*D. p.*

[a] Data from taxa other than *D. papillatum* taken from (EHRLICH *et al.* 2021)
[b] A, Archaea, B, Bacteria, E, Eukaryota, *D. p.*, *D. papillatum*, this report.

Supplementary Table S14. Codon frequency of *D. papillatum* nuclear protein-coding genes[a]

F	TTT	0.22	S	TCT	0.09	Y	TAT	0.14	C	TGT	0.18
F	TTC	0.77	S	TCC	0.19	Y	TAC	0.85	C	TGC	0.81
L	TTA	0.02	S	TCA	0.08	*	TAA	0.18	*	TGA	0.53
L	TTG	0.15	S	TCG	0.30	*	TAG	0.28	W	TGG	1.00
L	CTT	0.11	P	CCT	0.15	H	CAT	0.20	R	CGT	0.09
L	CTC	0.30	P	CCC	0.26	H	CAC	0.80	R	CGC	0.28
L	CTA	0.03	P	CCA	0.12	Q	CAA	0.30	R	CGA	0.12
L	CTG	0.37	P	CCG	0.46	Q	CAG	0.69	R	CGG	0.28
I	ATT	0.20	T	ACT	0.11	N	AAT	0.18	S	AGT	0.06
I	ATC	0.69	T	ACC	0.33	N	AAC	0.81	S	AGC	0.26
I	ATA	0.10	T	ACA	0.15	K	AAA	0.28	R	AGA	0.08
M	ATG	1.00	T	ACG	0.39	K	AAG	0.71	R	AGG	0.13
V	GTT	0.14	A	GCT	0.14	D	GAT	0.21	G	GGT	0.13
V	GTC	0.39	A	GCC	0.31	D	GAC	0.78	G	GGC	0.43
V	GTA	0.07	A	GCA	0.15	E	GAA	0.40	G	GGA	0.15
V	GTG	0.38	A	GCG	0.38	E	GAG	0.59	G	GGG	0.27

[a] Coding regions of 21,747 assembled mRNAs carrying a 5′ spliced leader

From among the three usual stop codons, UGA appears to be the most frequent in *D. papillatum* (**Supplementary Table S14**). Yet, in addition to the set of genes encoding tRNAs that load standard amino acids, the genome assembly also contains two distinct genes specifying **selenocysteine** tRNA (tRNA^{Sec}(UGA)). In the nascent polypeptide chain, selenocysteine incorporation at UGA codons that otherwise signal translation termination is directed by a particular structural signal in the mRNA (COMMANS AND BÖCK 1999). However, automated gene predictors do not recognize such signals and will infer truncated open reading frames in genes containing selenocysteine codons. Therefore, we inspected the gene models of two proteins known to contain Sec, notably Selenoprotein SelG (DIPPA_01860) and glutathione peroxidase BsaA (DIPPA_26877). Compared to homologs from other taxa, the reading frame of DIPPA_01860 was slightly shorter in the C-terminal region, whereas that of DIPPA_26877 was considerably shorter in its N-terminal region. This is because the start codon chosen as the N-terminus of the latter gene was the first ATG codon downstream of the TGA codon. Replacing the TGA-stop codons in the conceptual translations by "X" clearly improved the alignment of the *Diplonema* inferred proteins with counterparts from other eukaryotes (**Supplementary Figure S14**). This strongly suggests that the cytosolic translation in *D. papillatum* indeed decodes certain TGA codons as Sec.

Supplementary Figure S14. Multiple sequence alignments of selenocysteine-containing proteins. The residue selenocysteine is represented by an 'X'. **A**, Selenoprotein SelG. *D. discoideum*, *Dictyostelium discoideum* (sp|Q55EX3); *D. melanogaster*, *Drosophila melanogaster* (sp|Q7Z2C4), and the tentative homolog from *D. papillatum*. **B**, Glutathione peroxidase BsaA/GPX. *S. aureus*, *Staphylococcus aureus* (sp|Q6GHD0); *A. thaliana*, *Arabidopsis thaliana* (sp|O04922); *B. taurus*, *Bos taurus* (sp|P00435); and the homolog from *D. papillatum*.

Finally, one tRNA gene was identified to carry a UUA anticodon, classifying it as an ochre suppressor codon. However, the predicted codon recognition of this gene is at odds with the observation that the TAA termination codon is used frequently in *D. papillatum* nuclear genes (as well as mitochondrial genes that are translated exclusively with imported, nucleus-encoded tRNAs). Since this potential suppressor-tRNA is an abundant transcript with an RNA-Seq read coverage higher than that of many regular tRNA genes, it must play another yet unknown role, for instance, as a regulator of certain biological processes (RAINA AND IBBA 2014). It should be noted that we only detected transcripts comprising the 5′ half of tRNA(UUA). Such tRNA fragments could arise by premature transcription (e.g., termination at modified bases) during RNA-Seq library construction since we also observed this phenomenon in certain regular *Diplonema* tRNAs. However, the 5′ half of tRNA(UUA) could also be a tRNA-derived RNA fragment (tRFs) produced by specific RNase processing (MEGEL

et al. 2015) and function for example, as a signaling molecule in stress response or as a regulator of gene silencing (MARTINEZ *et al.* 2017).

METHODS

Construction of covariance models used for the identification of spliceosomal RNAs

Most covariance models (CM) available in the RFAM database ((KALVARI *et al.* 2021); https://rfam.xfam.org/) have been built with sequences related to biochemically well-investigated species such as human, yeast and *Arabidopsis*, i.e., they are biased towards animals, fungi, and plants. The bias in taxon sampling makes these models less sensitive when attempting to identify structural RNAs in evolutionarily remote protists, leaving some homologs unrecognized due to structural variation or low sequence conservation. To improve CM sensitivity, we have assembled a broad, taxonomically balanced collection of nuclear genome sequences from GenBank, including select representatives from each of the major eukaryotic groups. For example, Metazoa are represented by five taxa: human, *C. elegans*, a cnidarian, a demosponge, and a placozoan. The final collection includes a total of 207 eukaryotic genome assemblies. The first step of building more sensitive CMs involved searching our genome collection with *cmsearch* (Infernal package; (NAWROCKI AND EDDY 2013)) using a given RFAM model. The –A option of *cmsearch* was applied to produce structured nucleotide sequence alignments of the search results, from which only the first best hit was retained for each species. The alignment was visually inspected, manually corrected for apparent errors, and used to build a new CM with a balanced eukaryotic taxon sampling. For further refinement, the process of searching, alignment, manual curation and CM building was repeated three times. The final CMs almost always have elevated scores for detecting homologs across eukaryotes, particularly in protists. The CMs of spliceosomal U RNAs built by us are available upon request.

Protein sequence search and identification

The identification of protein function was performed as detailed in the **Supplementary Information**: Section 2. Assembly and annotation of the nuclear genome and transcriptome of *Diplonema papillatum*.

Spliced-leader *trans*-splicing acceptor site detection

We detected SLTS sites by exploiting soft-clipped regions of RNA-Seq reads mapped onto the nuclear genome assembly. The SAM file of poly-A RNA reads for a given contig was parsed to retrieve soft-clipped sequences that exactly matched at least 11 of the 3′ terminal nucleotides of the SL sequence (5′–ACCAACGATTTAAAAGCTACAGTTTCT**GTACTTTATTG**–3′).
To each SLTS site we assigned a score corresponding to the number of instances that the soft-clipped sequence matched. If multiple SLTS sites were detected within a 5-kbp region, only a single representative (the one with the highest score) site was selected. The portion of the read that had not been soft-clipped was used to determine the position of the SLTS site in the contig sequence. Because the acceptor site sequence for the SL RNA is AG in the pre-mRNA and the terminal nucleotide of the SL sequence is a G, the SLTS site annotation was placed at the G of the mapped AG acceptor site. Although the G nucleotide is not encoded by the gene but originates from the trans-spliced SL, this choice of representing the SL allowed us to determine an SLTS site presence quickly, because all protein-coding gene models for which we detected an SLTS site start with a G. For statistical purposes, only unique SLTS sites were considered, i.e., if the region 60 bp up and downstream of the detected site was identical (e.g., tandem gene repeats), the site was counted only once.

Genome analysis — Splice-site collection and logo generation

Splice-site datasets of protein-coding genes were mainly collected from the VEuPathDB (https://veupathdb.org) (AMOS *et al.* 2021), except from datasets for *Drosophila melanogaster* (https://flybase.org; (LARKIN *et al.* 2021)), *Saccharomyces cerevisiae* (https://www.yeastgenome.org; (NG *et al.* 2020)), and *Oxytricha trifallax* (https://oxy.ciliate.org; (SWART *et al.* 2013)). Based on available annotations, we retrieved exon-intron boundaries together with up- and downstream flanking regions of ~30 bp, and extracted a sequence representative for each splice site. The sequence logos were created with WebLogo v3.7.4 (CROOKS *et al.* 2004).

AUTHOR CONTRIBUTIONS

Conceptualization, Data curation, Investigation, Formal analysis, Visualization, Writing, original draft – B.F.L., G.B., M.V.; **Writing, review & editing** – all co-authors.

REFERENCES

Alacid, E., N. A. T. Irwin, V. Smilansky, D. S. Milner, E. S. Kilias *et al.*, 2022 A diversified and segregated mRNA spliced-leader system in the parasitic Perkinsozoa. Open Biol 12: 220126.

Amos, B., C. Aurrecoechea, M. Barba, A. Barreto, E. Y. Basenko *et al.*, 2021 VEuPathDB: the eukaryotic pathogen, vector and host bioinformatics resource center. Nucleic Acids Res.

Bangs, J. D., P. F. Crain, T. Hashizume, J. A. McCloskey and J. C. Boothroyd, 1992 Mass spectrometry of mRNA cap 4 from trypanosomatids reveals two novel nucleosides. J Biol Chem 267: 9805-9815.

Commans, S., and A. Böck, 1999 Selenocysteine inserting tRNAs: an overview. FEMS Microbiology Reviews 23: 335-351.

Crooks, G. E., G. Hon, J. M. Chandonia and S. E. Brenner, 2004 WebLogo: a sequence logo generator. Genome Res 14: 1188-1190.

Dillon, L. A., K. Okrah, V. K. Hughitt, R. Suresh, Y. Li *et al.*, 2015 Transcriptomic profiling of gene expression and RNA processing during *Leishmania major* differentiation. Nucleic Acids Res 43: 6799-6813.

Eddy, S. R., 2011 Accelerated profile HMM searches. PLoS Comput Biol 7: e1002195.

Ehrlich, R., M. Davyt, I. López, C. Chalar and M. Marín, 2021 On the track of the missing tRNA genes: a source of non-canonical functions? Front Mol Biosci 8: 643701.

Frey, K. A.-O., and B. A.-O. Pucker, 2020 Animal, fungi, and plant genome sequences harbor different non-canonical splice sites. Cells 9: 458.

Gray, M. W., G. Burger, R. Derelle, V. Klimeš, M. M. Léger *et al.*, 2020 The draft nuclear genome sequence and predicted mitochondrial proteome of *Andalucia godoyi*, a protist with the most gene-rich and bacteria-like mitochondrial genome. BMC Biol 18: 22.

Günzl, A., 2010 The pre-mRNA splicing machinery of trypanosomes: complex or simplified? Eukaryot Cell 9: 1159-1170.

Hastings, K. E., 2005 SL trans-splicing: easy come or easy go? Trends Genet 21: 240-247.

Jackman, J. E., and J. D. Alfonzo, 2013 Transfer RNA modifications: nature's combinatorial chemistry playground. Wiley Interdiscip Rev RNA 4: 35-48.

Kalvari, I., E. P. Nawrocki, N. Ontiveros-Palacios, J. Argasinska, K. Lamkiewicz *et al.*, 2021 Rfam 14: expanded coverage of metagenomic, viral and microRNA families. Nucleic Acids Res 49: D192-d200.

Keller, M., L. H. Tessier, R. L. Chan, J. H. Weil and P. Imbault, 1992 In *Euglena*, spliced-leader RNA (SL-RNA) and 5S rRNA genes are tandemly repeated. Nucleic Acids Res 20: 1711-1715.

Larkin, A., S. J. Marygold, G. Antonazzo, H. Attrill, G. Dos Santos *et al.*, 2021 FlyBase: updates to the *Drosophila melanogaster* knowledge base. Nucleic Acids Res 49: D899-d907.

Larue, G. E., M. Eliáš and S. W. Roy, 2021 Expansion and transformation of the minor spliceosomal system in the slime mold *Physarum polycephalum*. Curr Biol 31: 3125-3131.e3124.

Lasda, E. L., and T. Blumenthal, 2011 Trans-splicing. Wiley Interdiscip Rev RNA 2: 417-434.

Martinez, G., S. G. Choudury and R. K. Slotkin, 2017 tRNA-derived small RNAs target transposable element transcripts. Nucleic Acids Res 45: 5142-5152.

Matzov, D., M. Taoka, Y. Nobe, Y. Yamauchi, Y. Halfon *et al.*, 2020 Cryo-EM structure of the highly atypical cytoplasmic ribosome of *Euglena gracilis*. Nucleic Acids Res 48: 11750-11761.

Megel, C., G. Morelle, S. Lalande, A. M. Duchêne, I. Small *et al.*, 2015 Surveillance and cleavage of eukaryotic tRNAs. Int J Mol Sci 16: 1873-1893.

Michaeli, S., 2011 Trans-splicing in trypanosomes: machinery and its impact on the parasite transcriptome. Future Microbiol 6: 459-474.

Nawrocki, E. P., and S. R. Eddy, 2013 Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29: 2933-2935.

Ng, P. C., E. D. Wong, K. A. MacPherson, S. Aleksander, J. Argasinska *et al.*, 2020 Transcriptome visualization and data availability at the *Saccharomyces* Genome Database. Nucleic Acids Res 48: D743-d748.

Patel, A. A., and J. A. Steitz, 2003 Splicing double: insights from the second spliceosome. Nat Rev Mol Cell Biol 4: 960-970.

Raina, M., and M. Ibba, 2014 tRNAs as regulators of biological processes. Frontiers in Genetics 5.

Rubio, M. A., K. W. Gaston, K. M. McKenney, I. M. Fleming, Z. Paris *et al.*, 2017 Editing and methylation at a single site by functionally interdependent activities. Nature 542: 494-497.

Rubio, M. A., F. L. Ragone, K. W. Gaston, M. Ibba and J. D. Alfonzo, 2006 C to U editing stimulates A to I editing in the anticodon loop of a cytoplasmic threonyl tRNA in *Trypanosoma brucei*. J Biol Chem 281: 115-120.

Russell, A. G., J. M. Charette, D. F. Spencer and M. W. Gray, 2006 An early evolutionary origin for the minor spliceosome. Nature 443: 863-866.
Santana, D. M., J. Lukeš, N. R. Sturm and D. A. Campbell, 2001 Two sequence classes of kinetoplastid 5S ribosomal RNA gene revealed among bodonid spliced leader RNA gene arrays. FEMS Microbiology Letters 204: 233-237.
Sharp, P. A., and C. B. Burge, 1997 Classification of introns: U2-type or U12-type. Cell 91: 875-879.
Sturm, N. R., D. A. Maslov, E. C. Grisard and D. A. Campbell, 2001 *Diplonema* spp. possess spliced leader RNA genes similar to the Kinetoplastida. J Eukaryot Microbiol 48: 325-331.
Swart, E. C., J. R. Bracht, V. Magrini, P. Minx, X. Chen *et al.*, 2013 The *Oxytricha trifallax* macronuclear genome: a complex eukaryotic genome with 16,000 tiny chromosomes. PLoS Biol 11: e1001473.
Tarn, W. Y., and J. A. Steitz, 1996a Highly diverged U4 and U6 small nuclear RNAs required for splicing rare AT-AC introns. Science 273: 1824-1832.
Tarn, W. Y., and J. A. Steitz, 1996b A novel spliceosome containing U11, U12, and U5 snRNPs excises a minor class (AT-AC) intron in vitro. Cell 84: 801-811.
Turunen, J. J., E. H. Niemelä, B. Verma and M. J. Frilander, 2013 The significant other: splicing by the minor spliceosome. Wiley Interdiscip Rev RNA 4: 61-76.
UniProtKB, 2021 UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res 49: D480-d489.
Yan, C., J. Hang, R. Wan, M. Huang, C. C. Wong *et al.*, 2015 Structure of a yeast spliceosome at 3.6-angstrom resolution. Science 349: 1182-1191.
Zhang, X., C. Yan, J. Hang, L. I. Finci, J. Lei *et al.*, 2017 An atomic structure of the human spliceosome. Cell 169: 918-929.e914.
Zhao, Y., W. Dunker, Y. T. Yu and J. Karijolich, 2018 The role of noncoding RNA pseudouridylation in nuclear gene expression events. Front Bioeng Biotechnol 6: 8.

5. Untranslated regions of nuclear genes

INTRODUCTION

The term UnTranslated Region (UTR) refers to the sequence region upstream of the start codon (5′ UTR) and downstream of the stop codon (3′ UTR) of a protein-coding gene; these regions are part of the mature transcript. The 5′ UTR and the 3′ UTR include signals for translation initiation and termination, respectively, and harbor binding sites for small RNAs and proteins that regulate mRNA localization, stability and translation. One of the recently discovered roles of UTRs is the regulation of protein features that are not specified in the amino acid sequence (HINNEBUSCH *et al.* 2016; MAYR 2019; RENZ *et al.* 2020).

In contrast to most eukaryotes that primarily control nuclear gene expression *via* transcription initiation, kinetoplastids —and likely all euglenozoans— transcribe genes in bulk (i.e., polycistronically) and regulate the expression of individual genes post-transcriptionally (reviewed in (CLAYTON 2019)). In these organisms, the control of mRNA stability seems essential. Translation and decay rates of an mRNA are first and foremost influenced by its 3′ UTR, which serves as a landing platform for a wide variety of RNA-binding proteins (RBPs) (KOLEV *et al.* 2014). Messenger RNAs of kinetoplastids have been known to carry 3′ UTRs of similar extended length as those from vertebrates and much longer than reported for other eukaryotes (KOLEV *et al.* 2014; DILLON *et al.* 2015).

RESULTS AND DISCUSSION

To explore the potential of nuclear mRNAs from *Diplonema* to interact with regulatory RNA-binding proteins, we analyzed the length of UTRs and compared them across several eukaryotic species. *Diplonema* stood out in two aspects. While 5′ UTRs are quite short (~70 bp median length) as seen in most eukaryotes, 3′ UTRs are remarkably long, with a median size of ~860 bp. Second, for most organisms, the 3′ UTR tends to be approximately twice as long as the 5′ UTR, but in *Diplonema*, the two median values differ by a factor of ~13 (**Supplementary Table S15, Supplementary Figure S15**). The corresponding analyses were also performed with the subset of about 300 expert-curated gene models. In the case of 5′ UTRs, the values for the curated and non-curated gene set differed only marginally (65 bp vs 70 bp), whereas the automatically predicted 3′ UTRs were about 30% too short (~720 bp vs 970 bp).

In euglenozoans, including diplonemids, 5′ UTRs of genes can be easily inferred *via* detection of spliced-leader acceptor (SLTS) sites. Still, when 3′ UTRs are long, the correct determination of their 3′ terminus is challenging unless RNA-Seq reads in the kbp-size range are available. In kinetoplastids, for example, 3′ UTR sizes have been underestimated by automated annotations because of the difficulty in mapping short transcriptomic reads to repetitive and low-complexity regions that occur in 3′ UTRs (CLAYTON 2019). Because of their importance for gene regulation, mapping the 3′ UTRs will be critical for future functional studies in all euglenozoans.

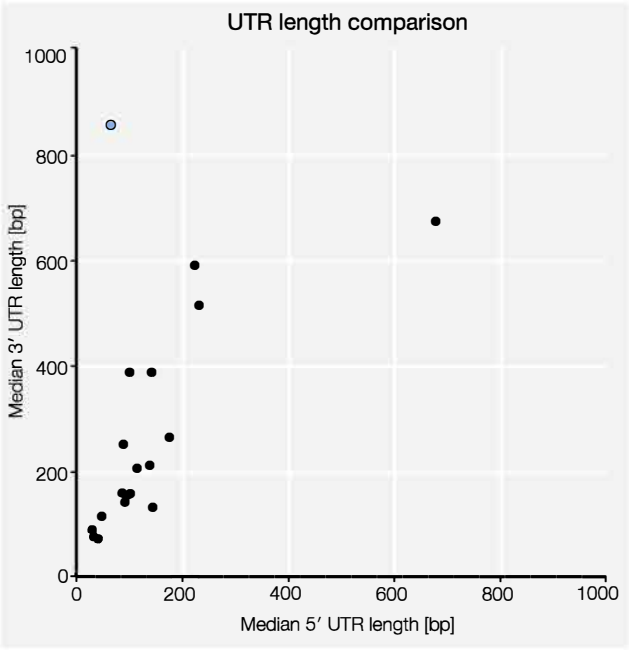
It is possible that what appears to be 3′ UTRs may contain unrecognized coding sequences. For the three longest contigs (see also the **Supplementary Information**: Section 2. Assembly and annotation of the nuclear genome and transcriptome of *Diplonema papillatum*, in particular the section on the expert validation of gene models), we screened 3′ UTRs for additional ORFs >100 codons long. A vast majority consist of very short ORFs that are probably spurious because no homologs were detected in other organisms, including diplonemids (not considering fragments of mobile-element ORFs). While we currently cannot rule out that genuine ORFs occur in long 3′ UTRs, this is probably not a frequent phenomenon in *Diplonema.* Alternatively, the short ORFs might serve a regulatory function, similar to short (mostly upstream) ORFs in fungi or plants (HELLENS *et al.* 2016; SCHLESINGER AND ELSÄSSER 2022). Resolving this issue will require the future application of experimental techniques such as ribosome profiling and deep proteomics.

Alternative processing events add another layer of complexity to the picture. For example, in *Leishmania*, the predominantly used SLTS and poly-adenylation (PA) sites are generally closer to CDS boundaries than the alternative sites, which seem to play important roles in the cell (DILLON *et al.* 2015). In *D. papillatum* we detected alternative SLTS sites in ~10% of genes. In addition, the observed variation in RNA-Seq coverage at the 3′ end of certain *Diplonema* genes may reflect alternative transcript processing events, indicating that gene regulation in this organism is at least as complex as in trypanosomes.

Supplementary Table S15. Length of UTRs in nuclear genes from *D. papillatum* and other eukaryotes.

Organism	Lineage	5'UTR median length [bp] [a]	3'UTR median length [bp] [b]	Gene number [c]	Source [d]
Aspergillus nidulans	Fungi; Ascomycota	116	208	6,037	FungiDB
Saccharomyces cerevisiae	Fungi; Ascomycota	49	116	2,849	SGD
Schizosaccharomyces pombe	Fungi; Ascomycota	177	267	4,524	FungiDB
Cryptococcus neoformans	Fungi; Basidiomycota	93	143	6,181	FungiDB
Allomyces macrogynus	Fungi; Blastocladiomycota	145	133	9,088	FungiDB
Rhizophagus irregularis	Fungi; Glomeromycota	100	157	11,216	FungiDB
Spizellomyces punctatus	Fungi; Chytridiomycota	103	159	6,331	FungiDB
Homo sapiens	Metazoa; Chordata	225	593	21,838	UTRDB; (GRILLO *et al.* 2010)
Drosophila melanogaster	Metazoa; Arthropoda	140	214	16,671	FlyBase
Phytophthora parasitica	SAR; Oomycota	88	160	6,986	FungiDB
Saprolegnia diclina	SAR; Oomycota	42	74	9,053	FungiDB
Toxoplasma gondi [e]	SAR; Apicomplexa	680	675	5,194	ToxoDB; (HASSAN *et al.* 2012)
Vitrella brassicaformis	SAR; Chromerida	90	253	16,306	CryptoDB
Oxytricha trifallax	SAR; Ciliophora	34	78	17,040	(SWART *et al.* 2013)
Leishmania major	Discoba; Euglenozoa	233	517	8,841	(DILLON *et al.* 2015)
Trypanosoma brucei	Discoba; Euglenozoa	102	389	5,474	TriTrypDB
Diplonema papillatum [f]	Discoba; Euglenozoa	66	857	15,369	this work

[a] Only 5' UTRs >10 bp were considered.
[b] Only 3' UTRs >20 bp were considered.
[c] The number of genes used for the analysis, selected by the criteria that both UTRs were annotated and had a 5' and 3' UTR length of ≥10 bp and ≥20 bp, respectively.
[d] References: FungiDB, https://fungidb.org/; SGD, http://www.yeastgenome.org ; UTRDB, http://utrdb.ba.itb.cnr.it/home/download ; FlyBase, https://flybase.org/ ; ToxoDB, https://toxodb.org/ ; CryptoDB, https://cryptodb.org/ ; TriTrypDB, https://tritrypdb.org/.
[e] *Toxoplasma gondi* is exceptional for having very long 5' and 3' UTRs with nearly identical median-length values.
[f] The numbers for *D. papillatum* refer to uncurated genes.



Supplementary Figure S15. Length of UTRs in nuclear genes from *D. papillatum* and other eukaryotes. The plot compares the length of 5′ and 3′ UTRs; *Diplonema* is highlighted in blue. For details, see (**Supplementary Table S15**).

MATERIALS AND METHODS

Data on untranslated regions (UTRs) of protein-coding genes were collected from several databases and publications (see **Supplementary Table S15**). Several filtering steps were implemented to ensure consistency across datasets. First, we considered only 5′ and 3′ UTRs of ≥10 bp and ≥20 bp length, respectively. Second, both 5′ and 3′ UTR had to be mapped for a gene to be considered. To avoid duplication, we focused on primary (predominantly used) mRNA processing sites; 100% identical UTR sequences were removed from *Saccharomyces cerevisiae* and *Homo sapiens* data, but otherwise, for practical reasons, alternative UTRs were kept. Lastly, we included only species for which the protein-coding genes that passed all filtering stages represented at least 50% of all genes in the corresponding nuclear genome. In addition to *Diplonema*, the final dataset contained 15 species representing the Fungi, Metazoa, SAR and Discoba lineages. We then calculated the median 5′ and 3′ UTR lengths for comparative purposes.

AUTHOR CONTRIBUTIONS

Conceptualization, Data curation, Investigation, Formal analysis, Visualization, Writing, original draft – M.V.; **Writing, review & editing –** all co-authors.

REFERENCES

Clayton, C., 2019 Regulation of gene expression in trypanosomatids: living with polycistronic transcription. Open Biol 9**:** 190072.

Dillon, L. A., K. Okrah, V. K. Hughitt, R. Suresh, Y. Li *et al.*, 2015 Transcriptomic profiling of gene expression and RNA processing during *Leishmania major* differentiation. Nucleic Acids Res 43**:** 6799-6813.

Grillo, G., A. Turi, F. Licciulli, F. Mignone, S. Liuni *et al.*, 2010 UTRdb and UTRsite (RELEASE 2010): a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. Nucleic Acids Res 38**:** D75-80.

Hassan, M. A., M. B. Melo, B. Haas, K. D. Jensen and J. P. Saeij, 2012 *De novo* reconstruction of the *Toxoplasma gondii* transcriptome improves on the current genome annotation and reveals alternatively spliced transcripts and putative long non-coding RNAs. BMC Genomics 13**:** 696.

Hellens, R. P., C. M. Brown, M. A. W. Chisnall, P. M. Waterhouse and R. C. Macknight, 2016 The emerging world of small ORFs. Trends Plant Sci 21: 317-328.
Hinnebusch, A. G., I. P. Ivanov and N. Sonenberg, 2016 Translational control by 5'-untranslated regions of eukaryotic mRNAs. Science 352: 1413-1416.
Kolev, N. G., E. Ullu and C. Tschudi, 2014 The emerging role of RNA-binding proteins in the life cycle of *Trypanosoma brucei*. Cell Microbiol 16: 482-489.
Mayr, C., 2019 What are 3' UTRs doing? Cold Spring Harb Perspect Biol 11.
Renz, P. F., F. Valdivia-Francia and A. Sendoel, 2020 Some like it translated: small ORFs in the 5′UTR. Experimental Cell Research 396: 112229.
Schlesinger, D., and S. J. Elsässer, 2022 Revisiting sORFs: overcoming challenges to identify and characterize functional microproteins. Febs j 289: 53-74.
Swart, E. C., J. R. Bracht, V. Magrini, P. Minx, X. Chen *et al.*, 2013 The *Oxytricha trifallax* macronuclear genome: a complex eukaryotic genome with 16,000 tiny chromosomes. PLoS Biol 11: e1001473.

6. Repetitive sequences in the nuclear genome of *Diplonema papillatum* (assembly v_1.0)

INTRODUCTION

Nuclear DNA of almost all eukaryotes contains a considerable portion of **intergenic regions**, from 70% in *Arabidopsis thaliana* (Arabidopsis Genome AGI 2000) to 30% in *Saccharomyces cerevisiae* (WOOD *et al.* 2001) and *Trypanosoma brucei* (ASLETT *et al.* 2010). A major portion of intergenic regions consists of sequence motifs that occur multiple times throughout the genome.

Repetitive DNA is commonly subdivided into two major families, tandem repeats and dispersed repeats (RICHARD *et al.* 2008). **Tandem repeats** are defined as a sequence motif of two or more nucleotides repeated multiple times adjacently head to tail. In more general terms, tandem repeats are a subclass of low-complexity regions (LCRs), since the latter category is defined as arrays of sequence patterns with biased composition (GGT-GTT-TTG; or GGT-GGT-GGT) and thus comprises arrays of identical repeat units. Long tandem repeats can contain multiple copies of a protein-coding gene or of the rDNA cluster consisting of the small subunit (SSU) rRNA, 5.8S rRNA, and large subunit (LSU) rRNA. Tandem repeats not including genes are often referred to as Satellite DNA.

Satellite DNA is subdivided into three subclasses according to the length of the repeat unit (note, however, that there is no strict convention as to size ranges): unit sizes of 2-10 bp define **microsatellites**, 11 to 99 bp-long units are categorized as **minisatellites**, and arrays of 100 bp and beyond as **macrosatellites**. Microsatellites, also denoted 'short tandem repeats' (STRs), include the much-studied trinucleotide repeats occurring within protein-coding regions, since retractions and expansions of such arrays have been linked to human disease (RAMAKRISHNAN AND GUPTA 2021). Minisatellites are typically located in intergenic and subtelomeric regions, whereas macrosatellites are prevalent in centromeric regions of nuclear chromosomes (MELTERS *et al.* 2013).

The second family of repetitive DNA encompasses the **dispersed** (or interspersed) **repeats**, i.e., sequence-motif copies that occur in different locations of the genome. Many of the dispersed repeat units have no apparent origin or function, and others include tRNAs, rDNA, gene copies, and transposons. The notion of 'junk DNA' or 'selfish elements' alludes to repetitive sequences being vestiges of self-replicating and mobile elements (OHNO 1972; ORGEL *et al.* 1980). However, in the meantime, certain genomic repeats have been shown to play a role in genetic variation and gene regulation, and they have been suggested to shape the three-dimensional folding of nuclear DNA (MEHROTRA AND GOYAL 2014). Locating and characterizing tandem repeats by bioinformatics methods remains a challenge because the various unit copies may contain substitutions and indels, and because of complex, nested structures of repeat arrays. As imperfect, degenerated repeat units are modelled differently by the various repeat-detection algorithms (e.g., RepeatMasker (SMIT AND GREEN), RepeatScout (PRICE *et al.* 2005) and Tandem Repeats Finder (BENSON 1999)), it is not surprising that the results for the same sequence but analyzed by different software often differ.

RESULTS AND DISCUSSION

Several publications have used RepeatMasker (SMIT AND GREEN) to analyze repeats in a genome, since it returns information about low-complexity regions, simple repeats, and satellites in a readily readable form. However, the RepeatMasker documentation explicitly states that the program is not suited for this task; its purpose is masking repeat regions in the genome prior to structural annotation to avoid spurious matches in database searches. To get an appreciation of the repeat content in the *D. papillatum* nuclear genome, we employed dedicated tools to identify tandem and interspersed repeats.

With less than 6%, **tandem repeats** of unit length 2 to 50 bp cover only a small portion of the *Diplonema* nuclear genome (**Supplementary Table S16**). The longest tandem array consists of 409 copies of a 48-bp unit that extends to nearly 20 kbp (tig00023097_5, positions 169,920-189,538) and is located in an intergenic region without any evidence for transcription. Homopolymers (1-bp units) of either A, C, G, or T were also found. Mononucleotide tracts longer than 600 bp turned out to be artifacts introduced by single PacBio reads. Still, shorter tracts have reliable long-read support, such as the 169-bp long T-run in contig tig00023828_2, positions 11,191-11,359.

Dispersed repeats represent the predominant type in the *D. papillatum* nuclear genome and constitute as much as 46% of the assembly. The number of distinct dispersed repeats detected by the RepeatScout tool is nearly 10,000 with a length of up to nearly 20 kbp long (e.g., R=44: 19,085 bp). The unit with the largest number of copies (≥90% identity) is R=0, a 226-bp long motif that occurs 6,030 times in the genome.

Tandem repeats rarely contain coding regions of validated protein-coding genes, but instead include parts or complete spurious gene models that lack support by RNA-Seq data. This contrasts with structural RNAs, which are often

arranged tandemly. In addition, tandem arrays can be part of dispersed repeats that occur in multiple copies across the genome. For example, half of the dispersed repeat R=121 (15,715 bp) is made up of the rDNA repeat unit that is composed of the genes for small subunit (SSU or 18S) rRNA, 5.8S rRNA, and large subunit (LSU or 28S) rRNA; (the genes are referred to as *rnl*, *rns*, *rrn5*, and *rn5.8*). R=121 is found nine times in the genome assembly, when considering copies of ≥90% identity. Similarly, the dispersed repeats R=75 (18,233 bp) and R=163 (7,201 bp), which occur in the assembly 4 and 7 times (at ≤90% identity), respectively, include tandem repeats of alternating 5S rRNA, spliced-leader RNA and spliceosomal U2 RNA-genes. Structural RNA genes identified in the *D. papillatum* nuclear genome assembly and their arrangement in repeat arrays are described in more detail in the **Supplementary Information**: Section 4. Intron splicing and structural RNAs.

About 2% of dispersed repeat motifs include small **pieces** of **rRNA** genes together with unrelated sequences of unknown identity. Among the four rRNA genes, fragments of *rnl* are most abundant, followed by *rns*, with pieces of 30-200 bp occurring ~2,400 and 1,100 times in the genome sequence, respectively.

Since **introns** in *Diplonema* nuclear genes can be several dozen kbp long, we investigated whether they include long dispersed repeats. One of the longest intronic regions covered by dispersed repeats is included in the gene DIPPA_27122 encoding a homolog of cholesterol-7-desaturase. The longest repeat unit that falls in the 9,167-bp long intron 1 is R=510 (7,674 bp), covering 84% of the intron 1. When considering copies of ≥90% sequence identity, R=510 occurs in 21 other positions in the *Diplonema* genome, totalling to six located in introns and 15 in untranscribed, intergenic regions. Introns of similar sequence but located in distinct genes are suggestive of containing **mobile genetic elements**. Conceptual translation indeed revealed a 1,866 amino acid-long reading frame (ORF; included in R=510 and named orf_R_510) carrying three protein domains diagnostic for mobile elements: an Exo_endo_phos domain in the N-terminal third of the sequence, an RT_like superfamily domain in the middle, and an RNase_H_like domain in the C-terminal third of the protein. A Pfam domain search retrieved RVT_1, RVT_3, Exo_endo_phos, and RNase_H. The search for free-standing sequences (i.e., outside introns) in the inferred *D. papillatum* proteome that are similar to orf_R_510 retrieved proteins from 1,008 distinct gene models. The corresponding function annotations indicated either 'Retrovirus-related Pol polyprotein', 'Ribonuclease HI', 'RNA-directed DNA polymerase from mobile element', or 'LINE-1 reverse transcriptase homolog'. A search for nucleotide sequences in the *D. papillatum* nuclear genome that can be conceptually translated into orf_R_510-like proteins retrieved another 500 matches, of which 21 displayed 100% protein-sequence identity.

The above result is consistent with the extraordinary expansion of transposable element (TE)-ORFs detected in the analysis of gene family evolution in *D. papillatum*, indicating that the *Diplonema* nuclear genome is permeated with ORFs typically encoded by transposable elements. Indeed, the analysis of the function assignments of the inferred proteins revealed signatures of both transposon classes—retrotransposons and DNA transposons— and a total of 12 subclasses. Retrotransposon ORFs make up the majority of subclasses (75%) and the predominant portion of instances (95%; **Supplementary Table S17**). Retrospectively, the observed repetitiveness and complex repeat arrangement of the *D. papillatum* nuclear genome explains why the sequence assembly engines did not generate longer and fewer contigs. In turn, the fragmentation of the current assembly represents a limitation in the rigorous characterization of repeats.

Supplementary Table S16. Repeat analysis[a].

Tool	Interspersed classified repeats	Dispersed unclassified repeats	Satellite DNA/ Tandem repeats
TandemRepeatsFinder	/	/	Count: 100,426 LenRange: 2-50 bp CumulLen: 17,042,328 bp **(6%)** ArrayLenRange: 25-29,859 bp
RepeatMasker	Count: 2,412[b] CumulLen=264,094 bp	/	Count: 93,490 LenRange: 1-6 bp[c] CumulLen: 5,388,838 bp[d]
RepeatScout (repeat unit detection) & glsearch (mapping of units to genome)	/	Count: 9,783 LenRange: 51 - 19,085 bp MedLen=236 bp CumulLen: 128,855,866 bp **(46%)**	/

[a] **Count**: number of motif instances; **LenRange**: length range of repeat motifs (units); **MedLen**=median length of repeat units; **CumulLen**, total (cumulative) length of all motif instances, after merging; **%** of total genome (size 280,385,187 bp); **ArrayLenRange**: length range of tandem arrays.
[b] RepeatMasker-reported types of retrotransposons and mobile DNA elements known from human: 1,577 LINEs, 321 SINEs, 22 LTRs; 492 DNA elements. A compilation of Euglenozoan transposable elements is not available.
[c] RepeatMasker uses a fixed unit length of 6 bp, referred to as simple repeats.
[d] These numbers are considerably smaller than NSEG and TandemRepeatsFinder's results because RepeatMasker has not been designed to detect repeats or low-complexity DNA regions comprehensibly.

Supplementary Table S17. Similarity of inferred *D. papillatum* proteins to ORFs of transposable elements

Transposon class	Element designation	Count	Explicit function annotation
Retrotransposons (Class I transposons)	Copia-like	2	product=Retrovirus-related Pol polyprotein from transposon 17.6
	BS	115	product=Probable RNA-directed DNA polymerase from transposon BS; note=Reverse transcriptase (RNA-dependent DNA polymerase); product=Probable RNA-directed DNA polymerase from transposon BS; note=Endonuclease/Exonuclease/phosphatase family; product=Probable RNA-directed DNA polymerase from transposon BS; note=Astacin (Peptidase family M12A);
	LINE	245	product=LINE-1 reverse transcriptase homolog; product=LINE-1 retrotransposable element ORF2 protein;
	Opus	1	product=Retrovirus-related Pol polyprotein from transposon opus
	R1	1	product=Retrovirus-related Pol polyprotein from type-1 retrotransposable element R1;
	R2	304	product=Retrovirus-related Pol polyprotein from type-1 retrotransposable element R2;
	R2DM	212	product=Retrovirus-related Pol polyprotein from type-2 retrotransposable element R2DM;
	X	266	product=Probable RNA-directed DNA polymerase from transposon X-element; note=Reverse transcriptase (RNA-dependent DNA polymerase)
	SLACS	66	product=Retrotransposable element SLACS 132 kDa protein; product=SLACS reverse transcriptase, putative [Trypanosoma equiperdum]; product=SLACS reverse transcriptase, putative [Trypanosoma theileri];
All retrotransposons		1,212	
DNA transposons[b] (Class II transposons)	MULE	46	product=hypothetical protein; note= MULE transposase domain; product=hypothetical protein; note=Transposase, Mutator family;
	TATE	12	product=TATE DNA Transposon [Trypanosoma theileri]; note=Phage integrase family;
	Tc5	1	product=hypothetical protein; note=Tc5 transposase DNA-binding domain;
All DNA transposons		69	

[b] The nine *D. papillatum* proteins carrying the annotation "product=hypothetical protein; note=containing Helitron helicase-like domains" are not listed here, because the relationship of these proteins to Helitron transposons is uncertain.

NUMTs in the *D. papillatum* nuclear genome

A particular subtype of dispersed repeats in the nuclear genome of *D. papillatum* is NUclear MiTochondrial segments (NUMTs). These regions originate from intracellular transfer of the mitochondrial DNA (mtDNA) to the nucleus. NUMTs have been reported in almost all eukaryotic genomes studied so far, but the number and sizes of transferred mtDNA segments vary considerably (KLEINE *et al.* 2009). It is important to distinguish NUMTs from originally mitochondrion-encoded genes that have migrated to the nucleus, a process that took place during the transformation of the endosymbiotic alpha-proteobacterium to an integral organelle of the eukaryotic cell. These relocated genes code for mitochondrial components, implying that the corresponding gene products are targeted to organelles. NUMTs in contrast, are considered not to be expressed ((HAZKANI-COVO *et al.* 2010) but see (NOUTSOS *et al.* 2007; ROGERS AND GRIFFITHS-JONES 2012; WANG *et al.* 2014)). Here we examined to what degree the incorporation of mtDNA has contributed to the large size of the *Diplonema* nuclear genome.

Our search for NUMTs in the *D. papillatum* nuclear genome took into consideration the unusual **mitochondrial genome architecture and gene structure** in this organism, which appears to be a shared feature of diplonemids (KIETHEGA *et al.* 2011; VALACH *et al.* 2017; KAUR *et al.* 2020). Mitochondrial DNA in *Diplonema* is organized in about 80 distinct circular chromosomes of 7-8 kbp length. Only 40-550 bp of these circles is coding, constituting fragments of genes, also called 'modules'. Gene modules are flanked by short stretches of unique sequence ('flanking sequence'). In contrast, the remaining 90% of a mitochondrial chromosome, referred to as 'constant region', is common to all chromosomes in the *Diplonema* mitochondrial genome (KIETHEGA *et al.* 2013; MOREIRA *et al.* 2016; VALACH *et al.* 2016). Constant regions are characterized by tandem repeats and low sequence complexity.

Approximately **1.2%** of the ***D. papillatum* nuclear genome** assembly v1.0 consists of NUMTs. This percentage is likely an underestimation because NUMTs are known to accumulate mutations leading to divergent sequences that are difficult to spot. Therefore, only the relatively recently transferred NUMTs will be recognized. The NUMTs that we detected in the *Diplonema* nuclear genome fall into two categories (i) "anchored" NUMTs representing high-confidence matches including mitochondrial gene modules; and (ii) constant-region NUMTs (cr-NUMTs) that exclusively contain portions of the constant region of mitochondrial chromosomes.

We retrieved 60 **anchored NUMTs** with a cumulative length of ~82.8 kbp. These NUMTs contain sequences from 25 distinct modules. As many as 11 NUMTs represent entire mitochondrial chromosomes; the accuracy of the nuclear genome assembly in these regions was confirmed by inspecting long PacBio reads. Additional 26 NUMTs consist of a part of a gene module, and another 23 also contain segments of flanking or constant regions. NUMTs comprising partial mitochondrial chromosomes add up to ~9.7 kbp (**Additional File 2** numts). Most anchored NUMTs are located in intergenic regions of nuclear contigs or in an intron of a nuclear gene; intronic location also applied to two of the NUMTs consisting of full-length mitochondrial chromosomes. Six anchored NUMTs reside within UTRs of nuclear genes and are transcribed together with the rest of the host genes. Interestingly, about one third of anchored NUMTs are arranged in tandem repeats. The longest array, residing in contig tig00023746_1, is 485 bp long and consists of eight copies of a ~40-bp long portion of module X18, which is an unidentified mitochondrial gene or gene fragment (MOREIRA *et al.* 2016). A somewhat shorter array (379 bp) occurs in tig00023753 and contains four copies of a 95-bp piece of module #1 of the mitochondrial LSU RNA (*rnl*-m1).

Searches for **cr-NUMTs** retrieved 1,359 instances with a median size of 128 bp and totaling nearly 260 kbp. Certain cr-NUMTs are also arranged in tandem. The longest array consists of 164 copies of a 68-bp motif and makes up an 11-kbp region within an intron of the gene DIPPA_33001. Nevertheless, most cr-NUMTs are dispersed individually across the nuclear genome, just as their anchored counterparts. They occur predominantly in intergenic regions (1,097 instances) and non-coding segments of genes (168 in introns, 95 in UTRs). Only five cr-NUMTs overlap by up to 120 bp a protein-coding region, usually at a CDS terminus (**Additional File 2** numts).

The total length and proportion of the *Diplonema* nuclear DNA whose origin can be traced to mtDNA is within the range published for **other organisms**, especially animals and plants (MICHALOVOVA *et al.* 2013; KO AND KIM 2016; CALABRESE *et al.* 2017). However, it should be noted that the reported contribution of NUMTs to nuclear genomes is generally an approximation because the underlying assemblies, unless built from long reads, are probably not entirely correct. Moreover, even within a given species, NUMTs might vary across ecotypes, as was recently documented in *A. thaliana*. The commonly studied ecotype Columbia-0 contains an enormous NUMT corresponding to the entire mitochondrial genome, which was confirmed by fluorescence *in situ* hybridization (FISH) on metaphase chromosomes (STUPAR *et al.* 2001). In contrast, recent nuclear genome sequencing of the two ecotypes Niederzenz-1 and Landsberg by a long-read technology revealed the absence of these insertions (PUCKER *et al.* 2019).

METHODS

Repeat identification

RepeatMasker v.4.1.1. (http://www.repeatmasker.org/) was executed in sensitive mode, run with rmblastn v.2.9.0+. TandemRepeatsFinder v.4.09 (BENSON 1999) was used with the command line parameters 2 7 7 80 10 50 2000 -h -m –ngs, i.e. with a maximum unit length of 2,000 bp. The corresponding output was filtered for further analyses to retain the repeat units of up to 50 bp. Low-complexity repeats including satellites with units <500 bp were identified with the tool nseg v1.0.0 (Wooton and Federhen, 1993), but only used for comparison with the output of TandemRepeatsFinder, without further following up the nseg results. *De novo* detection of dispersed repeats was performed with RepeatScout v.1.0.5 (PRICE *et al.* 2005). The first step is the generation of an k-mer (lmer) table

```
$ build_lmer_table -s sequence Dp_v1.0.fasta -freq lmer_table_v1.0.txt
```

that serves as input for RepeatScout

```
$ RepeatScout -sequence Dp_v1.0.fasta -output repS_v1.0.fasta -freq lmer_table_v1.0.txt
```

The output (`repS_v1.0.fasta`) is then filtered by the script filter-stage-1.prl that is provided with the RepeatScout distribution to remove from the initially determined repeat motifs those that consists of tandem repeats and low-complexity regions:

```
$ cat repS_v1.0.fasta | /share/supported/apps/RepeatScout-1.0.5/filter-stage-1.prl  > repS_v1.0-filter1.fasta
```

The output is a collection of distinct dispersed-repeat sequences but does not include their positions in the genome sequence. The authors of RepeatScout recommend mapping the repeat sequence to the genome by using RepeatMasker and providing the repS sequences as database. We generated a gff file that lists repS positions in the genome with the following command:

```
$ RepeatMasker -s -pa 38 -engine ncbi -dir .  -lib repS_v1.0-filter1.fasta  -gff -xsmall Dp_v1.0.fasta
```

However, the result was unsatisfactory because certain repeats were not identified in all expected positions. For finding repeat copies exhaustively, we used a global-query and local-database similarity search, notably glsearch of the fasta suite v.36.3.8g (PEARSON 2016). Glsearch allows defining the sequence identity threshold of repeat copies reported:

```
$ glsearch36 -T 40 -E 1.0e-5 -m8CB repS_v1.0-filter1.fasta  Dp_v1.0.fasta  >glsearch-out-repS_v1.0-filter1-vs-Dp_v1.0-e-5
```

The chosen generous E-value allows detection of copies of the shortest repS motifs (50 bp) at a minimum identity of ~60%. The output file was then filtered to extract motif copies of a given percentage of identity. We experimented with 75%, but

the subsequent analyses were performed with repeat copies of at least 90% identity. From the file `glsearch-out-repS_v1.0-filter1-vs-Dp_v1.0-0.9ident` we generated bed and gff files with in-house scripts, and calculated the number of occurrences (frequency) at which the repeat units occur in the genome sequence. To calculate the length and proportion of the genome covered by repeats, intervals covered by different repeats were merged with the BEDTools utility merge (BEDTools v.2-2.29.2 (QUINLAN 2014):

```
$ bedtools merge -i repS_v1.0-filter1-0.9ident.bed -c 4 -o distinct > repS_v1.0-filter1-0.9ident-merged.bed
```

To determine the overlap of repS copies with genes, exons, and introns, we generated from the conventional genome gff file (Dp_v1.0.gff3) one that lists only protein-coding genes and indicates introns explicitly. For that we used the GenomeTools system v.1.6.1 (http://genometools.org/tools.html) embedded in an awk script:

```
$ awk 'BEGIN {FS=OFS="\t"}s {split($9, a, "[;=]"); $9="gene_id \""a[4]"\"; transcript_id \""a[4]"\""}1'
Dp_v1.0.gff3 | gt gtf_to_gff3 -tidy 2>/dev/null | gt gff3 -sort -tidy -retainids -addintrons |grep -v '^#' >
out
```

Subsequently, we replaced the gene identifiers in the `out` file, which were introduced by the command `gt gtf_to_gff3`, with our gene identifiers DIPPA_<number>. The final gff file `Dp_v1.0-DIPPA+introns.gff3` contains explicit intron, exon, and gene features. This latter file was used to determine the overlaps of individual repeats with exons or introns employing BEDTools intersect:

```
$ bedtools intersect -wo -a repS_v1.0-filter1-0.9ident.bed -b Dp_v1.0-DIPPA+introns.gff3 > repS_v1.0-filter1-
0.9ident-intersect-genes
```

To determine the portions of genes, exons, or introns covered by repeats, the unmerged bed file was used as input file specified by –a. Finally, selected, long dispersed repeat motifs were conceptually translated with a home-made script and protein sequences >= 1,000 residues were retained. For example, the protein sequence file `orf_R=510.faa` represents the longest reading frame contained in the dispersed repeat R=510 detected by RepeatScout. Sequences in the inferred *Diplonema* proteome with similarity to `orf_R=510.faa` were detected with phmmer v.6.1 (EDDY 2011) using an E-value threshold of 1.0 e-50. Protein domains were retrieved from the output of BLAST searches (ALTSCHUL *et al.* 1990) in the non-redundant database at the National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov/) and from a search in the protein-family database Pfam (http://pfam.xfam.org/; (EL-GEBALI *et al.* 2019)) using hmmscan via the HMM webserver v.2.41.2 employing HMM release 3.3.2; http://hmmer.org/ (EDDY 2011)) of EMBL-EBI (https://www.ebi.ac.uk/Tools/hmmer/search/hmmscan) (POTTER *et al.* 2018).

Ribosomal RNA genes and gene fragments included in the genome sequence were searched with BLAST v2.2.26 (ALTSCHUL *et al.* 1990).

The identification of **transposable-element** types described in the main text was performed by extracting distinctive keywords from the automated function annotation. More specifically, we parsed the 'product' field (populated via function transfer from matches with UniProt/SwissProt proteins) and the 'note' field (containing information derived from matches with unreviewed domains) of the gff3 file. The search terms included 'gag protein', 'ingi', 'LINE', 'L1Tc', 'mobile element', 'polyprotein', 'reverse transcriptase', 'retrotranspos', 'retrovirus', 'ribonuclease HI', 'TcC31', 'SLACS', 'transcriptase'. The retrieved annotations were human-inspected.

To validate the automated function annotation '(putative) **trans-sialidase'**, initially assigned to 330 genome-inferred proteins, we performed a local HMM search with phmmer against the collection of reviewed trans-sialidases in Swissprot and further by HMMscan against all PFAM protein domains using web service at https://www.ebi.ac.uk/Tools/hmmer/. We realized that the function assignment of ~95% 'trans-sialidase' proteins was erroneous due to annotation transfer from incorrectly annotated proteins in GenBank.

NUMT identification

The 82 mitochondrial gene modules identified previously (MOREIRA *et al.* 2016) were searched against the nuclear genome assembly v1.0 using the Discontiguous MegaBLAST tool of the BLAST+ v2.2.28 package (CAMACHO *et al.* 2009). The scoring scheme was optimized for ~90% identity searches (i.e., match|mismatch|open gap|extend gap = 2|−3|5|2) (STATES AND BOTSTEIN 1991), with an initial E-value threshold of 10e-5 and a size threshold of 34 bp (i.e., longer than the 3× search-word size of 11 bp). Matches with contigs lacking gene annotations were ignored as these contigs might originate from incomplete removal of mitochondrial contigs from the nuclear genome assembly. To extend the matches to mitochondrial gene modules, we extracted from the contig the sequence interval that spans 8 kbp upstream and downstream of the match, 8 kbp being the maximum length of a mitochondrial chromosome. This nuclear genomic interval was then used as a query sequence to search by BLAST against a comprehensive mitochondrial sequence database. The target database consisted of all complete mitochondrial chromosome sequences (three and two from A and B class, respectively), all 81 cassettes, and all 18 mature transcripts (KIETHEGA *et al.* 2013; MOREIRA *et al.* 2016; VALACH *et al.* 2018). To verify whether a nuclear genomic region that included a NUMT was correctly assembled, we examined whether multiple distinct

long PacBio reads covered the NUMT and connected it to unique and reliable nuclear sequences. PacBio reads were aligned with the nuclear genome sequence using Minimap2 v2.17 (LI 2018), and the resulting alignments were visually inspected. We considered only reads of a minimum length of 5 kbp; support by one or two such reads was counted as weak evidence, and support by five or more reads counted as strong evidence for the candidate being located in the nuclear genome. All other pairwise and multiple sequence alignments were performed using MAFFT v7.388 (KATOH AND STANDLEY 2013). For searching cr-NUMTs derived from constant regions of mitochondrial chromosomes, we used the corresponding sequences of the complete chromosomes A46 (encoding *nad7*-m6 and *nad9*-m3; GenBank Acc. nr. HQ288824) and B03 (*cox1*-m4; EU123537) as representatives of the A and B class, respectively. The same parameters of BLAST+ and filtering were used as for the module-anchored searches described above. Overlapping hits were merged. For example, when A- and B-class mitochondrial chromosomes share a highly similar sequence, they will match a similar but not necessarily identical nuclear region; in such cases, the entire matched region was considered as a single NUMT.

AUTHOR CONTRIBUTIONS

Conceptualization, Data curation, Formal analysis, Investigation, Writing, original draft – G.B., M.V.; **Writing, review & editing** – all co-authors.

REFERENCES

AGI, 2000 Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408: 796-815.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman, 1990 Basic local alignment search tool. J Mol Biol 215: 403-410.

Aslett, M., C. Aurrecoechea, M. Berriman, J. Brestelli, B. P. Brunk *et al.*, 2010 TriTrypDB: a functional genomic resource for the Trypanosomatidae. Nucleic Acids Res 38: D457-462.

Benson, G., 1999 Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27: 573-580.

Calabrese, F. M., D. L. Balacco, R. Preste, M. A. Diroma, R. Forino *et al.*, 2017 NumtS colonization in mammalian genomes. Sci Rep 7: 16357.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos *et al.*, 2009 BLAST+: architecture and applications. BMC Bioinformatics 10: 421.

Eddy, S. R., 2011 Accelerated profile HMM searches. PLoS Comput Biol 7: e1002195.

El-Gebali, S., J. Mistry, A. Bateman, S. R. Eddy, A. Luciani *et al.*, 2019 The Pfam protein families database in 2019. Nucleic Acids Res 47: D427-d432.

Hazkani-Covo, E., R. M. Zeller and W. Martin, 2010 Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. PLoS Genet 6: e1000834.

Katoh, K., and D. M. Standley, 2013 MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30: 772-780.

Kaur, B., K. Záhonová, M. Valach, D. Faktorová, G. Prokopchuk *et al.*, 2020 Gene fragmentation and RNA editing without borders: eccentric mitochondrial genomes of diplonemids. Nucleic Acids Res 48: 2694-2708.

Kiethega, G. N., M. Turcotte and G. Burger, 2011 Evolutionary conserved *cox1* trans-splicing without cis-motifs. Mol Biol Evol 28: 2425-2458.

Kiethega, G. N., Y. Yan, M. Turcotte and G. Burger, 2013 RNA-level unscrambling of fragmented genes in *Diplonema* mitochondria. RNA Biology 10: 301-313.

Kleine, T., U. G. Maier and D. Leister, 2009 DNA transfer from organelles to the nucleus: the idiosyncratic genetics of endosymbiosis. Annu Rev Plant Biol 60: 115-138.

Ko, Y. J., and S. Kim, 2016 Analysis of nuclear mitochondrial DNA segments of nine plant species: size, distribution, and insertion loci. Genomics Inform 14: 90-95.

Li, H., 2018 Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34: 3094-3100.

Mehrotra, S., and V. Goyal, 2014 Repetitive sequences in plant nuclear DNA: types, distribution, evolution and function. Genomics Proteomics Bioinformatics 12: 164-171.

Melters, D. P., K. R. Bradnam, H. A. Young, N. Telis, M. R. May *et al.*, 2013 Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biology 14: R10.

Michalovova, M., B. Vyskot and E. Kejnovsky, 2013 Analysis of plastid and mitochondrial DNA insertions in the nucleus (NUPTs and NUMTs) of six plant species: size, relative age and chromosomal localization. Heredity (Edinb) 111: 314-320.

Moreira, S., M. Valach, M. Aoulad-Aissa, C. Otto and G. Burger, 2016 Novel modes of RNA editing in mitochondria. Nucleic Acids Res 44: 4907-4919.
Noutsos, C., T. Kleine, U. Armbruster, G. DalCorso and D. Leister, 2007 Nuclear insertions of organellar DNA can create novel patches of functional exon sequences. Trends Genet 23: 597-601.
Ohno, S., 1972 So much "junk" DNA in our genome. Brookhaven Symp Biol 23: 366-370.
Orgel, L. E., F. H. Crick and C. Sapienza, 1980 Selfish DNA. Nature 288: 645-646.
Pearson, W. R., 2016 Finding protein and nucleotide similarities with FASTA. Curr Protoc Bioinformatics 53: 3.9.1-3.9.25.
Potter, S. C., A. Luciani, S. R. Eddy, Y. Park, R. Lopez *et al.*, 2018 HMMER web server: 2018 update. Nucleic Acids Res 46: W200-w204.
Price, A. L., N. C. Jones and P. A. Pevzner, 2005 De novo identification of repeat families in large genomes. Bioinformatics 21 Suppl 1: i351-358.
Pucker, B., D. Holtgräwe, K. B. Stadermann, K. Frey, B. Huettel *et al.*, 2019 A chromosome-level sequence assembly reveals the structure of the *Arabidopsis thaliana* Nd-1 genome and its gene set. PLoS One 14: e0216233.
Quinlan, A. R., 2014 BEDTools: the swiss-army tool for genome feature analysis. Curr Protoc Bioinformatics 47: 11.12.11-34.
Ramakrishnan, S., and V. Gupta, 2021 Trinucleotide repeat disorders in *StatPearls*. StatPearls Publishing Copyright © 2021, StatPearls Publishing LLC., Treasure Island (FL).
Richard, G. F., A. Kerrest and B. Dujon, 2008 Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol Mol Biol Rev 72: 686-727.
Rogers, H. H., and S. Griffiths-Jones, 2012 Mitochondrial pseudogenes in the nuclear genomes of *Drosophila*. PLoS One 7: e32593.
Smit, A. F. A., and P. Green, RepeatMasker, pp.
States, D. J., and D. Botstein, 1991 Molecular sequence accuracy and the analysis of protein coding regions. Proc Natl Acad Sci U S A 88: 5518-5522.
Stupar, R. M., J. W. Lilly, C. D. Town, Z. Cheng, S. Kaul *et al.*, 2001 Complex mtDNA constitutes an approximate 620-kb insertion on *Arabidopsis thaliana* chromosome 2: implication of potential sequencing errors caused by large-unit repeats. Proc Natl Acad Sci U S A 98: 5099-5103.
Valach, M., A. Léveillé-Kunst, M. W. Gray and G. Burger, 2018 Respiratory chain Complex I of unparalleled divergence in diplonemids. J Biol Chem 293: 16043-16056.
Valach, M., D. Moreira, D. Faktorová, J. Lukeš and G. Burger, 2016 Post-transcriptional mending of gene sequences: looking under the hood of mitochondrial gene expression in diplonemids. RNA Biol 13: 1204-1211.
Valach, M., S. Moreira, S. Hoffmann, P. F. Stadler and G. Burger, 2017 Keeping it complicated: mitochondrial genome plasticity across diplonemids. Sci Rep 7: 14166.
Wang, D., Z. Qu, D. L. Adelson, J. K. Zhu and J. N. Timmis, 2014 Transcription of nuclear organellar DNA in a model plant system. Genome Biol Evol 6: 1327-1334.
Wood, V., K. M. Rutherford, A. Ivens, M. A. Rajandream and B. Barrell, 2001 A re-annotation of the *Saccharomyces cerevisiae* genome. Comp Funct Genomics 2: 143-154.

7. Polycistronic transcription units in the nuclear genome of *Diplonema papillatum*

INTRODUCTION

In most eukaryotes, each nuclear protein-coding gene has its own promoter, with the consequence that neighboring genes may or may not be located on the same DNA strand. In contrast, all nuclear genomes of kinetoplastids analyzed to date have gene clusters of up to 100 kbp in length, grouping up to a hundred members arranged in the same orientation (EL-SAYED *et al.* 2005; JACKSON *et al.* 2016). Genes within these clusters are frequently co-transcribed, with the corresponding gene arrays referred to as Polycistronic Transcription Units (PTUs; (MARTÍNEZ-CALVILLO *et al.* 2003)). In *Trypanosoma*, promoters were pinpointed systematically upstream of PTUs, but some promoters were also found within PTU arrays (KOLEV *et al.* 2010). Polycistronic primary transcripts are processed by a specialized spliceosome via *trans*-splicing that involves the attachment of a short (39 nt-long in *Trypanosoma*) spliced-leader (SL) RNA —decorated by a hyper-modified cap— to the 5′ end of each individual mRNA (reviewed in (MICHAELI 2011)). Messenger RNA maturation is finalized by cleaving off the portion downstream of the transcript's 3′ UTR followed by the addition of a poly(A) tail (reviewed in (CLAYTON AND MICHAELI 2011)).

RESULTS AND DISCUSSION

Inspection of the annotated *Diplonema* genome assembly showed that many predicted protein-coding genes in a given contig were arranged head-to-tail, suggesting co-transcription. Therefore, we examined in more detail clusters of five or more protein-coding genes (referred to here as 'unidirectional gene arrays') occurring in contigs larger than 50 kbp.

About 87% (1,159) of these contigs carried at least one unidirectional gene array, with the median size of approximately nine genes in a row and with the longest array having >120 members (for details, see **Additional File 3** polycistTranscripts). The longer contigs, in particular, contained several shorter arrays in a row having the same orientation, most often separated by a single gene model on the other strand. Closer scrutiny of several dozen such cases revealed that virtually all 'trend-disrupting' genes lacked evidence for transcription. Further, most of them were related to open reading frames contained in transposons or dispersed repeat elements that are uniquely found in the *D. papillatum* genome sequence (see also the **Supplementary Information**: Section 2. Assembly and annotation of the nuclear genome and transcriptome of *Diplonema papillatum*, in particular, the section on the expert validation of gene models).

Based on the three longest contigs whose gene models we manually curated, we estimated the extent of polycistronic transcription units in the genome of *D. papillatum*. For example, of the initial 11 strand switches in contig tig00022679, only a single one remained after curation, leading to arrays comprising up to 52 genes. By extrapolation, the *Diplonema* genome is believed to contain two to three times longer unidirectional gene arrays compared to the prediction by the automated structural annotation. Among the expert-confirmed arrays, the longest extends to more than 1 Mbp (tig00022654_12).

In *Trypanosoma*, arrays of about a hundred unidirectional genes are transcribed into several polycistronic RNAs (KOLEV *et al.* 2010). Therefore, it is possible that the gene arrays in *Diplonema*, too, contain multiple promoters. These are likely located in intergenic regions that stand out for their large size. For example, contig tig00022654_12 includes six intergenic regions that are >10 kbp long. Given these predictions, it will be interesting to examine transcription initiation and termination positions of *Diplonema* nuclear genes by experimental means. For example, it was shown in kinetoplastids that modified bases provide the signal for transcription termination (VAN LUENEN *et al.* 2012; SCHULZ *et al.* 2016).

In conclusion, *D. papillatum* and kinetoplastids appear to share an organization of nuclear genes into sizeable PTUs, and a relatively small number of RNAP II transcription start sites per chromosome. While the organizational principle may be the same in the two sister taxa, the actual gene order is very different, since we did not detect any signs of synteny between the genomes of *D. papillatum* on the one hand and *Trypanosoma brucei*, *Leishmania tarentolae* or *Bodo saltans* on the other.

MATERIALS AND METHODS

The orientation of protein-coding genes along a contig was extracted from the genome annotation file (gff). The identification of spliced-leader trans-splicing sites is described in the **Supplementary Information** : Section 2. Assembly and annotation of the nuclear genome and transcriptome of *Diplonema papillatum*.

AUTHOR CONTRIBUTIONS

Conceptualization, Investigation – G.B., M.V.; **Formal analysis, Writing, original draft** – M.V.; **Writing, review & editing** – all co-authors.

REFERENCES

Clayton, C., and S. Michaeli, 2011 3' processing in protists. Wiley Interdiscip Rev RNA 2**:** 247-255.

El-Sayed, N. M., P. J. Myler, G. Blandin, M. Berriman, J. Crabtree *et al.*, 2005 Comparative genomics of trypanosomatid parasitic protozoa. Science 309**:** 404-409.

Jackson, A. P., T. D. Otto, M. Aslett, S. D. Armstrong, F. Bringaud *et al.*, 2016 Kinetoplastid phylogenomics reveals the evolutionary innovations associated with the origins of parasitism. Current Biology : CB 26**:** 161-172.

Kolev, N. G., J. B. Franklin, S. Carmi, H. Shi, S. Michaeli *et al.*, 2010 The transcriptome of the human pathogen *Trypanosoma brucei* at single-nucleotide resolution. PLoS Pathog 6**:** e1001090.

Martínez-Calvillo, S., S. Yan, D. Nguyen, M. Fox, K. Stuart *et al.*, 2003 Transcription of *Leishmania major* Friedlin chromosome 1 initiates in both directions within a single region. Mol Cell 11**:** 1291-1299.

Michaeli, S., 2011 Trans-splicing in trypanosomes: machinery and its impact on the parasite transcriptome. Future Microbiol 6**:** 459-474.

Schulz, D., M. Zaringhalam, F. N. Papavasiliou and H. S. Kim, 2016 Base J and H3.V regulate transcriptional termination in *Trypanosoma brucei*. PLoS Genet 12**:** e1005762.

van Luenen, H. G., C. Farris, S. Jan, P. A. Genest, P. Tripathi *et al.*, 2012 Glucosylated hydroxymethyluracil, DNA base J, prevents transcriptional readthrough in *Leishmania*. Cell 150**:** 909-921.

8. DNA modifications (5mC and J)

INTRODUCTION

Genomic DNA consists primarily of four bases, adenine (A), guanine (G), cytosine (C), and thymine (T). However, derivatives of this basic alphabet have been documented in numerous organisms, with around 50 naturally occurring modified bases identified so far (SOOD *et al.* 2019; DAI *et al.* 2021). Out of these, the two most frequently encountered modified bases are 5-methyl-cytosine (5mC) and N6-methyl-adenine (6mA) in eukaryotic and prokaryotic genomes, respectively, which serve as epigenetic marks (SUN *et al.* 2015; DE MENDOZA *et al.* 2019).

Most unconventional bases occur across the tree of life, but some display a constrained phylogenetic distribution. A prime example of the latter category is the hyper-modified thymine derivative, β-D-glucopyranosyl-oxymethyluracil, referred to as base J. This nucleotide is restricted to euglenozoans ((GOMMERS-AMPT *et al.* 1993; VAN LEEUWEN *et al.* 1998; DOOIJES *et al.* 2000), reviewed in (BORST AND SABATINI 2008)). The metabolism of base J has been extensively studied in trypanosomes and leishmanias, in which its biosynthesis relies on three protein families: JBP1/2, JGT, and JBP3. JBP1 and JBP2 are DNA-binding thymidine hydroxylases (TH). JBP2 generates *de novo* the modification of thymine to hydroxymethyl-uracil (5hmU), and JBP1 then propagates this base across a larger region (CROSS *et al.* 1999; DIPAOLO *et al.* 2005). Subsequently, the intermediate 5hmU is transformed by the glycosyltransferase JGT to base J (BULLARD *et al.* 2014). The most recently identified J-interacting protein JBP3 binds to DNA enriched in J and recruits additional factors to the modified sites forming a regulatory protein complex (KIEFT *et al.* 2020).

RESULTS

In *Diplonema* nuclear DNA, modified bases were experimentally identified more than two decades ago, notably the taxonomically wide-spread 5mC and the euglenozoan-specific base J (VAN LEEUWEN *et al.* 1998). We searched in the *D. papillatum* genome assembly for homologs of proteins known to be implicated in the metabolism of these two minor DNA bases.

The nuclear genome of *Diplonema* codes for five distinct proteins of the cytosine-5 DNA methyltransferase (DNMT) family, a class of enzymes responsible for the biosynthesis of 5mC (DIPPA_02667, DIPPA_11195, DIPPA_01326, DIPPA_70042, and DIPPA_70054). The latter three belong to the DNMT2 subfamily and most likely act as tRNA methyltransferases. Interestingly, DIPPA_02667 and DIPPA_11195, which are members of the more prominent DNMT3/5 family, affiliate in phylogenetic analyses with *de novo* DNA methyltransferases from bacteria that are also sporadically found in protists and certain fungi (BEWICK *et al.* 2019). This distribution suggests that *D. papillatum* acquired these genes *via* horizontal transfer. We also detected homologs of a dozen oxidative demethylases belonging to the AlkB family. Some of these predicted enzymes might be specifically responsible for the C5-demethylation of nuclear DNA.

We also searched in the inferred *Diplonema* proteome for homologs of the three protein families involved in the base J metabolism from kinetoplastids. *D. papillatum* (and all other examined diplonemids and *Euglena*) seems to lack JBP2, but encodes two paralogs of JBP1, namely DIPPA_70009 (DpJBP1B) and DIPPA_70011 (DpJBP1A). The latter possesses in addition to TH and J-DNA-binding domains, an atypical C-terminal methyltransferase domain (PF13489). The top candidate for being a JGT homolog is DIPPA_30303, which encodes a glycosyltransferase.

Lastly, we identified two JBP3 homologs in *Diplonema* (as well as in other diplonemids), which we refer to as JBP3A (DIPPA_17973) and JBP3B (DIPPA_28081). The latter protein is more closely related to the sole JBP3 in kinetoplastids. *Leishmania* JGT and JBP3 were shown to form a complex together with the proteins PP1, Wdr82/Swd2, and PNUTS (KIEFT *et al.* 2020). However, among dozens of potential homologs, *Diplonema* has no clear orthologs of PP1 and Wdr82, and further appears to lack PNUTS. The base J metabolism either has been rewired, or its components have considerably diverged during euglenozoan evolution.

DISCUSSION

In eukaryotes, C5 methylation plays various roles ranging from gene and transposon silencing to nucleosome positioning (HUFF AND ZILBERMAN 2014; SCHMITZ *et al.* 2019). Given the abundance of non-transcribed repetitive and transposable elements in the *D. papillatum* genome, we posit that 5mC is involved in their transcriptional repression.

Base J, a hallmark feature of euglenozoans, often has a predominantly telomeric localization but is also found elsewhere in the nuclear genome (GENEST *et al.* 2015). While this base functions as a critical genomic silencing marker, its precise role varies. In *Leishmania*, the rare extra-telomeric J bases participate in transcriptional termination of polycistronic units (VAN LUENEN *et al.* 2012), while in trypanosomes, the role of base J is auxiliary to a histone H3 variant (SCHULZ *et al.*

2016). Given the more than 100 kbp-long polycistron-like gene clusters in *Diplonema* (see main text), we predict that in this protist, base J has the same primary function as in kinetoplastids. At the same time, *D. papillatum* differs in two key aspects from kinetoplastids: the richness of silent repetitive DNA and transposable elements in the nuclear genome and the assortment of proteins involved in the base J metabolism. Hence, investigating the diplonemid system promises to provide new insights into the adaptability of this epigenetic pathway.

METHODS

The identification of protein function was performed as described in the **Supplementary Information**: Section 2. Assembly and annotation of the nuclear genome and transcriptome of *Diplonema papillatum*.

AUTHOR CONTRIBUTIONS

Conceptualization, Investigation, Formal analysis, Writing, original draft – M.V.; **Writing, review & editing** – all co-authors.

REFERENCES

Bewick, A. J., B. T. Hofmeister, R. A. Powers, S. J. Mondo, I. V. Grigoriev *et al.*, 2019 Diversity of cytosine methylation across the fungal tree of life. Nat Ecol Evol 3**:** 479-490.

Borst, P., and R. Sabatini, 2008 Base J: discovery, biosynthesis, and possible functions. Annu Rev Microbiol 62**:** 235-251.

Bullard, W., J. Lopes da Rosa-Spiegler, S. Liu, Y. Wang and R. Sabatini, 2014 Identification of the glucosyltransferase that converts hydroxymethyluracil to base J in the trypanosomatid genome. J Biol Chem 289**:** 20273-20282.

Cross, M., R. Kieft, R. Sabatini, M. Wilm, M. de Kort *et al.*, 1999 The modified base J is the target for a novel DNA-binding protein in kinetoplastid protozoans. Embo j 18**:** 6573-6581.

Dai, Y., B. F. Yuan and Y. Q. Feng, 2021 Quantification and mapping of DNA modifications. RSC Chem Biol 2**:** 1096-1114.

de Mendoza, A., R. Lister and O. Bogdanovic, 2019 Evolution of DNA methylome diversity in eukaryotes. J Mol Biol.

DiPaolo, C., R. Kieft, M. Cross and R. Sabatini, 2005 Regulation of trypanosome DNA glycosylation by a SWI2/SNF2-like protein. Mol Cell 17**:** 441-451.

Dooijes, D., I. Chaves, R. Kieft, A. Dirks-Mulder, W. Martin *et al.*, 2000 Base J originally found in kinetoplastida is also a minor constituent of nuclear DNA of *Euglena gracilis*. Nucleic Acids Res 28**:** 3017-3021.

Genest, P. A., L. Baugh, A. Taipale, W. Zhao, S. Jan *et al.*, 2015 Defining the sequence requirements for the positioning of base J in DNA using SMRT sequencing. Nucleic Acids Res 43**:** 2102-2115.

Gommers-Ampt, J. H., A. J. Teixeira, G. van de Werken, W. J. van Dijk and P. Borst, 1993 The identification of hydroxymethyluracil in DNA of *Trypanosoma brucei*. Nucleic Acids Res 21**:** 2039-2043.

Huff, J. T., and D. Zilberman, 2014 Dnmt1-independent CG methylation contributes to nucleosome positioning in diverse eukaryotes. Cell 156**:** 1286-1297.

Kieft, R., Y. Zhang, A. P. Marand, J. D. Moran, R. Bridger *et al.*, 2020 Identification of a novel base J binding protein complex involved in RNA polymerase II transcription termination in trypanosomes. PLoS Genet 16**:** e1008390.

Schmitz, R. J., Z. A. Lewis and M. G. Goll, 2019 DNA methylation: shared and divergent features across eukaryotes. Trends Genet 35**:** 818-827.

Schulz, D., M. Zaringhalam, F. N. Papavasiliou and H. S. Kim, 2016 Base J and H3.V regulate transcriptional termination in *Trypanosoma brucei*. PLoS Genet 12**:** e1005762.

Sood, A. J., C. Viner and M. M. Hoffman, 2019 DNAmod: the DNA modification database. J Cheminform 11**:** 30.

Sun, Q., S. Huang, X. Wang, Y. Zhu, Z. Chen *et al.*, 2015 N6-methyladenine functions as a potential epigenetic mark in eukaryotes. Bioessays 37**:** 1155-1162.

van Leeuwen, F., M. C. Taylor, A. Mondragon, H. Moreau, W. Gibson *et al.*, 1998 beta-D-glucosyl-hydroxymethyluracil is a conserved DNA modification in kinetoplastid protozoans and is abundant in their telomeres. Proc Natl Acad Sci U S A 95**:** 2366-2371.

van Luenen, H. G., C. Farris, S. Jan, P. A. Genest, P. Tripathi *et al.*, 2012 Glucosylated hydroxymethyluracil, DNA base J, prevents transcriptional readthrough in *Leishmania*. Cell 150**:** 909-921.

9. RNA interference (RNAi)

INTRODUCTION

RNA interference (RNAi) is a process thought to have evolved as a defence mechanism against retrotransposons and invading double-stranded (ds) RNA viruses (MALONE AND HANNON 2009). The last eukaryotic common ancestor (LECA) must have been already equipped with a minimal eukaryotic RNA silencing pathway comprising an RNase III-like endonuclease Dicer, PIWI-domain containing proteins (Argonaute [Ago] and Piwi), and an RNA-dependent RNA polymerase (RdRP) (BURROUGHS *et al.* 2014; SWARTS *et al.* 2014). Mechanistically, Dicer processes long dsRNAs into small interfering (si) RNAs that are then loaded onto PIWI-domain proteins. The siRNAs recognize target RNA molecules through complementary base pairing. Subsequently, the target transcript is cleaved, thereby silencing the expression of the corresponding gene. RdRP can either initiate or enhance the response by amplifying the amount of the dsRNA that triggers the cascade.

RESULTS

The *Diplonema* inferred proteome contains all the elementary components of a functional RNAi pathway (**Supplementary Table S18**). The genome assembly includes two homologs of the Dicer-like protein (DCL), notably DpDCL1 (DIPPA_01275) and DpDCL2 (DIPPA_00339). DCL1 is an unusually large gene spanning three contigs; the protein sequence was manually assembled with the help of transcriptome data. Other diplonemids appear to contain only a single Dicer—one that most closely resembles DpDCL1. DpDCL2 likely originated in *D. papillatum* by gene duplication and divergence, as this protein lacks a C-terminal domain that all other diplonemid DCLs otherwise share.

A total of eight PIWI domain-containing proteins were identified in the *D. papillatum* inferred proteome. The five genes coding for Piwi proteins (DIPPA_27834, DIPPA_27835, DIPPA_27840, DIPPA_27841, and DIPPA_70076) are identical in sequence and arranged in a tandem array; the correctness of the assembly at the corresponding locus was verified by examining PacBio reads. In contrast, three members of the Argonaute family (DIPPA_09821, DIPPA_23819, and DIPPA_13995) are distinct in sequence (pairwise identity 93%) and are encoded in different genomic loci. Other diplonemids seem to possess a similar assortment of homologs.

The *Diplonema* genome also encodes a single homolog of RdRP (DIPPA_03987). The protein is characterized by three distinct regions: an N-terminal helicase domain (Pfam families DEAD (PF00270) and Helicase_C (PF00271)), a central RdRP domain (Pfam family RdRP (PF05183)), and a C-terminal low-complexity region. While kinetoplastids are known to lack RdRP (KOLEV *et al.* 2011), other discobids, such as the heterolobosean *Naegleria gruberi*, have undergone an expansion of this family (BURROUGHS *et al.* 2014).

Lastly, we searched for homologs of two novel kinetoplastid-specific RNAi factors discovered in *Trypanosoma brucei*, i.e., TbRIF4, which is responsible for generating single-stranded siRNAs from duplexes, and TbRIF5, which is a cofactor of *Trypanosoma* DCL1 (BARNES *et al.* 2012). In *Diplonema* (and in several other diplonemids), we found a single homolog of TbRIF4 (DIPPA_70075, another large gene spanning three contigs in the assembly). Still, a counterpart of the Dicer cofactor was not detected.

Supplementary Table S18. Proteins involved in RNAi of *D. papillatum*.

Dicer	Argonaute		RdRP	Co-factors
	Piwis	Agos		
DIPPA_01275 (DCL1)	DIPPA_27834	DIPPA_09821	DIPPA_03987	DIPPA_70075 (RIF4)
DIPPA_00339 (DCL2)	DIPPA_27835	DIPPA_23819		
	DIPPA_27840	DIPPA_13995		
	DIPPA_27841			
	DIPPA_70076			

DISCUSSION

The predicted RNAi-implicated protein repertoire in *D. papillatum*, and diplonemids in general, indicates a less streamlined pathway compared to kinetoplastids. In particular the finding of a signal-amplifying RdRP gene in the *Diplonema* nuclear genome suggests a more conventional RNAi machinery.

In addition, all examined diplonemids possess multiple Piwi and Ago proteins, which contrasts with kinetoplastids. For instance, in *T. brucei*, its sole AGO1 is responsible for all forms of mRNA silencing, irrespective of whether siRNA duplexes are generated by the cytoplasmic Dicer-like (DCL) enzyme 1 or the nucleus-located DCL2 (KOLEV *et al.* 2011). The function of the PIWI-only TbPWI1 is unclear, but the protein is apparently not involved in RNAi (DURAND-DUBIEF AND BASTIN 2003). Curiously, in phylogenetic trees, kinetoplastid Piwi (Kiwi) and Ago proteins cluster together to the exclusion of Argonautes and Piwis from all other eukaryotes (SWARTS *et al.* 2014). Based on our phylogenetic analysis, all diplonemid Ago proteins belong to the same family, including their kinetoplastid counterparts (**Supplementary Figure S16,** 'euAgo' group, for euglenozoan Ago). In contrast, the Piwi proteins of diplonemids associate with the RNAi-implicated members of the Piwi-like family from other eukaryotes. This indicates that not only Agos, but also Piwis likely contribute to silencing responses in diplonemids.

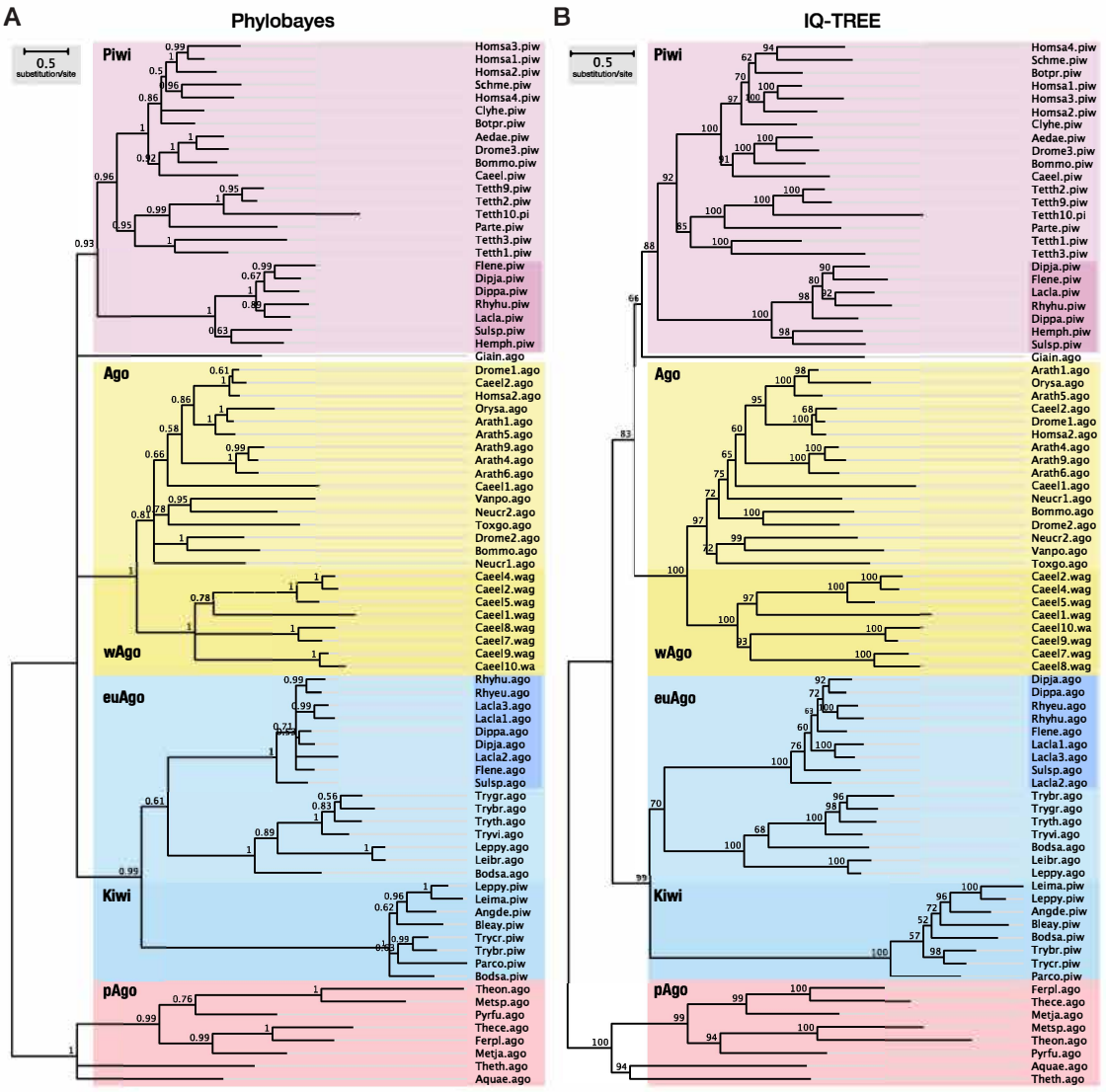
Traditionally, RNAi has been considered to be involved in post-transcriptional gene regulation and transposon repression. Both processes are critical for *Diplonema* given polycistronic transcription and the large population of transposable elements lingering in its nuclear genome. However, while the *Diplonema* RNAi machinery might well suppress transcripts originating from mobile elements, it has apparently not contained their massive spread throughout the nuclear genome. Alternatively, RNAi in *Diplonema* could be predominantly involved in other processes such as small-RNA-driven translational silencing, genome methylation, or centromere formation, functions that have been documented in various organisms, including animals, plants, fungi and certain protists alike (KETTING 2011; GUTBROD AND MARTIENSSEN 2020). Determining more precisely what roles the individual RNAi pathway components perform in *Diplonema* will require experimental scrutiny.

METHODS

Protein identification and function assignment across diplonemids were performed by searches with BLAST and profile Hidden Markov Models (HMMs) as described in the Supplementary Information on functional genome annotation. To determine the phylogenetic relationships among Piwi and Ago proteins, sequences from diplonemids and selected eukaryotes and prokaryotes from a previous study (SWARTS *et al.* 2014) were pre-aligned with Muscle (EDGAR 2004) and realigned with an HMM search (built with the initial Muscle alignment) using hmmsearch (HMMER 3.3, Nov 2019) (EDDY 2011). Only amino acids aligned with a PP value of 1.0 were retained for phylogenetic analysis, which was performed by a Bayesian and maximum likelihood (ML) approach. For the former, we used PhyloBayes (LARTILLOT *et al.* 2013) by running four independent chains, six gamma categories and the CAT-GTR model. For the ML approach, we used IQ-TREE with default parameters and the option to calculate 1,000 ultrafast bootstrap replicates (MINH *et al.* 2020).

AUTHOR CONTRIBUTIONS

Conceptualization, Data curation, Investigation, Writing, original draft – C.B., M.V.; **Formal analysis, Visualization** – B.F.L., M.V.; **Writing, review & editing** – all co-authors.



Supplementary Figure S16. Phylogeny of Piwi-domain containing proteins. The phylogenetic tree was constructed with Bayesian (**A**) and maximum likelihood (**B**) methods using the Piwi-domain sequences of proteins across eukaryotes and from several prokaryotes as outgroups (pAgos). Posterior probabilities (Phylobayes) and UF bootstrap support values (IQ-TREE) are indicated next to each branch. Both approaches resolve the tree into several clusters of related sequences, namely: conventional Piwi proteins from various eukaryotes (Piwi), conventional Ago proteins from various eukaryotes (Ago), nematode-specific Ago protein expansion (wAgo), kinetoplastid and diplonemid Ago proteins (euAgo), kinetoplastid-specific Piwi proteins (Kiwi), and prokaryotic Ago proteins (pAgo). Diplonemid taxa are highlighted in darker background shades.

REFERENCES

Barnes, R. L., H. Shi, N. G. Kolev, C. Tschudi and E. Ullu, 2012 Comparative genomics reveals two novel RNAi factors in *Trypanosoma brucei* and provides insight into the core machinery. PLoS Pathog 8: e1002678.

Burroughs, A. M., Y. Ando and L. Aravind, 2014 New perspectives on the diversification of the RNA interference system: insights from comparative genomics and small RNA sequencing. Wiley Interdiscip Rev RNA 5: 141-181.

Durand-Dubief, M., and P. Bastin, 2003 TbAGO1, an argonaute protein required for RNA interference, is involved in mitosis and chromosome segregation in *Trypanosoma brucei*. BMC Biol 1: 2.

Eddy, S. R., 2011 Accelerated profile HMM searches. PLoS Comput Biol 7: e1002195.

Edgar, R. C., 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792-1797.

Gutbrod, M. J., and R. A. Martienssen, 2020 Conserved chromosomal functions of RNA interference. Nat Rev Genet 21: 311-331.

Ketting, R. F., 2011 The many faces of RNAi. Dev Cell 20: 148-161.

Kolev, N. G., C. Tschudi and E. Ullu, 2011 RNA interference in protozoan parasites: achievements and challenges. Eukaryot Cell 10: 1156-1163.

Lartillot, N., N. Rodrigue, D. Stubbs and J. Richer, 2013 PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. Syst Biol 62: 611-615.

Malone, C. D., and G. J. Hannon, 2009 Small RNAs as guardians of the genome. Cell 136: 656-668.

Minh, B. Q., H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams *et al.*, 2020 IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol 37: 1530-1534.

Swarts, D. C., K. Makarova, Y. Wang, K. Nakanishi, R. F. Ketting *et al.*, 2014 The evolutionary journey of Argonaute proteins. Nat Struct Mol Biol 21: 743-753.

10. The cytosolic ribosome of *Diplonema papillatum*

Gel electrophoresis under denaturing conditions demonstrated that the *Diplonema* cytosolic ribosome (cytoribosome) contains the usual four eukaryotic rRNA species: 18S, 5.8S, 26S, and 5S. The bipartite *Diplonema* large subunit (LSU) rRNA (5.8S + 26S) stands in marked contrast to the cytosolic LSU rRNAs of other euglenozoans, where the 26S rRNA component is further fragmented into six distinct pieces in kinetoplastids (GRAY 1981; CAMPBELL *et al.* 1987; SPENCER *et al.* 1987; HASHEM *et al.* 2013; LIU *et al.* 2016) and 13 separate pieces in the euglenid *Euglena gracilis* (SCHNARE *et al.* 1990; SCHNARE AND GRAY 1990; HAŁAKUC *et al.* 2022). Although some of the additional processing sites are in a similar site in kinetoplastid and euglenid LSU rRNAs (SCHNARE *et al.* 1990), these positionally equivalent processing events must have arisen independently in the two lineages, given that diplonemids are evolutionarily intermediate between kinetoplastids and euglenids (HAŁAKUC *et al.* 2022).

The *Diplonema* genome encodes all 33 small subunit (SSU) ribosomal proteins (RPs) found in the human cytoribosome. In addition, the *Diplonema* SSU is predicted to contain an ortholog of a novel RNA-binding protein that has been identified by cryo-EM in the SSU of kinetoplastids (BRITO QUERIDO *et al.* 2017) and *E. gracilis* (MATZOV *et al.* 2020). We also identified all 47 LSU RPs found in the human cytoribosome.

To gain insight into the organization and evolution of cyt-RP genes in *Diplonema*, we carried out a detailed manual examination of SSU and LSU cyt-RP genes (of which there are in total 109 and 155, respectively), CDS and their predicted amino acid sequences. Several themes emerged from the study of identified cyt-RP genes.

- As in other eukaryotes (LECOMPTE *et al.* 2002), multiple copies exist for each of the *Diplonema* SSU and LSU cyt-RPs. Gene copies (2–9) may be localized in a single genomic scaffold (e.g., five copies of uS3 on tig00023122_3) or distributed among two or more scaffolds. When copies are located on the same scaffold, they are invariably found in the same transcriptional orientation and generally separated by several thousand base pairs.
- Just over half of the 109 SSU and 155 LSU cyt-RP gene copies (57 and 86, respectively) contain introns (1–3 per copy). With few exceptions, intron-containing copies have the same number of introns, and these are invariably located at the same position within the CDS. With very few exceptions (e.g., I3 in duplicate uS7 copies), positionally equivalent introns differ considerably in size and display no significant sequence similarity with one another. Intron size generally ranges from a few hundred to a few thousand bp, with an unusually long intron (23,385 bp) in one uS13 gene copy (vs. 416 bp in the other).
- In duplicate copies of the uS8 gene, a single intron is present in the 3′-UTR, inserted after position 14 in copy #1 and position 16 in copy #2. Comparison of genome and transcriptome sequence data provides evidence of alternative splicing of the copy #1 transcript, in which the 3′ UTR intron removed is either 233 or 301 nt long (both GT...AG).
- In comparing CDS sequences of different copies of the same gene, no consistent pattern of sequence conservation/divergence emerges. In some cases, CDSs are identical, in other cases quite divergent. For example, duplicate copies of the uS8 gene are present in two separate genomic scaffolds. The duplicates on each scaffold have an identical CDS, but the CDSs between duplicate pairs differ by 55 SNPs (86% identity) and predicted protein sequences differ at 9 positions (93% identity). In other cases, the CDSs in multiple copies differ by only a single or a few SNPs, all of which are at silent 1st or 3rd codon positions, whereas in other instances SNPs are relatively numerous (e.g., 21 SNPs and 95.5% identity between the duplicate uS13 copies). But again, all SNPs are silent so that the predicted amino acid sequence is unaffected.
- Yet another pattern is exemplified by the duplicate copies of uS2 (DIPPA_07808 and DIPPA_05951), whose CDSs differ by 162 SNPs (82% identity) and predicted protein sequences differ at 61 positions (75% identity). Nevertheless, despite their divergent sequences, BLASTp against NCBI nr with both protein variants retrieves the same top hit ('40S small subunit ribosomal protein uS2 [Euglena gracilis]').
- Another type of variation arises from terminal extensions or truncations. For instance, the two uS14 copies, located on different contigs, have a length of 150 aa (DIPPA_22991) and 100 aa (DIPPA_33894). Notably, all the genes discussed here are transcribed at an exceptionally high level.
- Overall, CDS identity between multiple copies of the same gene ranges from ~80% to 100% and inferred amino acid sequence between ~87% and 100%, although the latter is generally >95%. These results indicate the emergence of paralogs by gene duplication and divergence within the diplonemid lineages rather than by horizontal acquisition of a new homolog.

In summary, detailed examination of *Diplonema* cyt-RP genes did not reveal a consistent pattern that would allow one to draw meaningful conclusions about the mechanism of duplication and sequence divergence of these genes. The data do indicate the existence of sequence variants of a substantial number of individual *Diplonema* cyt-RPs. Although further work would be necessary to confirm transcription of individual cyt-RP gene variants that differ in sequence by only a few nucleotides, current evidence from transcript quantification indicates that essentially all copies of a particular cyt-RP are expressed in the form of mRNAs. It remains to be determined what the protein levels are, whether expression is constitutive, and if certain variants are preferred under certain environmental or physiological conditions. These observations leave open the possibility that *Diplonema* cytoribosomes may be heterogeneous with respect to their composition of individual RPs. Compositional ribosome heterogeneity has been a subject of considerable conjecture and experimentation (GILBERT 2011; SLAVOV *et al.* 2015; SHI *et al.* 2017; GENUTH AND BARNA 2018; FERRETTI AND KARBSTEIN 2019; GHULAM *et al.* 2020; MARTINEZ-SEIDEL *et al.* 2020; NORRIS *et al.* 2021).

AUTHOR CONTRIBUTIONS

Conceptualization, Data curation, Writing, original draft – M.W.G.; **Investigation** – M.W.G., M.V.; **Writing, review & editing** – all co-authors.

REFERENCES

Brito Querido, J., E. Mancera-Martínez, Q. Vicens, A. Bochler, J. Chicher *et al.*, 2017 The cryo-EM structure of a novel 40S kinetoplastid-specific ribosomal protein. Structure 25: 1785-1794.e1783.

Campbell, D. A., K. Kubo, C. G. Clark and J. C. Boothroyd, 1987 Precise identification of cleavage sites involved in the unusual processing of trypanosome ribosomal RNA. J Mol Biol 196: 113-124.

Ferretti, M. B., and K. Karbstein, 2019 Does functional specialization of ribosomes really exist? Rna 25: 521-538.

Genuth, N. R., and M. Barna, 2018 Heterogeneity and specialized functions of translation machinery: from genes to organisms. Nat Rev Genet 19: 431-452.

Ghulam, M. M., M. Catala and S. Abou Elela, 2020 Differential expression of duplicated ribosomal protein genes modifies ribosome composition in response to stress. Nucleic Acids Res 48: 1954-1968.

Gilbert, W. V., 2011 Functional specialization of ribosomes? Trends Biochem Sci 36: 127-132.

Gray, M. W., 1981 Unusual pattern of ribonucleic acid components in the ribosome of *Crithidia fasciculata*, a trypanosomatid protozoan. Mol Cell Biol 1: 347-357.

Hałakuc, P., A. Karnkowska and R. Milanowski, 2022 Typical structure of rRNA coding genes in diplonemids points to two independent origins of the bizarre rDNA structures of euglenozoans. BMC Ecol Evol 22: 59.

Hashem, Y., A. des Georges, J. Fu, S. N. Buss, F. Jossinet *et al.*, 2013 High-resolution cryo-electron microscopy structure of the *Trypanosoma brucei* ribosome. Nature 494: 385-389.

Lecompte, O., R. Ripp, J. C. Thierry, D. Moras and O. Poch, 2002 Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. Nucleic Acids Res 30: 5382-5390.

Liu, Z., C. Gutierrez-Vargas, J. Wei, R. A. Grassucci, M. Ramesh *et al.*, 2016 Structure and assembly model for the *Trypanosoma cruzi* 60S ribosomal subunit. Proc Natl Acad Sci U S A 113: 12174-12179.

Martinez-Seidel, F., O. Beine-Golovchuk, Y. C. Hsieh and J. Kopka, 2020 Systematic review of plant ribosome heterogeneity and specialization. Front Plant Sci 11: 948.

Matzov, D., M. Taoka, Y. Nobe, Y. Yamauchi, Y. Halfon *et al.*, 2020 Cryo-EM structure of the highly atypical cytoplasmic ribosome of *Euglena gracilis*. Nucleic Acids Res 48: 11750-11761.

Norris, K., T. Hopes and J. L. Aspden, 2021 Ribosome heterogeneity and specialization in development. Wiley Interdiscip Rev RNA 12: e1644.

Schnare, M. N., J. R. Cook and M. W. Gray, 1990 Fourteen internal transcribed spacers in the circular ribosomal DNA of *Euglena gracilis*. J Mol Biol 215: 85-91.

Schnare, M. N., and M. W. Gray, 1990 Sixteen discrete RNA components in the cytoplasmic ribosome of *Euglena gracilis*. J Mol Biol 215: 73-83.

Shi, Z., K. Fujii, K. M. Kovary, N. R. Genuth, H. L. Röst *et al.*, 2017 Heterogeneous ribosomes preferentially translate distinct subpools of mRNAs genome-wide. Mol Cell 67: 71-83.e77.

Slavov, N., S. Semrau, E. Airoldi, B. Budnik and A. van Oudenaarden, 2015 Differential stoichiometry among core ribosomal proteins. Cell Rep 13: 865-873.

Spencer, D. F., J. C. Collings, M. N. Schnare and M. W. Gray, 1987 Multiple spacer sequences in the nuclear large subunit ribosomal RNA gene of *Crithidia fasciculata*. EMBO J 6: 1063-1071.

11. Meiosis in *Diplonema papillatum*?

INTRODUCTION

Meiosis is a prerequisite for sexual reproduction as it produces haploid gamete cells. In *D. papillatum* and other diplonemids, meiotic processes have not been observed and are not induced by starvation as in many other systems. This raises the questions of (i) whether *D. papillatum* has the potential to reproduce sexually and, if so, (ii) whether the crossover is mediated by the synaptonemal complex, which is the preferred pathway in most examined eukaryotic species.

At the cell-biology level, meiosis proceeds in two principal stages. First, homologous chromosomes pair with each other, followed by programmed double-strand breaks and strand exchange. Second, meiotic crossovers are formed, and then the joint molecules are resolved (KOHL AND SEKELSKY 2013). In organisms for which cytological evidence is lacking or genetics tools are not available, the potential for sexual reproduction is examined by scrutinizing the genome for homologs of genes involved in the meiosis of model organisms (*Saccharomyces cerevisiae*, *Schizosaccharomyces pombe, Drosophila,* and *C. elegans*). Below, we will briefly discuss the genes involved in the various steps of meiosis and report for which genes we detected homologs in the *D. papillatum* nuclear genome assembly version 1.

RESULTS AND DISCUSSION

Genes involved in early meiosis

Pairing of homologous chromosomes proceeds in three steps, each catalyzed or controlled by a specific set of genes conserved across eukaryotes (KOHL AND SEKELSKY 2013). In the following, the names of genes and proteins that are meiosis-specific are shown in red, and those that also act in mitosis or other processes are shown in turquoise; genes detected in the ***D. papillatum*** nuclear genome are set in bold (**Supplementary Table S19**).

The three steps of homologous chromosome pairing and the proteins involved in these steps are:

1. Programmed DNA double-strand breaks (DSBs), induced by the **SPO11** endonuclease;
2. Processing of DSBs and ssDNA regions, performed by the Replication Protein A (**RPA**), a heteromeric complex;
3. Formation of extended D-loops resulting in joint molecules, catalyzed by the recombinases **DMC1** and **RAD51**, and assisted by **HOP2** and **MND1**.

In yeast, REC102, REC104 and REC114 have been shown to interact with **SPO11** (SASANUMA *et al.* 2007; KOHL AND SEKELSKY 2013), but they appear to have no homologs outside the fungi.

The second part of meiosis

Subsequent meiotic crossovers and the resolution of the joint molecule may follow either of two distinct crossover pathways, Pathway I or II. Most eukaryotes are equipped for both pathways but prefer using one or the other.

Crossover Pathway I is characterized by the formation of a synaptonemal complex during meiotic prophase I, a structure that is thought to stabilize meiotic pairing. This pathway is the preferred route in, e.g., *Caenorhabditis* and *Drosophila* ((KOHL AND SEKELSKY 2013) and references therein). Genes encoding components of the synaptonemal complex, such as the lateral-element proteins RED1 and HOP1 and the recombination-nodule proteins ZIP1, ZIP2, ZIP3, and SPO22 are specific for Pathway I. Further involved in this pathway are the DNA helicase MER3 and the mutS-homologs **MSH4** and **MSH5** (PAGE AND HAWLEY 2004). The latter two proteins also participate in general DNA repair. Note that eukaryotes typically encode four distinct mutS-homologs (yeast even has a total of six), which are highly similar to each other in their protein sequence. In the *Diplonema* nuclear genome, we were unable to retrieve any of the homologs of the meiosis-specific Pathway-I components RED1, HOP1 and ZIP1-4, except a moderately supported MER3 (see phylogeny below). However, we detected homologs of **MSH4** and **MSH5**, and further MSH2, MSH3, MSH6 that are all probably involved in general DNA repair.

Crossover Pathway II, the exclusive mode of crossover formation in the fission yeast, uses mitotic DNA-repair functions involving non-interfering crossovers. Thus, Pathway-II-specific meiosis genes do not exist. The complex that resolves meiotic crossovers in *S. pombe* is composed of **MUS81** and MMS4. Experimental evidence shows that *Tetrahymena thermophila* generates crossovers by the Class II pathway. According to their gene repertoire, i.e., absence of Pathway I genes and presence of **MUS81**, other ciliates are probably also capable of Class II crossovers (see (CHI *et al.* 2014) and references therein). Note that MMS4 has not been detected in ciliate genomes, suggesting that it is not essential for Pathway II. In the *Diplonema* nuclear genome, we detected a homolog of **MUS81**, but no significant hit with MMS4.

Meiotic divisions

Essentially all eukaryotes go through haploid and diploid phases, and depending on the predominant stage, a species is classified as being haploid or diploid. Consequently, there is no difference in the meiosis of haploid and diploid organisms. The (either interim or long-term) diploid undergoes two rounds of specialized cell divisions, Meiosis I and II, without intervening S-phase. Prior to entry into Meiosis I, homologous chromosomes (originating from each parent) pair and undertake meiotic recombination, thereby generating new genetic combinations in the offspring. During this step, homologous chromosomes are segregated into daughter cells, whereas sister chromatids are segregated during Meiosis II. In the case of diploid organisms, the haploid products of Meiosis II function as sexual spores that fuse to form a zygote, which then multiplies by mitosis. In the case of haploid species, the haploid product of Meiosis II corresponds to the 'adult' organism that directly multiplies by mitosis.

Meiosis II largely resembles mitotic division. One of the differences found, e.g., in yeast and human, is that the cohesin subunit **Rad21** (termed SSC1 in *S. cerevisiae*), which otherwise functions in mitosis, is substituted by a specialized meiotic form, named REC8 (WASSMANN 2013).

Distinction of closely related homologs

Several meiosis proteins have closely related counterparts and are therefore easily confused. These include **DMC1** and **RAD51**; MSH2, MSH3, **MSH4**, **MSH5** and MSH6; **RAD21** and REC8; and MER3 and other ATP-dependent DNA and RNA helicases. All of the above function assignments of *D. papillatum* proteins have been confirmed by protein-domain content and arrangement. In the case of confounding paralogs, we performed phylogenetic analyses and transferred the function annotation of a given SwissProt protein to the *Diplonema* protein, if they group together in a well-supported clade.

Diplonema **DMC1** and **RAD51** could be clearly distinguished based on their affiliation with the corresponding kinetoplastid homologs (**Supplementary Figure S17A**). Similarly, MSH2 to MSH6 from diverse eukaryotic groups including *D. papillatum* form coherent clades (**Supplementary Figure S17B**). To validate the DNA-helicase MER3 candidates of *Diplonema*, we added to the dataset DEAD/DEAH box RNA helicases and U5 small nuclear RNA helicase that have significant sequence similarity to MER3. Of the initial three MER3 candidates in *Diplonema*, one affiliates with the MER3 clade, including confirmed orthologs from plants, human, and yeast (**Supplementary Figure S17C**).

Evidence of meiosis in Discoba

Finding orthologs of genes whose involvement in meiosis has been demonstrated experimentally in model organisms does not prove that *D. papillatum* is indeed capable of sexual reproduction. The same caution applies, e.g., to the discoban *Trichomonas* (MALIK *et al.* 2007) and the dinoflagellate *Symbiodinium*. In these organisms, 'meiosis' genes had been identified, but otherwise, cytological or genetic evidence for sexual processes is lacking (CHI *et al.* 2014). . In addition, meiosis genes can be present in a genome despite the absence of sexual reproduction, as was demonstrated in *Giardia intestinalis* where **DMC1**, **SPO11**, and HOP1 act during *parasexual* genetic recombination (CARPENTER *et al.* 2012). Further, in *Acanthamoeba*, meiosis genes are expressed constitutively, which led to the proposition that the corresponding gene products are involved in biological processes other than meiosis (MACIVER *et al.* 2019).

Although protists had historically been considered to propagate only clonally, cytological evidence supporting meiotic processes exists for quite a number of groups (reviewed in (HEYWOOD 1976)). Among Discoba, conventional meiotic division has been demonstrated in trypanosomes by examining the inheritance of genetic markers. Meiosis is further supported by the finding that *Trypanosoma brucei* temporally expresses **DMC1**, HOP1, **MND1**, and **SPO11**, and forms a synaptonemal complex that was demonstrated by immunofluorescence experiments ((PEACOCK *et al.* 2011) and references therein).

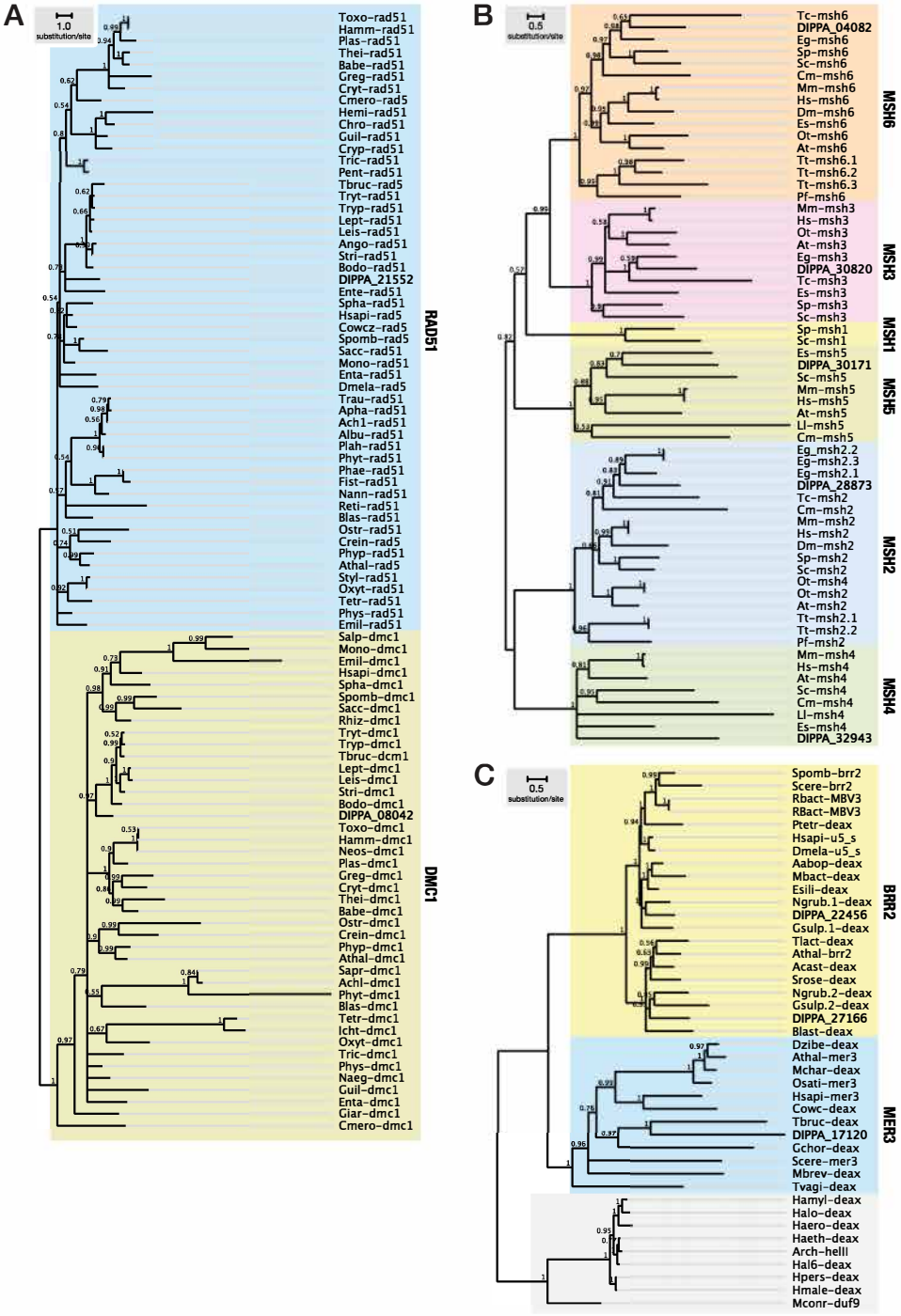
In euglenids, traditionally believed to reproduce exclusively mitotically, indications for meiosis come from cytological studies of *Hyalophacus* (LEEDALE 1967). For *Euglena gracilis*, corresponding experimental data are not available. Still, its genome includes a complete set of meiosis-specific genes (EBENEZER *et al.* 2017), although **DMC1**, REC8, and **SPO11** seem not to be transcribed (under the conditions tested (EBENEZER *et al.* 2019)).

Sexual reproduction of the heterolobosean *Naegleria* was suggested based on isoenzyme analysis (PERNIN *et al.* 1992). However, isoenzymes can arise from conditions other than meiosis, for example paralogs that arose via gene duplication or post-translational processing. Therefore, isoenzyme analyses are only conclusive in connection with genetic crosses and the analysis of offspring as conducted for *T. brucei* (SCHWEIZER *et al.* 1994). Finally, several studies have provided indirect evidence for sexual processes in protists, e.g., by searching for the existence of genes encoding proteins that function in cell and nuclear fusion (SPEIJER *et al.* 2015).

Supplementary Table S19. *D. papillatum* homologs of genes involved in meiosis

Protein Name/component[a]	Gene[b]	Distribution[c]	Synonym	Dp protein ID	TPM (RNA-Seq)[d]
Meiotic recombination	DMC1*	broad	LIM15	DIPPA_08042	32.47744 (7,852)
Mei-specific/SC lateral element	HOP1*	broad	ASY1	DIPPA_03672	2.049079 (622)
Homol-pairing/Class II pathway	HOP2*	broad	/	DIPPA_04049	0.756920 (101)
ATP-dep DNA helicase/SC?	MER3	broad	HFM1	DIPPA_17120	0.138232 (224)
Meiotic nuclear division	MND1*	broad	/	DIPPA_03210	0.74791 (96)
DNA mismatch repair/SC	MSH4	broad	/	DIPPA_32943	0.426741 (408)
DNA mismatch repair/SC	MSH5	broad	/	DIPPA_30171	0.564389 (360)
Meiotic recombination	REC114	Ascom	REC7	/	/
Meiotic recombination	REC8*	broad	SYN2	/	/
Reductive division/SC axial element	RED1	Ascom	REC10	/	/
Meiotic recombination	SPO11*	broad	REC12	DIPPA_09542	0.507504 (130)
SC central element	ZIP1	Ascom	/	/	/
SC central element	ZIP2	Ascom	/	/	/
SC central element	ZIP3	broad	/	/	/
SC central element	ZIP4	Ascom	SPO22	/	/
Complex formation of SPO11, SC	REC102	Ascom	/	/	/
Meiosis specific	MEI4	Opisth	/	/	/
Recombination protein	MER2	Ascom	REC107	/	/
DSB-repair	MRE11	broad	/	DIPPA_08235	11.030102 (2,396)
DNA mismatch repair	MSH1	broad	/	/	/
DNA mismatch repair	MSH2	broad	/	DIPPA_28873	3.458633 (2,452)
DNA mismatch repair	MSH3	broad	/	DIPPA_30820	7.495374 (5,815)
DNA mismatch repair	MSH6	broad	/	DIPPA_04082	1.134347 (881)
Cross-over junction endonuclease	MUS81	broad	/	DIPPA_25323	1.026156 (1,926)
Sisterchromatid separation	RAD21	broad	SSC1	DIPPA_12003	24.029910 (9,023)
DNA repair	RAD50	Ascom	/	DIPPA_22997	3.062462 (1,702)
				DIPPA_20691	3.134435 (1,742)
				DIPPA_12463	5.720618 (2,516)
Homologous-recombination repair	RAD51	broad	/	DIPPA_21552	3.939032 (969)
Antiviral SKI8	REC103	Ascom	SKI8	/	/
Meiotic recombination	REC104	Ascom	/	/	/
Meiotic recombination	REC114	Opisth	/	/	/
Replication factor A protein 1	RPA1	broad	RPA 70 kDa	DIPPA_33311	22.036753 (8,573)
Replication factor A protein 2	RPA2	broad	RPA 32 kDa	DIPPA_08046	9.324137 (1,402)
DNA repair	XRS2	Ascom	/	/	/

[a] Protein name/function from SwissProt or literature; SC, synaptonemal complex.
[b] Asterisks indicate proteins considered to be a core meiotic component (Peacock *et al.* 2011; Chi *et al.* 2014); underscoring of protein names indicates homologs that require phylogenetic analysis for accurate function assignment; **red**, meiosis-specific proteins; **blue**, proteins that participate in meiosis and mitosis or other processes.
[c] Taxonomic distribution. Ascom, Ascomycetes; Opisth, Opisthokonts.
[d] Transcripts per Million (TPM) and RNA-Seq reads mapped, i.e., read count (in brackets). For details on transcript quantification based on RNA-Seq read mapping, see **Supplementary Information**: Section 2. Assembly and annotation of the nuclear genome and transcriptome of *Diplonema papillatum*.


A
1.0
substitution/site
Toxo-rad51
Hamm-rad51
Plas-rad51
Thei-rad51
Babe-rad51
Greg-rad51
Cryt-rad51
Cmero-rad5
Hemi-rad51
Chro-rad51
Guil-rad51
Cryp-rad51
Tric-rad51
Pent-rad51
Tbruc-rad5
Tryt-rad51
Tryp-rad51
Lept-rad51
Leis-rad51
Ango-rad51
Stri-rad51
Bodo-rad51
DIPPA_21552
Ente-rad51
Spha-rad51
Hsapi-rad5
Cowcz-rad5
Spomb-rad5
Sacc-rad51
Mono-rad51
Enta-rad51
Dmela-rad5
Trau-rad51
Apha-rad51
Ach1-rad51
Albu-rad51
Plah-rad51
Phyt-rad51
Phae-rad51
Fist-rad51
Nann-rad51
Reti-rad51
Blas-rad51
Ostr-rad51
Crein-rad5
Phyp-rad51
Athal-rad5
Styl-rad51
Oxyt-rad51
Tetr-rad51
Phys-rad51
Emil-rad51
RAD51
Salp-dmc1
Mono-dmc1
Emil-dmc1
Hsapi-dmc1
Spha-dmc1
Spomb-dmc1
Sacc-dmc1
Rhiz-dmc1
Tryt-dmc1
Tryp-dmc1
Tbruc-dcm1
Lept-dmc1
Leis-dmc1
Stri-dmc1
Bodo-dmc1
DIPPA_08042
Toxo-dmc1
Hamm-dmc1
Neos-dmc1
Plas-dmc1
Greg-dmc1
Cryt-dmc1
Thei-dmc1
Babe-dmc1
Ostr-dmc1
Crein-dmc1
Phyp-dmc1
Athal-dmc1
Sapr-dmc1
Achl-dmc1
Phyt-dmc1
Blas-dmc1
Tetr-dmc1
Icht-dmc1
Oxyt-dmc1
Tric-dmc1
Phys-dmc1
Naeg-dmc1
Guil-dmc1
Enta-dmc1
Giar-dmc1
Cmero-dmc1
DMC1
B
0.5
substitution/site
Tc-msh6
DIPPA_04082
Eg-msh6
Sp-msh6
Sc-msh6
Cm-msh6
Mm-msh6
Hs-msh6
Dm-msh6
Es-msh6
Ot-msh6
At-msh6
Tt-msh6.1
Tt-msh6.2
Tt-msh6.3
Pf-msh6
MSH6
Mm-msh3
Hs-msh3
Ot-msh3
At-msh3
Eg-msh3
DIPPA_30820
Tc-msh3
Es-msh3
Sp-msh3
Sc-msh3
MSH3
Sp-msh1
Sc-msh1
Es-msh1
DIPPA_30171
MSH1
Sc-msh5
Mm-msh5
Hs-msh5
At-msh5
Ll-msh5
Cm-msh5
MSH5
Eg_msh2.2
Eg-msh2.3
Eg-msh2.1
DIPPA_28873
Tc-msh2
Cm-msh2
Mm-msh2
Hs-msh2
Dm-msh2
Sp-msh2
Sc-msh2
Ot-msh4
Ot-msh2
At-msh2
Tt-msh2.1
Tt-msh2.2
Pf-msh2
MSH2
Mm-msh4
Hs-msh4
At-msh4
Sc-msh4
Cm-msh4
Ll-msh4
Es-msh4
DIPPA_32943
MSH4
C
0.5
substitution/site
Spomb-brr2
Scere-brr2
Rbact-MBV3
RBact-MBV3
Ptetr-deax
Hsapi-u5_s
Dmela-u5_s
Aabop-deax
Mbact-deax
Esili-deax
Ngrub.1-deax
DIPPA_22456
Gsulp.1-deax
Tlact-deax
Athal-brr2
Acast-deax
Srose-deax
Ngrub.2-deax
Gsulp.2-deax
DIPPA_27166
Blast-deax
BRR2
Dzibe-deax
Athal-mer3
Mchar-deax
Osati-mer3
Hsapi-mer3
Cowc-deax
Tbruc-deax
DIPPA_17120
Gchor-deax
Scere-mer3
Mbrev-deax
Tvagi-deax
MER3
Hamyl-deax
Halo-deax
Haero-deax
Haeth-deax
Arch-helII
Hal6-deax
Hpers-deax
Hmale-deax
Mconr-duf9

Supplementary Figure S17. Phylogenetic analysis of proteins involved in meiosis. Posterior likelihood values of bipartitions are indicated. The function assignment of clades (vertical protein name) is based on that of the proteins in UniProtKB 'reviewed (SwissProt)' that are part of the corresponding clades. Branches of clades uniting kinetoplastid and *D. papillatum* sequences are colored in purple. Protein candidates from *D. papillatum* (DIPPA_) are shown in bold font. **A, DMC1 and RAD51.** Taxon names are abbreviated as follows: Ach1, Achl, *Achlya hypogyna;* Albu, *Albugo laibachii;* Ango, *Angomonas deanei;* Apha, *Aphanomyces astaci;* Athal, *Arabidopsis thaliana;* Babe, *Babesia microti;* Blas, *Blastocystis sp.;* Bodo, *Bodo saltans;* Chro, *Chroomonas mesostigmatica;* Cowcz, *Capsaspora owczarzaki;* Crein, *Chlamydomonas reinhardtii;* Cryp, *Cryptomonas paramecium;* Cryt, *Cryptosporidium andersoni;* Cmero, *Cyanidioschyzon merolae;* Dmela, *Drosophila melanogaster;* Emil, *Emiliania huxleyi;* Enta, *Entamoeba histolytica;* Ente, *Enterospora canceri;* Fist, *Fistulifera solaris;* Giar, *Giardia intestinalis;* Greg, *Gregarina niphandrodes;* Guil, *Guillardia theta;* Hamm, *Hammondia hammondi;* Hemi, *Hemiselmis andersenii;* Hsapi, *Homo sapiens;* Icht, *Ichthyophthirius multifiliis;* Leis, *Leishmania major;* Lept, *Leptomonas pyrrhocoris* and *seymouri;* Mono, *Monosiga brevicollis;* Nann, *Nannochloropsis gaditana;* Ostr, *Ostreococcus lucimarinus;* Oxyt, *Oxytricha trifallax;* Phae, *Phaeodactylum tricornutum;* Phyp, *Physcomitrella patens;* Phyt, *Phytophthora nicotianae;* Phys, *Physarum polycephalum;* Plah, *Plasmopara halstedii;* Reti, *Reticulomyxa filosa;* Rhiz, *Rhizophagus irregularis;* Sacc, *Saccharomyces cerevisiae;* Salp, *Salpingoeca rosetta;* Sapr, *Saprolegnia diclina;* Spha, *Sphaeroforma arctica;* Spomb, *Schizosaccharomyces pombe;* Tetr, *Tetrahymena thermophila;* Trau, *Thraustotheca clavata;* Stri, *Strigomonas culicis;* Styl, *Stylonychia lemnae;* Tryp, *Trypanosoma cruzi,* Tryt, *Trypanosoma theileri.* RAD51 in *Giardia* is probably a DMC1 ortholog. **B. MSH2 to MSH6**. Taxon names are abbreviated as follows: At, *Arabidopsis thaliana;* Cm, *Cyanidioschyzon merolae;* Dm, *Drosophila melanogaster;* Es, *Ectocarpus siliculosus;* Eg, *Euglena gracilis;* Hs, *Homo sapiens;* Ll, *Leishmania major;* Mm, *Mus musculus;* Ot, *Ostreococcus tauri;* Pf, *Plasmodium malariae and falciparum;* Sc, *Saccharomyces cerevisiae;* Tc, *Trypanosoma brucei, cruzi cruzi and theileri;* Tt, *Tetrahymena thermophila.* **C, MER3.** Taxon names are abbreviated as follows: Aabop, *Aureococcus anophagefferens;* Acast, *Acanthamoeba castellanii;* Athal, *rabidopsis thaliana;* Blast, *Blastocystis sp.;* Cowc, *Capsaspora owczarzaki;* Dzibe, *Durio zibethinus;* Esili, *Ectocarpus siliculosus;* Gchor, *Gracilariopsis chorda;* Csulp, *Galdieria sulphuraria;* Haero, *Haloplanus aerogenes;* Haeth, *Halorubrum aethiopicum;* Hal6, *Halorubrum sp.;* Hamyl, *Halogranum amylolyticum;* Hmale, *Halopenitus malekzadehii;* Hpers. *Halopenitus persicus;* Hsapi, *Homo sapiens;* Mbact, *Myxococcaceae bacterium;* Mbrev, *Monosiga brevicollis;* Hchar, *Momordica charantia;* Mconr, *Methanocella conradii;* Ngrub, *Naegleria gruberi;* Osati, *Oryza sativa;* Ptetr, *Paramecium tetraurelia;* RBact, Rbact, *Rickettsiales bacterium;* Scere, *Saccharomyces cerevisiae;* Spomb, *Schizosaccharomyces pombe;* Tbruc, *Trypanosoma brucei brucei;* Tlact, *Tieghemostelium lacteum;* Tvagi, *Trichomonas vaginalis.*

METHODS

Homology detection by pairwise sequence alignment and Hidden Markov model searches

We analyzed the proteins inferred from version 1.0 of the *D. papillatum* assembly. As query 'meiosis' proteins, we used those whose involvement in meiosis has been demonstrated (asterisks indicate proteins considered to be specific for meiosis): DMC1*, HOP1*, HOP2*, MER3, MND1*, MSH4, MSH5, REC114, REC8*, RED1, SPO11*, ZIP1, ZIP2, ZIP3, ZIP4, REC102, MEI4, MER3, MRE11, MSH1, MSH2, MSH3, MSH6, MUS81, RAD21, RAD50, RAD51, REC103, REC104, REC114, RPA1, RPA2, and XRS2. Sequences were downloaded from UniProtKB release 2021_3 (https://www.uniprot.org/help/uniprotkb). For similarity searches, we employed local blast and fasta as a rapid test. In addition, profile HMMs were generated for searches with HMMERsearch v3.3 from the HMMER suite (EDDY 2009) in the *Diplonema* inferred proteome. Further, we performed HMMscan of *D. papillatum* proteins against the PFAM profile database v35.0 (http://pfam.xfam.org) to test if candidates retrieved by similarity search do indeed contain the same conserved protein domains as the confirmed meiosis proteins.

Distinction of orthologs and paralogs

Meiosis proteins that are easily confused, notably DMC1 and RAD51, MSH1 to MSH6, RAD21 and REC8, and MER3 and other ATP-dependent DNA helicases, were distinguished *via* reciprocal blast and inspection of the protein-domain content and arrangement and *via* phylogenetic inference. For the latter analysis, proteins were pre-aligned with Muscle v3.8.155 and a profile HMM was built from the multiple alignment using HMMbuild with default parameters. The profile HMM served for building a final multiple protein alignment with HMMalign, default parameters, which then was used to construct a phylogenetic tree with Phylobayes v4.1c (LARTILLOT *et al.* 2009) using the options -cat -gtr -dgam 6 –dc, essentially as described previously (VALACH *et al.* 2017). The function assignment of clades as shown in the figures is based on that of the UniProt proteins labelled 'reviewed (SwissProt)' that make part of a clade.

AUTHOR CONTRIBUTIONS

Conceptualization, Data curation, Investigation, Formal analysis, Writing, original draft – G.B.; **Visualization** – M.V.; **Writing, review & editing –** all co-authors.

REFERENCES

Carpenter, M. L., Z. J. Assaf, S. Gourguechon and W. Z. Cande, 2012 Nuclear inheritance and genetic exchange without meiosis in the binucleate parasite *Giardia intestinalis*. J Cell Sci 125: 2523-2532.

Chi, J., F. Mahe, J. Loidl, J. Logsdon and M. Dunthorn, 2014 Meiosis gene inventory of four ciliates reveals the prevalence of a synaptonemal complex-independent crossover pathway. Mol Biol Evol 31: 660-672.

Ebenezer, T. E., M. Zoltner, A. Burrel, A. Nenarokova, A. M. G. N. Vanclová *et al.*, 2017 Unlocking the biological potential of *Euglena gracilis*: evolution, cell biology and significance to parasitism. bioRxiv: 228015.

Ebenezer, T. E., M. Zoltner, A. Burrell, A. Nenarokova, A. M. G. Novák Vanclová *et al.*, 2019 Transcriptome, proteome and draft genome of *Euglena gracilis*. BMC Biol 17: 11.

Eddy, S. R., 2009 A new generation of homology search tools based on probabilistic inference. Genome Inform 23: 205-211.

Heywood, P., 1976 Algal sexuality. Nature 259: 425.

Kohl, K. P., and J. Sekelsky, 2013 Meiotic and mitotic recombination in meiosis. Genetics 194: 327-334.

Lartillot, N., T. Lepage and S. Blanquart, 2009 PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25: 2286-2288.

Leedale, G. F., 1967 Euglenida-euglenophyta. Annu Rev Microbiol 21: 31-48.

Maciver, S. K., Z. Koutsogiannis and A. de Obeso Fernández Del Valle, 2019 'Meiotic genes' are constitutively expressed in an asexual amoeba and are not necessarily involved in sexual reproduction. Biol Lett 15: 20180871.

Malik, S. B., A. W. Pightling, L. M. Stefaniak, A. M. Schurko and J. M. Logsdon, Jr., 2007 An expanded inventory of conserved meiotic genes provides evidence for sex in *Trichomonas vaginalis*. PLoS One 3: e2879.

Page, S. L., and R. S. Hawley, 2004 The genetics and molecular biology of the synaptonemal complex. Annu Rev Cell Dev Biol 20: 525-558.

Peacock, L., V. Ferris, R. Sharma, J. Sunter, M. Bailey *et al.*, 2011 Identification of the meiotic life cycle stage of *Trypanosoma brucei* in the tsetse fly. Proc Natl Acad Sci U S A 108: 3671-3676.

Pernin, P., A. Ataya and M. L. Cariou, 1992 Genetic structure of natural populations of the free-living amoeba, *Naegleria lovaniensis*. Evidence for sexual reproduction. Heredity (Edinb) 68: 173-181.

Sasanuma, H., H. Murakami, T. Fukuda, T. Shibata, A. Nicolas *et al.*, 2007 Meiotic association between Spo11 regulated by Rec102, Rec104 and Rec114. Nucleic Acids Res 35: 1119-1133.

Schweizer, J., H. Pospichal, G. Hide, N. Buchanan, A. Tait *et al.*, 1994 Analysis of a new genetic cross between two East African *Trypanosoma brucei* clones. Parasitology 109 (Pt 1): 83-93.

Speijer, D., J. Lukeš and M. Eliáš, 2015 Sex is a ubiquitous, ancient, and inherent attribute of eukaryotic life. Proc Natl Acad Sci U S A 112: 8827-8834.

Valach, M., S. Moreira, S. Hoffmann, P. F. Stadler and G. Burger, 2017 Keeping it complicated: mitochondrial genome plasticity across diplonemids. Sci Rep 7: 14166.

Wassmann, K., 2013 Sister chromatid segregation in meiosis II: deprotection through phosphorylation. Cell Cycle 12: 1352-1359.

12. CAZyme-coding genes in *Diplonema papillatum*

INTRODUCTION

The complement of Carbohydrate-Active enzymes (CAZymes) of an organism can provide important clues about its metabolism and "lifestyle". CAZymes are highly diverse, with currently 173 Glycoside Hydrolase (GH) families, 42 Polysaccharide Lyase (PL) families, 20 Carbohydrate Esterase (CE) families, and 114 GlycosylTransferase (GT) families (DRULA *et al.* 2022) (http://www.cazy.org). CAZymes may also include, in addition to catalytic domains, Carbohydrate-Binding Modules (CBMs), which currently form 89 families. To get a glimpse at the nutrient resources and carbon storage of *D. papillatum*, we searched in its genome for homologs known from model organisms to be involved in the assembly and breakdown of glycans. We also compared the spectrum of CAZymes from *Diplonema* to that of two other euglenozoans: the free-living heterotroph *Bodo saltans* (Kinetoplastea) and the photoautotroph *Euglena gracilis* (Euglenida) (JACKSON *et al.* 2016; OPPERDOES *et al.* 2016; EBENEZER *et al.* 2019).

RESULTS

CAZyme complement of *D. papillatum*

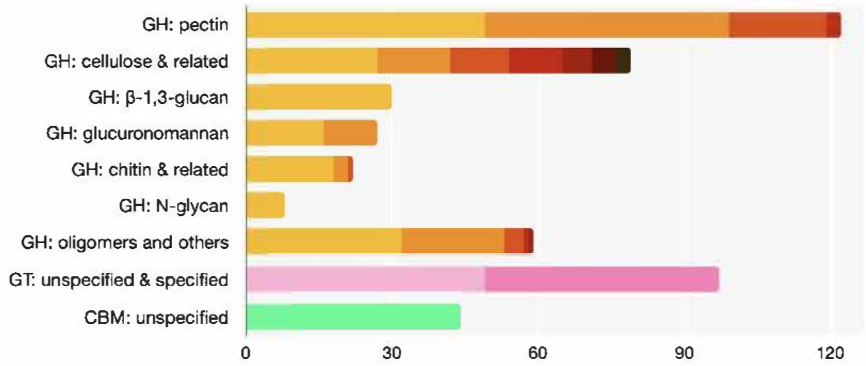
The nuclear genome assembly of *D. papillatum* contains as many as 489 CAZymes for metabolizing diverse polysaccharides (**Supplementary Figure S18, Additional File 4** cazymeList). The repertoire includes 52 out of the 173 described GH families, 4 out of 20 CEs, 3 out of 42 PLs, 36 out of 114 GTs, and 11 out of 89 CBMs. Essentially all CAZyme genes we detected are transcribed, and most are likely translated, as peptides of 11% of the enzymes were identified in low-depth mass-spectrometry data (**Additional File 4** cazymeList).

Based on the repertoire of CAZyme domains, *D. papillatum* appears to specialize strongly in plant biomass degradation. Genes involved in the breakdown of **pectin** (consisting primarily of methyl-esterified α-1,4-galacturonic acid units) are the largest and most diverse group. The corresponding 122 genes (all expert-validated) belong to nine distinct CAZyme families (GH28, GH53, GH54, GH78, GH145, CE8, CE13, PL1, and PL4). Most of these enzyme families have undergone significant expansions, exemplified by the GH28 family, which comprises 25 members. In addition, a total of 82 proteins belong to CAZyme families that degrade the β-1,4-linked glucose polymer **cellulose** (GH5, GH6, GH7, and GH45, as well and the carbohydrate-binding module family CBM1) or **hemicelluloses**, i.e., hemicellulose side chains (GH51), xylan (GH10), xyloglucan (GH74 and GH115), and β-linked xylo- and gluco-oligosaccharides (GH3). Further, *Diplonema* encodes a large set of enzymes (27 proteins from the families GH79, GH154, GH47, GH76, and GH92) for the digestion of the **sulfated glucuronomannan** (α-1,3-mannan with β-D-glucuronic acid side chains), which is the main polysaccharide present in the cell walls of diatoms (LE COSTAOUËC *et al.* 2017).

The presence of genes encoding family members of GH55, GH72, GH128, GH152, GH81, and potentially GH16 in the *Diplonema* genome assembly suggests that this protist can digest the β-1,3-glucan **laminarin**, which is the storage polysaccharide of micro- and macro-algae such as haptophytes, stramenopiles, and diatoms (MICHEL *et al.* 2010), the latter organismal group being the most abundant primary producer in the oceans (FALKOWSKI *et al.* 1998). Laminarin is estimated to make up ~10% of the carbon produced globally (BECKER *et al.* 2020).

The *D. papillatum* genome assembly also revealed 18 genes encoding GH18, GH19, and GH20 enzymes, shown in model organisms to break down **chitin** (polymer of N-acetylglucosamine), together with several GH56 and GH114 enzymes, assumed to degrade glycoaminoglycans. Thus, *Diplonema* should be able to break down these polysaccharides usually encountered in animals and fungi.

Finally, the genome encodes genes for 90 additional CAZymes (about equally divided between GTs and CBMs), whose substrates cannot be inferred with confidence. The CBM13 family is particularly large (25 members), with ~75% of proteins containing multiple copies of this non-catalytic module. Seven CMB13 modules are appended to a PL4 domain likely to cleave pectin, suggesting that the PL4-CMB13 proteins in *Diplonema* are involved in pectin degradation.



Supplementary Figure S18. Predicted carbohydrate-targeting proteins in *D. papillatum* containing at least one CAZyme domain. Proteins were grouped by their cognate substrate class. GH, glycosylhydrolases; GT, glycosyltransferases; CBM, carbohydrate-binding modules. The subdivision of the bars by different color shades represents the number of enzymes in the following subgroups. **GH: pectin** – pectin hydrolases, pectin lyases, pectin acetylesterases, and pectin methylesterases. **GH: cellulose & related** – cellulases, xylan-α-glucuronidases, xylan/cellulose and xylan/xyloglucan hydrolases, hemicellulases, β-glucan/β-xylan hydrolases, and β-mannanases. **GH: β-1,3-glucan** – no subgroups. **GH: glucuronomannan** – α-mannanases and β-glucuronidases. **GH: chitin & related** – chitinases, glycosaminoglycan, and glucosamine hydrolases. N-glycan – no subgroups. **GH: oligomers and others** – α-glycosidases (including α-mannosidases, α-fucosidases, α-glucosidases, and unspecified), β-glycosidases (including β-galactosidases, β-mannosidases, and unspecified), trehalases, an α-fucanase, and an invertase. **GT: unspecified & specified** – various substrates. **CBM: unspecified** – no subgroups (all substrates remain unspecified).

Comparative analyses of CAZymes profiles in Euglenozoa

The CAZyme portfolio of *D. papillatum* appears to lack enzymes from families the GH22, GH23, GH25, GH73, and CBM50 that target bacterial cell wall components, which points to a strictly eukaryotic diet. This finding contrasts with the gene repertoire and feeding behaviour of free-living members of *Diplonema*'s kinetoplastid sister group, represented by the marine and freshwater protist *B. saltans* (MITCHELL *et al.* 1988; OPPERDOES *et al.* 2016).

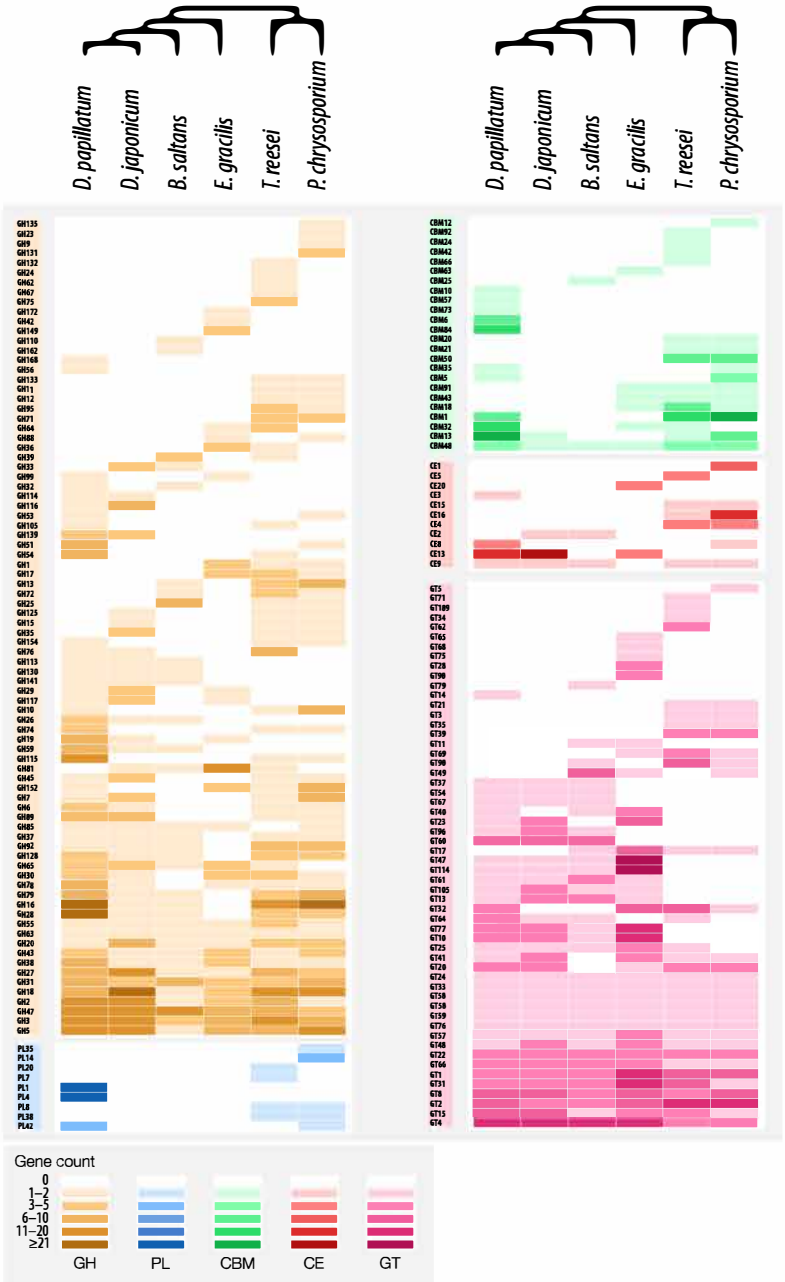
Further, the *Diplonema* genome assembly does not include any known α-glucan depolymerizing enzyme from families GH13 and GH77, which would suggest that this organism is unable to digest starch and glycogen, the carbon-storage compounds of Viridiplantae and Metazoa. However, we have demonstrated experimentally that *Diploma* readily utilizes amylopectin, a component of starch (see **Supplementary Information**: Section 13. Glycan and peptide assimilation by *Diplonema papillatum*). At the moment, it is impossible to identify the enzyme(s) responsible for this activity; they may be among the oligomer-targeting GH CAZymes or belong to novel families.

On the other hand, the absence of GH13 and GH77, together with the absence of the glycosyltransferase families GT3, GT5 and GT35, indicates that the form in which *Diplonema* stores carbon is not starch or glycogen. It is instead a β-1,3-glucan, as we found genes encoding members of the β-1,3-glucanase families GH55, GH72, GH128, GH152, and GH81. Supporting evidence for this form of carbon storage comes from the recent experimental identification of a β-1,3-glucan-containing polymer in *D. papillatum* (ŠKODOVÁ-SVERÁKOVÁ *et al.* 2020), a compound that has long been known from *Euglena*, called paramylon. *B. saltans* appears to store its carbon in the same form (reviewed in (RALTON *et al.* 2021)), which suggests that paramylon synthesis was already an attribute of the last common ancestor of Euglenozoa.

Lastly, we compared the distribution of CAZyme families in a systematic fashion across four euglenozoans, namely *D. papillatum*, its closest described relative *D. japonicum* (TASHYREVA *et al.* 2018), the free-living kinetoplastid *Bodo saltans* (JACKSON *et al.* 2016), and the euglenid *Euglena gracilis* (EBENEZER *et al.* 2019), as well as two representative, well-studied saprophytic fungi, *Trichoderma reesei* and *Phanerochaete chrysosporium* (**Supplementary Figure S19**). The *D. papillatum* genome encodes a number and diversity of GT families comparable to that of the above-listed species. However, it possesses a substantially larger and more diverse repertoire of carbohydrate-degrading enzymes, i.e., GH, CE, and PL families, as well as carbohydrate-recognizing proteins (CBM families). In particular, the 122-member pectin-targeting cohort of *D. papillatum* stands out, with the closest relative *D. japonicum* having only a quarter and the other four species less than 10% of that number. Enzymes and CBMs that target xylans and β-1,3-glucans are similarly enriched in *D. papillatum*. In contrast, *D. japonicum* has nearly twice as many chitin-metabolizing proteins (GH18, GH19, GH20) as *D. papillatum*.

The major limitation in comparing other diplonemids to *D. papillatum* is the lack of genomic information, with the proteome of *D. japonicum* being inferred from its reconstructed transcriptome alone. Although gene searches indicate that the coverage is good enough for retrieving well-conserved, low-expression genes (e.g., regulatory proteins involved in cell cycle, splicing, or endosome formation), the poly-A RNA-Seq sequencing depth is ~24 times lower than in our *D.*

papillatum dataset (~12 M vs. ~290 M read pairs). Thus, the *D. japonicum* CAZyme cohort inferred here is almost certainly underestimated. To get a more realistic measure of the scale of differences between the current and more genuine numbers, we compared the RNA-Seq coverage of transcripts between the two diplonemid species. Based on the bulk coverage difference and the number of reads mapped to individual CAZyme-encoding transcripts, we estimate that *D. japonicum* possesses 10–20% more CAZyme genes than the current tally (roughly equally distributed across functional categories).



Supplementary Figure S19. Repertoire of carbohydrate-targeting proteins across euglenozoans and fungi. Proteins belonging to the CAZyme classes GH, PL, CBM, CE, and GT from four free-living euglenozoans (*D. papillatum*, *D. japonicum*, *Bodo saltans*, and *Euglena gracilis*) and two model fungi (*Trichoderma reesei* and *Phanerochaete chrysosporium*) are compiled. Rows correspond to individual CAZyme families within the classes GH, PL, CBM, CE, and GT. Heatmap shading indicates gene counts in each genome as detailed in the key (bottom). Within CAZyme classes, families are ordered from top to bottom based on the increasing number of members detected in the family.

DISCUSSION

In contrast to the nuclear genomes of parasitic and photoautotrophic euglenozoans, the genome of *D. papillatum* encodes an extremely high number of diverse CAZymes for the degradation of polysaccharides, comparable to the CAZyme complement of saprophytic fungi (ALMÁSI *et al.* 2019; DÍAZ-ESCANDÓN *et al.* 2022). With 52 GH, 3 PL, 11 CBM, and 36 GT families, the CAZyme repertoire in *D. papillatum* is as diverse as in the model saprophytic ascomycete *Trichoderma reesei* (55 GH, 4 PL, 12 CBM, and 32 GT families; (MARTINEZ *et al.* 2008)) or the basidiomycete *Phanerochaete chrysosporium* (50 GH, 5 PL, 11 CBM, and 28 GT families; (MARTINEZ *et al.* 2004)). The only category widely present in fungi but lacking in *Diplonema* are lytic polysaccharide monooxygenases (LPMOs) that oxidatively cleave polysaccharides.

This remarkably versatile CAZyme outfit of *D. papillatum* argues against the notion that this organism follows a parasitic lifestyle in its natural environment. Like other diplonemids (TRIEMER AND OTT 1990; PROKOPCHUK *et al.* 2022), *D. papillatum* takes up nutrients *via* phagocytosis and/or osmotrophy, indicating that it feeds on the biomass produced by photosynthetic microeukaryotes colonizing the surface layers of the oceans. For example, it has the potential to digest all carbohydrate polymers, including glucuronomannan, chitin, and pectin which are typically found in the most abundant marine primary producers, the diatoms. In particular, the expansion of pectin-degrading families suggests that pectin has, over the course of evolution, become a preferred food for *D. papillatum*. In addition, many marine eukaryotes, including phototrophic diatoms (CHEN AND THORNTON 2015) and mixotrophic dinoflagellates (LARSSON *et al.* 2022), produce extracellular polymers such as the transparent exopolymer particles (TEPs). The major components of TEPs are yet uncharacterized, complex mixtures of acidic polysaccharides (reviewed in (PASSOW 2002; DECHO AND GUTIERREZ 2017)). It is conceivable that the enzyme families of *D. papillatum* predicted to degrade pectin (which is acidic as well) are also capable of breaking down TEPs. Together with the extensive repertoire of proteins with unknown carbohydrate targets, the substrate range of *Diplonema* is probably much broader than we currently appreciate.

METHODS

Generation of proteomes

The section **Supplementary Information**: Section 2. Assembly and annotation of the nuclear genome and transcriptome of *Diplonema papillatum* describes how the 'submission' proteome of *D. papillatum* was prepared. For additional diplonemids, the proteomes were generated as follows. First, publicly available poly-A RNA-Seq reads (KAUR *et al.* 2020) were cleaned using cutadapt v1.16 (http://journal.embnet.org/index.php/embnetjournal/article/view/200) and assembled using Trinity v2.4.0 (GRABHERR *et al.* 2011) with default parameters. We removed from the transcriptomes isoforms and incomplete reverse-complementary transcript fragments by two rounds of clustering using CD-HIT-EST v4.6 (FU *et al.* 2012). In the first round, all transcripts with 100% identity were clustered, and TransDecoder v5.5.0 (https://github.com/TransDecoder/TransDecoder) was used to infer coding sequences over 150 bp (i.e., length >50 amino acid residues), employing the standard genetic code. In the second round, the predicted coding sequences were clustered with CD-HIT-EST in local mode (option '-G 0') at 99% sequence identity and local overlap over 90% length between the cluster representative and embedded cluster members. Finally, PRINSEQ-lite v0.20.3 (SCHMIEDER AND EDWARDS 2011) was used to select only coding sequences longer than 450 bp (corresponding to proteins >150 aa) and protein sequences were regenerated using gotranseq v0.3.2 (https://github.com/feliixx/gotranseq) (standard genetic code, forward frame 1).

Protein identification and classification

Detection and assignment of CAZyme families in the inferred proteome of *D. papillatum* and the other seven diplonemids was performed using the methodology applied for the daily updates of the CAZy database (www.cazy.org), including expert validation (CANTAREL *et al.* 2009; LOMBARD *et al.* 2014).

Analyses of mass spectrometry data

We reanalyzed mass spectrometry data (ProteomeXchange ID: PXD025411) generated in the context of a prior study (ŠKODOVÁ-SVERÁKOVÁ *et al.* 2021). Raw data were first converted from the Thermo RAW format to mzML using ThermoRawFileParser v1.3.4 (HULSTAERT *et al.* 2020). Peptide searches were performed using MSFragger v3.5 (KONG *et al.* 2017), followed by filtering and scoring by Philosopher v4.4.0 (DA VEIGA LEPREVOST *et al.* 2020) and intensity-based quantification by IonQuant v1.8.0 (YU *et al.* 2020). Briefly, we searched for peptide-spectrum matches (PSMs) in a custom database of *D. papillatum* nuclear and mitochondrial proteins supplemented with frequent contaminants. Precursor and fragment mass tolerance were set to 20 ppm. We allowed up to two missed cleavage sites per protein for trypsin digestion. Carbamidomethylation of cysteine was specified as a fixed modification. Methionine oxidation, N-terminal protein acetylation, serine and threonine phosphorylation, and conversion of glutamine and glutamate at peptide N-termini to

pyrrolidone-carboxylic acid (PCA) were specified as variable modifications (up to three per peptide). Minimum and maximum peptide sizes were set to 700 and 5,000 Da, respectively. False discovery rates (FDR) for PSM and protein-identification probability were determined by the target-reversed decoy approach and set to 1%. Data processing and normalization by IonQuant were done for all ions using the topN strategy but without the match-between-runs option.

AUTHOR CONTRIBUTIONS

Conceptualization, Data curation, Formal analysis, Investigation, Writing, original draft – B.H., P.L., M.V.; **Visualization** – M.V.; **Writing, review & editing** – all authors.

REFERENCES

Almási, É., N. Sahu, K. Krizsán, B. Bálint, G. M. Kovács *et al.*, 2019 Comparative genomics reveals unique wood-decay strategies and fruiting body development in the Schizophyllaceae. New Phytol 224: 902-915.

Becker, S., J. Tebben, S. Coffinet, K. Wiltshire, M. H. Iversen *et al.*, 2020 Laminarin is a major molecule in the marine carbon cycle. Proc Natl Acad Sci U S A 117: 6599-6607.

Cantarel, B. L., P. M. Coutinho, C. Rancurel, T. Bernard, V. Lombard *et al.*, 2009 The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Res 37: D233-238.

Chen, J., and D. C. Thornton, 2015 Transparent exopolymer particle production and aggregation by a marine planktonic diatom (*Thalassiosira weissflogii*) at different growth rates. J Phycol 51: 381-393.

da Veiga Leprevost, F., S. E. Haynes, D. M. Avtonomov, H. Y. Chang, A. K. Shanmugam *et al.*, 2020 Philosopher: a versatile toolkit for shotgun proteomics data analysis. Nat Methods 17: 869-870.

Decho, A. W., and T. Gutierrez, 2017 Microbial Extracellular Polymeric Substances (EPSs) in ocean systems. Front Microbiol 8: 922.

Díaz-Escandón D., G. Tagirdzhanova, D. Vanderpool, C. C. G. Allen, A. Aptroot *et al.* 2022 Genome-level analyses resolve an ancient lineage of symbiotic ascomycetes. Curr Biol. 32: 5209-5218.

Drula, E., M. L. Garron, S. Dogan, V. Lombard, B. Henrissat *et al.*, 2022 The carbohydrate-active enzyme database: functions and literature. Nucleic Acids Res 50: D571-d577.

Ebenezer, T. E., M. Zoltner, A. Burrell, A. Nenarokova, A. M. G. Novák Vanclová *et al.*, 2019 Transcriptome, proteome and draft genome of *Euglena gracilis*. BMC Biol 17: 11.

Falkowski, P. G., R. T. Barber and V. V. Smetacek, 1998 Biogeochemical controls and feedbacks on ocean primary production. Science 281: 200-207.

Fu, L., B. Niu, Z. Zhu, S. Wu and W. Li, 2012 CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28: 3150-3152.

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson *et al.*, 2011 Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29: 644-652.

Hulstaert, N., J. Shofstahl, T. Sachsenberg, M. Walzer, H. Barsnes *et al.*, 2020 ThermoRawFileParser: modular, scalable, and cross-platform RAW file conversion. J Proteome Res 19: 537-542.

Jackson, A. P., T. D. Otto, M. Aslett, S. D. Armstrong, F. Bringaud *et al.*, 2016 Kinetoplastid phylogenomics reveals the evolutionary innovations associated with the origins of parasitism. Current Biology 26: 161-172.

Kaur, B., K. Záhonová, M. Valach, D. Faktorová, G. Prokopchuk *et al.*, 2020 Gene fragmentation and RNA editing without borders: eccentric mitochondrial genomes of diplonemids. Nucleic Acids Res 48: 2694-2708.

Kong, A. T., F. V. Leprevost, D. M. Avtonomov, D. Mellacheruvu and A. I. Nesvizhskii, 2017 MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. Nat Methods 14: 513-520.

Larsson, M. E., A. R. Bramucci, S. Collins, G. Hallegraeff, T. Kahlke *et al.*, 2022 Mucospheres produced by a mixotrophic protist impact ocean carbon cycling. Nat Commun 13: 1301.

Le Costaouëc, T., C. Unamunzaga, L. Mantecon and W. Helbert, 2017 New structural insights into the cell-wall polysaccharide of the diatom *Phaeodactylum tricornutum*. Algal Res 26: 172-179.

Lombard, V., H. Golaconda Ramulu, E. Drula, P. M. Coutinho and B. Henrissat, 2014 The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res 42: D490-495.

Martinez, D., R. M. Berka, B. Henrissat, M. Saloheimo, M. Arvas *et al.*, 2008 Genome sequencing and analysis of the biomass-degrading fungus *Trichoderma reesei* (syn. *Hypocrea jecorina*). Nat Biotechnol 26: 553-560.

Martinez, D., L. F. Larrondo, N. Putnam, M. D. Gelpke, K. Huang *et al.*, 2004 Genome sequence of the lignocellulose degrading fungus *Phanerochaete chrysosporium* strain RP78. Nat Biotechnol 22: 695-700.

Michel, G., T. Tonon, D. Scornet, J. M. Cock and B. Kloareg, 2010 Central and storage carbon metabolism of the brown alga *Ectocarpus siliculosus*: insights into the origin and evolution of storage carbohydrates in eukaryotes. New Phytol 188**:** 67-81.

Mitchell, G. C., J. H. Baker and M. A. Sleigh, 1988 Feeding of a freshwater flagellate, *Bodo saltans,* on diverse bacteria. J Protozool 35**:** 219-222.

Opperdoes, F. R., A. Butenko, P. Flegontov, V. Yurchenko and J. Lukeš, 2016 Comparative metabolism of free-living *Bodo saltans* and parasitic trypanosomatids. J Eukaryot Microbiol 63**:** 657-678.

Passow, U., 2002 Transparent exopolymer particles (TEP) in aquatic environments. Progress Oceanogr 55**:** 287-333.

Prokopchuk, G., T. Korytář, V. Juricová, J. Majstorović, A. Horák *et al.*, 2022 Trophic flexibility of marine diplonemids - switching from osmotrophy to bacterivory. ISME J 16**:** 1409-1419.

Ralton, J. E., M. F. Sernee and M. J. McConville, 2021 Evolution and function of carbohydrate reserve biosynthesis in parasitic protists. Trends Parasitol 37**:** 988-1001.

Schmieder, R., and R. Edwards, 2011 Quality control and preprocessing of metagenomic datasets. Bioinformatics 27**:** 863-864.

Škodová-Sveráková, I., G. Prokopchuk, P. Peña-Diaz, K. Záhonová, M. Moos *et al.*, 2020 Unique dynamics of paramylon storage in the marine euglenozoan *Diplonema papillatum*. Protist 171**:** 125717.

Škodová-Sveráková, I., K. Záhonová, V. Juricová, M. Danchenko, M. Moos *et al.*, 2021 Highly flexible metabolism of the marine euglenozoan protist *Diplonema papillatum*. BMC Biol 19**:** 251.

Tashyreva, D., G. Prokopchuk, J. Votýpka, A. Yabuki, A. Horák *et al.*, 2018 Life cycle, ultrastructure, and phylogeny of new diplonemids and their endosymbiotic bacteria. MBio 9**:** e02447-02417.

Triemer, R. E., and D. W. Ott, 1990 Ultrastructure of *Diplonema ambulator* Larsen & Patterson (Euglenozoa) and its relationship to *Isonema*. Eur J Protistol 25**:** 316-320.

Yu, F., S. E. Haynes, G. C. Teo, D. M. Avtonomov, D. A. Polasky *et al.*, 2020 Fast quantitative analysis of timsTOF PASEF data with MSFragger and IonQuant. Mol Cell Proteomics 19**:** 1575-1585.

13. Glycan and peptide assimilation by *Diplonema papillatum*

INTRODUCTION

The metabolism of *D. papillatum* has been examined experimentally regarding glycolysis and gluconeogenesis (MORALES *et al.* 2016), respiratory pathways (VALACH *et al.* 2018), carbon storage (ŠKODOVÁ-SVERÁKOVÁ *et al.* 2020), and adaptation to hypoxia (ŠKODOVÁ-SVERÁKOVÁ *et al.* 2021). A surprising result from this research was that *Diplonema* does not import glucose in any significant amount but instead takes up and preferentially metabolizes amino acids. This finding led to the conclusion that in its natural habitat, the primary energy source of this organism is not carbohydrates, as in the majority of heterotrophic eukaryotes, but rather poly- and oligo-peptides (MORALES *et al.* 2016; ŠKODOVÁ-SVERÁKOVÁ *et al.* 2020). The requirement of serum in the growth medium together with the prediction of numerous peptidase-encoding genes in the nuclear genome assembly (see the main text and **Supplementary Information**: Section 14. Secretome prediction), have corroborated this view.

The large ensemble of carbohydrate-metabolizing genes (CAZymes) detected in the inferred *Diplonema* proteome was, therefore, completely unexpected (see **Supplementary Information**: Section 12. CAZyme-coding genes in *Diplonema papillatum* and Section 14. Secretome prediction). About 70% of CAZyme genes carry domains that allow us to predict the likely substrates of these enzymes. A quarter of all CAZymes are homologs of enzymes degrading pectin, a heteropolysaccharide composed mainly of α–1,4-linked galacturonic acid (mostly methyl-esterified at its carboxyl groups) but also containing rhamnose, galactose, xylose, and arabinose units (CAFFALL AND MOHNEN 2009). Additional polysaccharides that *Diplonema* is presumably able to metabolize are cellulose (a β–1,4 glucan), hemi-celluloses (xylans, xylo-glucans), β–1,3 glucans (callose, laminarin, paramylon), glucuronomannans, mannans, and chitin, i.e., most cell-wall building blocks of diverse marine algae and plants.

Since many of the polysaccharide-degrading enzymes from *Diplonema* are predicted to be secreted outside the cell (see **Supplementary Information**: Section 14. Secretome prediction), we posit that this protist, in its natural habitat, uses its large enzymatic arsenal to break down the cell wall of microeukaryotic prey (for cell-wall composition, see (MYKLESTAD AND GRANUM 2009; SCHOLZ *et al.* 2014; LE COSTAOUËC *et al.* 2017; RAIMUNDO *et al.* 2017)). Such behaviour may serve different 'purposes': either to gain access to proteins inside the prey's cell or to exploit the prey's polycarbohydrates in the cell wall and intracellular storage (or both). For example, starch is a glycan stored by various marine plants and green algae (BALL *et al.* 2011), while laminarin and its variants are accumulated by diatoms, brown algae, and other photosynthetic microeukaryotes (CHABI *et al.* 2021).

We reasoned that if the only purpose of cell-wall degradation is to feed on the cytoplasm of the prey, then *D. papillatum* would not be able to use carbohydrates as the sole carbon source. To test this hypothesis, we performed growth experiments in media of various compositions, described below.

RESULTS

Nutrient assimilation experiments

D. papillatum was cultured in liquid medium supplemented with diverse carbon sources, including the cell-wall component pectin and the storage compound amylopectin; the latter is a polymer of α–1,4 and α–1,6 glucose units and constitutes the water-soluble component of starch. Compared to the control on medium without any carbon source, we observed that mono- and disaccharides, and polyols such as sorbitol, only poorly supported cell growth (**Supplementary Figure S20**). Cell counts were slightly higher in medium containing free amino acids, especially after eight days; we also noted that under these conditions, cells had large vacuoles. In contrast, when cultivated in the presence of serum, pectin or amylopectin, cell counts were ~2 times higher than of the control; **Supplementary Figure S20**). Cultures growing on tryptone and yeast extract achieved the highest titers (~3 to 4-fold increase; **Supplementary Figure S20**). The observed low cell proliferation on glycans compared to tryptone was likely due to the exhaustion of the internal nitrogen reserve that in *Diplonema* consists of free amino acids such as β-alanine and glutamate (MORALES *et al.* 2016; ŠKODOVÁ-SVERÁKOVÁ *et al.* 2021). We also observed that ammonium sulfate, which is added as a nitrogen source in synthetic media of numerous microbes (e.g., yeast), did not significantly influence the growth of *D. papillatum* (at 0.1% w/v) or even strongly inhibited it (at 0.5% w/v). Apparently, *Diplonema* is not able to use external ammonia as a nitrogen source.

Interestingly, *D. papillatum* cells grown on pectin and, particularly, amylopectin as a sole carbon source were significantly larger, and those from tryptone cultures considerably smaller, compared to cells cultivated in the standard serum-supplemented medium (**Supplementary Figure S21A–D**). However, the total biomass was quite similar because in tryptone-grown cultures, the number of cells was higher, whereas in pectin or amylopectin media, the cells had a lower titer (**Supplementary Figure S21E**).

The cytoplasm of large glycan-fed cells contained conspicuous granules readily visible by light microscopy (**Supplementary Figure S21**). The best candidates for the compound stored in these granules are (i) paramylon (a β-1,3-

glucan), which *Diplonema* can synthesize (ŠKODOVÁ-SVERÁKOVÁ *et al.* 2020), or alternatively (ii) lipids—known to be produced by many microeukaryotes and bacteria under nitrogen-limiting conditions (and considered for large-scale biotechnological production of biofuels (SUBRAMANIAM *et al.* 2010)). In contrast to glycans, lipids are readily detected by incubating cells with the Nile Red dye that stains specifically neutral lipids. Inspection of Nile Red-treated *D. papillatum* cells by fluorescence microscopy demonstrated that the granules formed by glycan-grown cells indeed represented lipid droplets (**Supplementary Figure S21F,G**).

Finally, a startling observation was that similar titers and cell sizes were obtained when *Diplonema* was cultivated on a mix of equal amounts of tryptone and a glycan (either pectin or amylopectin), compared with cultures on tryptone alone (**Supplementary Figure S20**). One possible explanation is that *Diplonema* prefers oligo-peptides over polysaccharides in situations where both substrates are readily available. It would be interesting to examine by metabolic labelling to what extent *D. papillatum* uses these substrates as an energy source or building blocks.

Expression of metabolic enzymes

Our RNA-Seq data indicated that essentially all genes encoding metabolic proteins produced transcripts. To test if these enzymes inferred from the *Diplonema* genome and transcriptome assemblies were indeed translated, we analyzed raw mass-spectrometry data, produced in another context, of total cellular proteins from *Diplonema* (ŠKODOVÁ-SVERÁKOVÁ *et al.* 2021). In that study, cells were cultivated in a standard medium (containing serum as a carbon and nitrogen source) or in a tryptone-supplemented medium, both in the presence or absence of oxygen. Our analysis shows that ~7% of all predicted CAZymes were expressed as proteins, which appears low, but can be explained by the moderate depth of the mass-spectrometry data (see **Supplementary Information**: Section 12. CAZyme-coding genes in *Diplonema papillatum*). More importantly, many CAZymes displayed similar levels under all tested conditions. Although polysaccharides were absent from the culture medium, we detected nine pectin-, seven cellulose-, and five glucuronomannan-degrading proteins, half of them at levels comparable to mitochondrial TCA-cycle enzymes (see **Supplementary Information**: Section 12. CAZyme-coding genes in *Diplonema papillatum*).

DISCUSSION

Our growth experiments with *D. papillatum* described above lead to two main conclusions. First, this organism prefers polymers (peptide- or carbohydrate-based) over their monomeric constituents, and second, it can utilize a carbohydrate-rich diet as efficiently as a peptide-rich one. Therefore, in its natural habitat, *Diplonema* most likely feeds on essentially all cellular components of its presumed eukaryotic prey, from cell wall to cytoplasm to storage glycans. The results obtained from the growth experiments are consistent with the large and diverse array of carbohydrate- and peptide-metabolizing enzymes inferred from the nuclear genome sequence, testifying to the extraordinary metabolic versatility of *Diplonema*.

β-1,3-glucans are abundant in marine habitats, serving as a cell-wall component (callose) in diatoms, haptophytes, and macroscopic brown algae, but also as a carbon storage material (with β–1,6 branching, laminarin) (RAIMUNDO *et al.* 2017; MYKLESTAD AND GRANUM 2009). One could argue that the presence of β-1,3-glucan-metabolizing enzymes in *Diplonema* simply reflects that paramylon is its carbon-storage compound rather than an ability to degrade the cell walls of prey (see **Supplementary Information**: Section 12. CAZyme-coding genes in *Diplonema papillatum*). However, multiple members of the corresponding CAZyme families from *D. papillatum* were predicted to be secreted outside the cell (see **Supplementary Information**: Section 14. Secretome prediction), which corroborates their role in digesting prey and in utilizing various β-1,3-glucans as an energy source.

Our finding that *D. papillatum* can feed on carbohydrates is at odds with previous studies, which reported that cells took up only negligible amounts of glucose (MORALES *et al.* 2016; ŠKODOVÁ-SVERÁKOVÁ *et al.* 2020; ŠKODOVÁ-SVERÁKOVÁ *et al.* 2021). In the corresponding experiments, this carbohydrate was added to the standard medium consisting of sea salts, 1% horse serum, and optionally 0.1% tryptone. The noted contradiction between the observed marginal glucose import and the presence of a gene encoding a putative sodium/glucose co-transporter (DIPPA_28570) led to the hypothesis that this transporter acted predominantly in intracellular transport (ŠKODOVÁ-SVERÁKOVÁ *et al.* 2021). While our growth experiments confirm that *Diplonema* does not readily feed on mono- and di-saccharides, we argue that the import rate of simple saccharides cannot be extrapolated to that of oligomers. Oligosaccharide importers are fairly frequent in many microbes, with some outer membrane transporters accommodating solutes of up to 2.5 kDa, corresponding to an atomic mass of 12-mers (FERREIRA *et al.* 2017; JECKELMANN AND ERNI 2020). Such transporters belong mainly to the ABC and MFS families. Since the genome of *D. papillatum* encodes ~100 members of each family (many carrying sugar-binding domains, see also **Supplementary Table S20**), this protist most likely imports carbohydrate oligomers. Lastly, we noted that the previous carbohydrate-assimilation experiments were conducted over a duration that might have been too short to detect the full extent of nutrient import (only 8 to 24h compared to multiple days in the study presented here). This aspect should be taken into account in the design of future experiments.

Another surprising finding was that *D. papillatum* can feed efficiently on amylopectin, although a gene encoding an amylase could not be detected in the genome assembly (see **Supplementary Information**: Section 12. CAZyme-coding genes in *Diplonema papillatum*). It is improbable that the amylopectin added to the growth medium was contaminated with proteins and/or lipids, because potato starch contains <0.1% proteins and lipids (per dry matter), and the amylopectin purified from starch contains even less of these compounds (SWINKELS 1985; DHITAL *et al.* 2011). In addition, such minute contaminants cannot explain the massive accumulation of carbon in the form of lipid droplets in *Diplonema* cells and the high biomass generated during growth on amylopectin. Therefore, we infer that *D. papillatum* can break down amylopectin, and that the genes encoding the corresponding enzymes are present in the genome but have not been recognized. The posited 'incognito' genes are likely among the 150 CAZymes of unknown substrate specificity detected in the *Diplonema* nuclear genome. Identifying these orphan genes might well unveil a new class of starch-degrading enzymes.

Finally, by analyzing protein mass spectrometry data, we observed that several CAZyme-encoding genes predicted by functional genome annotation of the *D. papillatum* nuclear genome sequence are highly expressed despite the absence of their glycan substrates in the culture medium. Constitutive expression or induced co-expression (e.g., during phagocytosis) of diverse nutrient-degrading genes would allow *Diplonema* to adapt rapidly to changes in the food landscape. In the same vein, *D. papillatum* also appears to barely modulate the expression of most enzymes upon changing oxygen levels in the medium (ŠKODOVÁ-SVERÁKOVÁ *et al.* 2021). While expressing genes that target substances absent at a given moment and place might seem wasteful, it is also an excellent strategy to be prepared for rapid shifts in marine environmental conditions. Thus, *D. papillatum* stands out for its highly versatile metabolic capability and its potential for rapid metabolic switching.

METHODS

Strain and regular culture conditions

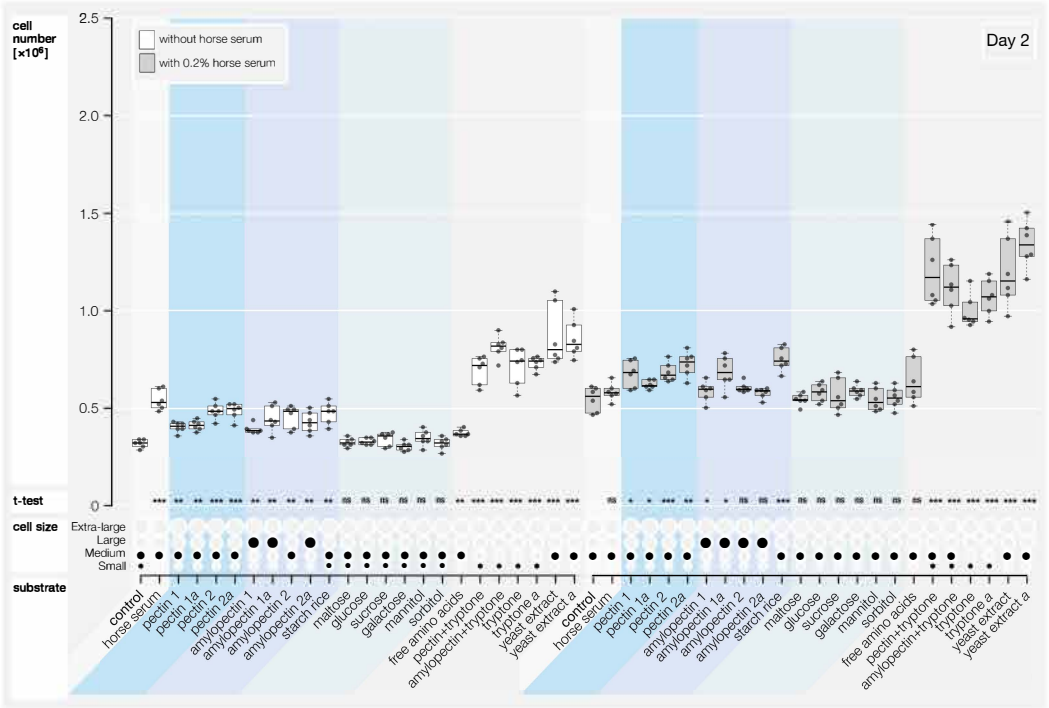
Diplonema papillatum (ATCC 50162) was originally obtained from the American Type Culture Collection (ATCC). As described earlier (VALACH *et al.* 2018), the strain was cultivated axenically without shaking at 15–22 °C in ocean salt medium (OS) containing 33 g/L Instant Ocean Sea Salt (*Instant Ocean*) supplemented with 1% (v/v) horse serum (*Wisent*). For extended cultivations, chloramphenicol (*Sigma-Aldrich*) was added to the medium at 40 mg/L to prevent bacterial contamination.

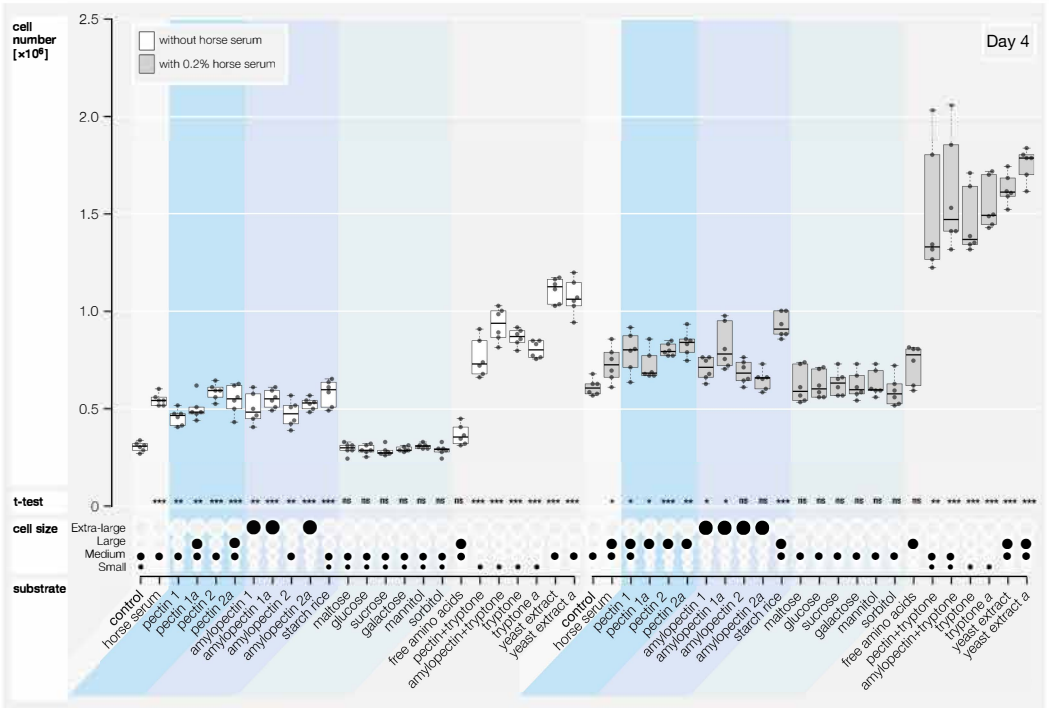
Nutrient assimilation experiments

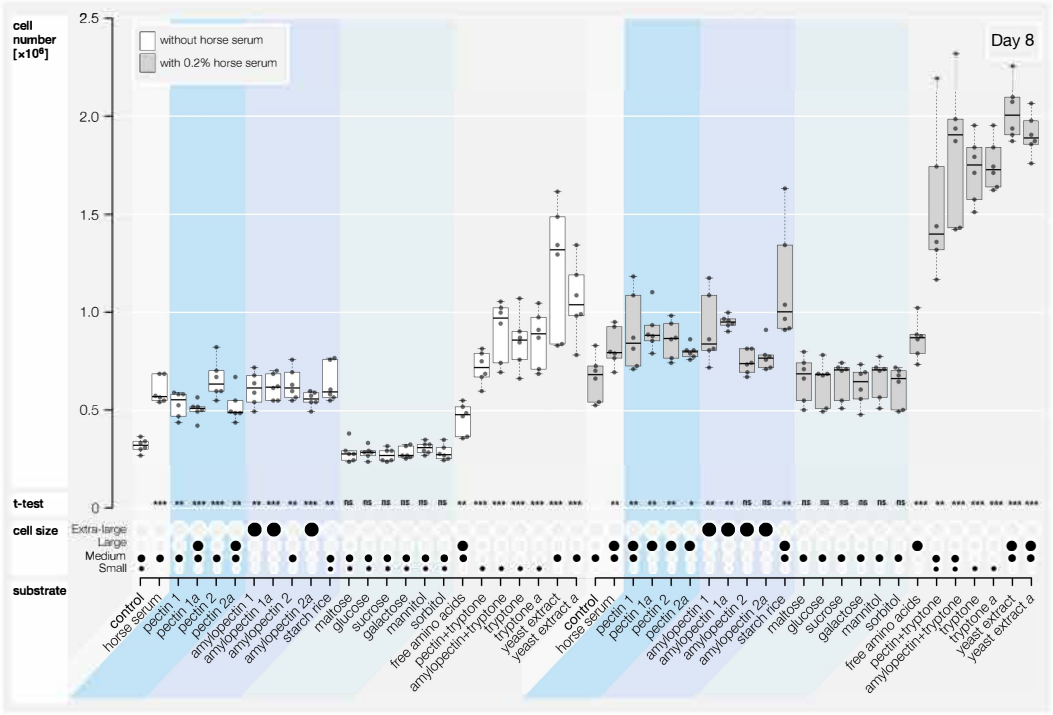
Most nutrient assimilation experiments were performed in multi-well plates, with each of the 48 wells holding 625 µL. Cells (~3×10^5 per well) from a preculture in the standard medium were extensively washed, then starved for 2–6 hr, and subsequently transferred to the new OS medium containing vitamins (biotin [1 µg/1 L], cyanocobalamin [2 µg/1 L], thiamine-HCl [20 µg/1 L]; *Thermo Scientific*), 20 mM HEPES-KOH, pH7.6 (*Bioshop Canada*), optionally horse serum to 0.2%. The following substrates served as primary energy sources: citrus pectin (purchased from two manufacturers: *TCI America* and *Pomona's Pectin/Green Link*); amylopectin (produced from waxy corn, *TCI America*, and from potato starch, *Fluka/Sigma-Aldrich*); mono- and disaccharides (glucose, galactose, maltose, saccharose; *Bioshop Canada*); sugar-derived poly-alcohols (polyols include mannitol, sorbitol; *Bioshop Canada*); yeast extract (a complex mixture of carbohydrates, proteins, ions, and cofactors; *BioBasic*); tryptone (trypsin-digested casein; *Difco/ThermoFisher*); and free amino acids (glycine, glutamate, glutamine; *Bioshop Canada*). Each substrate was added at the same weight-to-volume ratio (0.1% w/v) to provide a similar carbon content. Cells were monitored for up to eight days (~200 hr). Six biological replicates from three independent inocula were carried out. To determine the biomass, cells from the late exponential/early stationary phase were inoculated at a titer of ~5×10^5 per mL into 62.5 mL medium containing the substrate to be tested (horse serum, tryptone, pectin, or amylopectin). After 150–160 hr (~6–7 days), cells were counted under the microscope in triplicate, and the biomass was determined to calculate the wet weight per 10^8 cells. Four independent biological replicates were made.

Microscopy

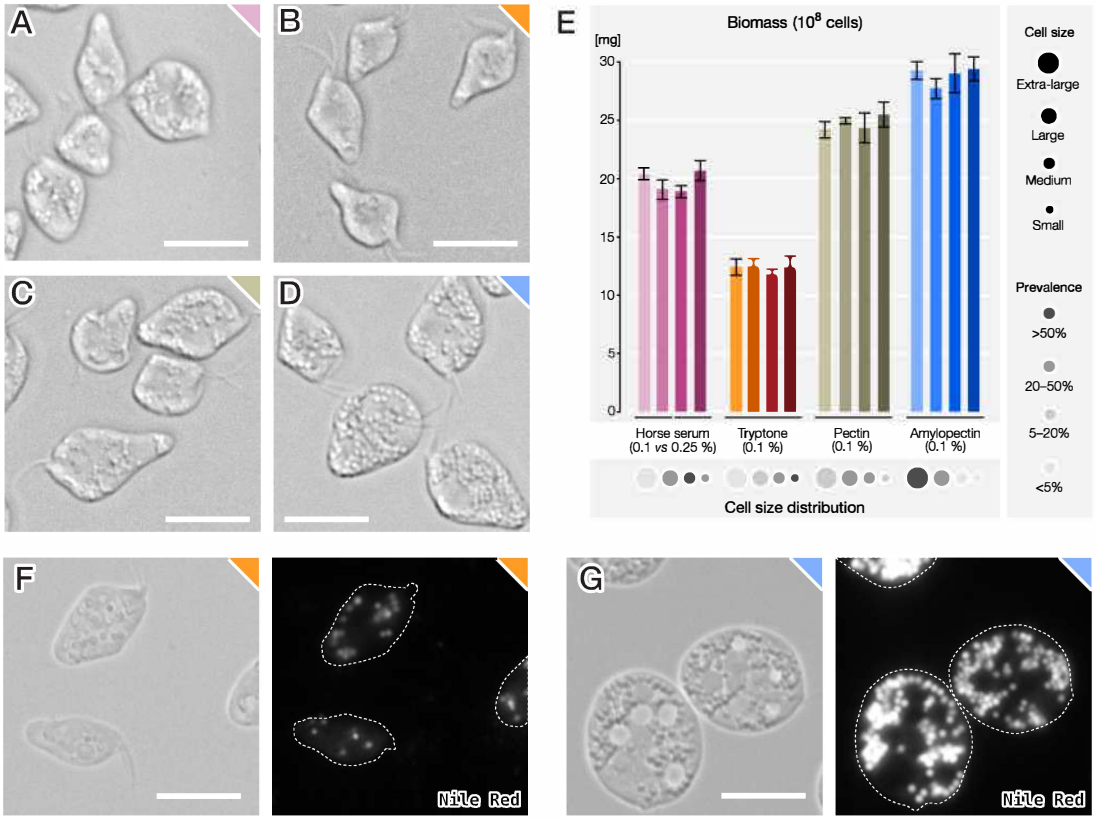
For conventional light microscopy, cells harvested from 6 to 7 day-old cultures were washed twice in plain OS medium (2,000×*g*, 3 min, 4 °C) and resuspended in the same medium or in an isotonic SoH buffer (1.2 M sorbitol, 20 mM HEPES, pH7.5). For the visualization of lipid droplets by fluorescence microscopy, cells were harvested after six days of cultivation, washed twice in plain OS medium, and then incubated (45 min, 22 °C) in 0.5× SoH buffer plus 16 g/L OS, to which Nile Red (AdipoRed staining reagent, *Lonza*) was added at a ratio of 1:20 (vol:vol). The hypotonicity of this buffer increased the assimilation of the dye and ensured a more even staining throughout the cell population. Cells, which stayed alive during the treatment, adhered more than usual to the slides. The slowed-down movement of the cells allowed visualization of sub-cellular structures at a higher resolution. Mounted cells were examined using an Eclipse Ts2R microscope (*Nikon*), and images were taken using a DS-Fi3 camera, analyzed by the NIS Elements BR software (*Nikon*), and post-processed with the Affinity Photo software v1.10.4 (*Serif*).


cell number [×10⁶]
without horse serum
with 0.2% horse serum
Day 2
t-test
cell size
Extra-large
Large
Medium
Small
substrate
control
horse serum
pectin 1
pectin 1a
pectin 2
pectin 2a
amylopectin 1
amylopectin 1a
amylopectin 2
amylopectin 2a
starch rice
maltose
glucose
sucrose
galactose
mannitol
sorbitol
free amino acids
pectin+tryptone
amylopectin+tryptone
tryptone
tryptone a
yeast extract
yeast extract a


cell number [×10⁶]
without horse serum
with 0.2% horse serum
Day 4
t-test
cell size
Extra-large
Large
Medium
Small
substrate
control
horse serum
pectin 1
pectin 1a
pectin 2
pectin 2a
amylopectin 1
amylopectin 1a
amylopectin 2
amylopectin 2a
starch rice
maltose
glucose
sucrose
galactose
mannitol
sorbitol
free amino acids
pectin+tryptone
amylopectin+tryptone
tryptone
tryptone a
yeast extract
yeast extract a



Supplementary Figure S20. **Growth of *D. papillatum* on various substrates.** Approximately 3×10^5 cells from mid-exponential phase were used to inoculate media containing the indicated substrates at 0.1% weight-to-volume (w/v) concentration. White and grey boxes indicate absence or presence of 0.2% horse serum, respectively, in the medium. The following substrates were tested: horse serum, pectin, and amylopectin, each from two different manufacturers, starch rice (~90% starch, ~10% proteins), maltose, glucose, sucrose, galactose, mannitol, sorbitol, free amino acids (glutamate, glutamine, glycine), a mixture of pectin or amylopectin with tryptone (1:1), tryptone, and yeast extract; *a*, autoclaved substrate. (For a detailed composition of the basic medium, which was also used for the 'control' sample, see the *Methods* section.) We made six biological replicates from three independent inocula. The graphs show cell counts after **2**, **4**, and **8** days post-inoculation. Circles of different diameters symbolize the observed cell sizes, with the predominant types indicated by the black fills. Growth rates were compared to the control sample. The two-tailed, paired Student's t-test probabilities are: *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$; ns, not significant). The boxplot whiskers extend to values below the 1.5-fold interquartile range beyond the 1st and 3rd quartile; the boxplot was generated using the BoxPlotR shiny application (Spitzer *et al.* 2014).



Supplementary Figure S21. ***D. papillatum*** **cell size and biomass depend on the substrate.** (**A–D**) Cells cultivated in a medium containing horse serum (**A**), tryptone (**B**), pectin (**C**), or amylopectin (**D**) as a sole carbon source were inspected by light microscopy. White bars correspond to 10 µm. (**E**) Biomass of *D. papillatum* cells grown in various substrates. Cells were counted in triplicate after six days and weighed to calculate their biomass (wet weight per 10^8 cells). Bars indicate the mean deviation of the cell counts for each of the four independent biological replicates. Note that the predominant cell size correlates with both biomass and the number of granules. (**F–G**) Granules, which accumulate especially during growth on glycans, are droplets of neutral lipids. Cells cultivated in a medium containing tryptone (**F**) or amylopectin (**G**) as a sole carbon source were stained with the Nile Red fluorescent reagent. Note that cells in **F–G** are rounder and larger than in **A–D**, because in the former, cells were incubated in a mildly hypotonic buffer to enhance staining. White bars correspond to 10 µm.

Supplementary Table S20. Sugar transporter genes identified in the predicted proteome of *D. papillatum*.

ID	Pfam Domain	Description	Transcripts Per Million (TPM) [a]	RNA-Seq Reads Mapped
DIPPA_16016.mRNA.1	PF00083	generic sugar transporters	3.023213	2406
DIPPA_24401.mRNA.1	PF00083	generic sugar transporters	1.472163	640
DIPPA_24417.mRNA.1	PF00083	generic sugar transporters	3.720192	509
DIPPA_33584.mRNA.1	PF00083	generic sugar transporters	1.558700	580
DIPPA_31431.mRNA.1	PF00083	generic sugar transporters	0.846433	426
DIPPA_31431.mRNA.2	PF00083	generic sugar transporters	0.921461	470
DIPPA_31439.mRNA.1	PF00083	generic sugar transporters	0.257623	274
DIPPA_11817.mRNA.1	PF00083	generic sugar transporters	2.829031	1601
DIPPA_27902.mRNA.1	PF00083	generic sugar transporters	0.958081	453
DIPPA_27956.mRNA.1	PF00083	generic sugar transporters	1.118480	526
DIPPA_27950.mRNA.1	PF00083	generic sugar transporters	4.544418	2141
DIPPA_02168.mRNA.1	PF00083	generic sugar transporters	0.396609	160
DIPPA_17708.mRNA.1	PF00083	generic sugar transporters	3.162026	1851
DIPPA_18224.mRNA.1	PF00083	generic sugar transporters	4.068439	1507
DIPPA_18217.mRNA.1	PF00083	generic sugar transporters	3.686166	1381
DIPPA_18239.mRNA.1	PF00083	generic sugar transporters	3.307667	1228
DIPPA_18361.mRNA.1	PF00083	generic sugar transporters	0.536121	273
DIPPA_18361.mRNA.2	PF00083	generic sugar transporters	0.561651	286
DIPPA_18361.mRNA.3	PF00083	generic sugar transporters	0.402806	170
DIPPA_18361.mRNA.4	PF00083	generic sugar transporters	0.543977	277
DIPPA_04043.mRNA.1	PF00083	generic sugar transporters	0.644033	256
DIPPA_04061.mRNA.1	PF00083	generic sugar transporters	0.915006	246
DIPPA_04061.mRNA.2	PF00083	generic sugar transporters	1.182485	429
DIPPA_34266.mRNA.1	PF00083	generic sugar transporters	2.955753	1335
DIPPA_35197.mRNA.1	PF00083	generic sugar transporters	3.277877	651
DIPPA_59166.mRNA.1	PF00083	generic sugar transporters	1.081599	440
DIPPA_62771.mRNA.1	PF00083	generic sugar transporters	0.942414	243
DIPPA_62771.mRNA.2	PF00083	generic sugar transporters	1.297940	461
TR56418__c0_g1_i1__m.7768	PF00083	generic sugar transporters	6.104713	5146
TR74098__c0_g1_i1__m.13842	PF00083	generic sugar transporters	3.513357	4601
TR116637__c0_g2_i1__m.27060	PF00083	generic sugar transporters	0.411792	74
TR121062__c0_g2_i1__m.28493	PF00083	generic sugar transporters	1.610895	1377
DIPPA_26420.mRNA.1	PF03083	sugar efflux transporters	2.915142	463
DIPPA_01763.mRNA.1	PF03083	sugar efflux transporters	3.803449	691
DIPPA_63739.mRNA.1	PF03083	sugar efflux transporters	1.190166	185
DIPPA_14432.mRNA.1	PF04142	nucleotide-sugar transporter	1.417889	344
DIPPA_16693.mRNA.1	PF04142	nucleotide-sugar transporter	0.786019	206
DIPPA_08985.mRNA.1	PF04142	nucleotide-sugar transporter	2.525701	694
DIPPA_26750.mRNA.1	PF04142	nucleotide-sugar transporter	16.764209	4351
DIPPA_01245.mRNA.1	PF04142	nucleotide-sugar transporter	5.568472	1318
DIPPA_18559.mRNA.1	PF04142	nucleotide-sugar transporter	2.487042	736
DIPPA_05515.mRNA.1	PF04142	nucleotide-sugar transporter	13.207931	3428
DIPPA_24052.mRNA.1	PF04142	nucleotide-sugar transporter	39.831333	8922
DIPPA_23546.mRNA.1	PF04142	nucleotide-sugar transporter	1.728252	431
DIPPA_07719.mRNA.1	PF04142	nucleotide-sugar transporter	0.615102	156
DIPPA_24216.mRNA.1	PF04142	nucleotide-sugar transporter	2.288946	689
DIPPA_09073.mRNA.1	PF04142	nucleotide-sugar transporter	2.096262	631
DIPPA_23121.mRNA.1	PF04142	nucleotide-sugar transporter	0.697580	135
DIPPA_59991.mRNA.1	PF04142	nucleotide-sugar transporter	34.402038	8492
DIPPA_65534.mRNA.1	PF04142	nucleotide-sugar transporter	9.882492	2498
TR66463__c0_g2_i1__m.11111	PF04142	nucleotide-sugar transporter	2.741823	1810
DIPPA_28795.mRNA.1	PF05631	12-TMH MFS-family sugar-transporters	0.978760	370
DIPPA_17988.mRNA.1	PF05631	12-TMH MFS-family sugar-transporters	2.002404	618
DIPPA_17988.mRNA.2	PF05631	12-TMH MFS-family sugar-transporters	2.006976	633

[a] For details on the calculation of transcript levels, see **Supplementary Information**: Section 2. Assembly and annotation of the nuclear genome and transcriptome of *Diplonema papillatum*.

AUTHOR CONTRIBUTIONS

Conceptualization – G.B., M.V.; **Data curation, Formal analysis, Investigation, Visualization, Writing, original draft** – M.V.; **Writing, review & editing** – all co-authors.

REFERENCES

Ball, S., C. Colleoni, U. Cenci, J. N. Raj and C. Tirtiaux, 2011 The evolution of glycogen and starch metabolism in eukaryotes gives molecular clues to understand the establishment of plastid endosymbiosis. J Exp Bot 62: 1775-1801.

Caffall, K. H., and D. Mohnen, 2009 The structure, function, and biosynthesis of plant cell wall pectic polysaccharides. Carbohydr Res 344: 1879-1900.

Chabi, M., M. Leleu, L. Fermont, M. Colpaert, C. Colleoni *et al.*, 2021 Retracing storage polysaccharide evolution in Stramenopila. Front Plant Sci 12: 629045.

Dhital, S., A. K. Shrestha, J. Hasjim and M. J. Gidley, 2011 Physicochemical and structural properties of maize and potato starches as a function of granule size. J Agric Food Chem 59: 10151-10161.

Ferreira, M. J., A. L. Mendes and I. de Sá-Nogueira, 2017 The MsmX ATPase plays a crucial role in pectin mobilization by *Bacillus subtilis*. PLoS One 12: e0189483.

Jeckelmann, J. M., and B. Erni, 2020 Transporters of glucose and other carbohydrates in bacteria. Pflugers Arch 472: 1129-1153.

Le Costaouëc, T., C. Unamunzaga, L. Mantecon and W. Helbert, 2017 New structural insights into the cell-wall polysaccharide of the diatom *Phaeodactylum tricornutum*. Algal Research 26: 172-179.

Morales, J., M. Hashimoto, T. A. Williams, H. Hirawake-Mogi, T. Makiuchi *et al.*, 2016 Differential remodelling of peroxisome function underpins the environmental and metabolic adaptability of diplonemids and kinetoplastids. Proc Biol Sci 283.

Myklestad, S. M., and E. Granum, 2009 Chapter 4.2 - Biology of (1,3)- β -glucans and related glucans in protozoans and chromistans, pp. 353-385 in *Chemistry, Biochemistry, and Biology of 1-3 Beta Glucans and Related Polysaccharides*, edited by A. Bacic, G. B. Fincher and B. A. Stone. Academic Press, San Diego.

Raimundo, S. C., S. Pattathil, S. Eberhard, M. G. Hahn and Z. A. Popper, 2017 β -1,3-Glucans are components of brown seaweed (Phaeophyceae) cell walls. Protoplasma 254: 997-1016.

Scholz, M. J., T. L. Weiss, R. E. Jinkerson, J. Jing, R. Roth *et al.*, 2014 Ultrastructure and composition of the *Nannochloropsis gaditana* cell wall. Eukaryot Cell 13: 1450-1464.

Škodová-Sveráková, I., G. Prokopchuk, P. Peña-Diaz, K. Záhonová, M. Moos *et al.*, 2020 Unique dynamics of paramylon storage in the marine euglenozoan *Diplonema papillatum*. Protist 171: 125717.

Škodová-Sveráková, I., K. Záhonová, V. Juricová, M. Danchenko, M. Moos *et al.*, 2021 Highly flexible metabolism of the marine euglenozoan protist *Diplonema papillatum*. BMC Biol 19: 251.

Spitzer, M., J. Wildenhain, J. Rappsilber and M. Tyers, 2014 BoxPlotR: a web tool for generation of box plots. Nat Methods 11: 121-122.

Subramaniam, R., S. Dufreche, M. Zappi and R. Bajpai, 2010 Microbial lipids from renewable resources: production and characterization. J Ind Microbiol Biotechnol 37: 1271-1287.

Swinkels, J. J. M., 1985 Composition and properties of commercial native starches. Starch - Stärke 37: 1-5.

Valach, M., A. Léveillé-Kunst, M. W. Gray and G. Burger, 2018 Respiratory chain Complex I of unparalleled divergence in diplonemids. J Biol Chem 293: 16043-16056.

14. Secretome prediction

INTRODUCTION

In unicellular eukaryotes, studying the secretome —the totality of proteins actively transported outside of the cell— provides insight into the organism's feeding behaviour and ecological role. Transport 'outside' of the cell means not only secretion into the environment. Certain microeukaryotic groups, such as ciliates and Discoba (including diplonemids), secret proteins also into their cytostome/cytopharynx, an invagination of the plasma membrane specialized for food ingestion. *Diplonema* is among the few microeukaryotes in which the ultrastructure of the feeding apparatus has been examined in great detail (TRIEMER AND OTT 1990; TRIEMER 1992; MONTEGUT-FELKNER AND TRIEMER 1996; TASHYREVA *et al.* 2018).

Most proteins translated in the cytosol follow the secretory pathway to reach their final destination; only those targeted to the nucleus, mitochondria, and peroxisomes take other distinct routes, while cytosolic proteins stay behind. The secretory pathway involves co-translational translocation of proteins into the endoplasmic reticulum (ER) and then transport into the Golgi complex (Golgi). According to the prevailing view, which is based on the 'bulk flow hypothesis', the transport outside the cell is the default track. In contrast, proteins that remain in the ER or the Golgi, or are destined to the vacuoles, carry a specific targeting code. Secreted proteins have a classical N-terminal signal peptide (SP) motif required for entering the secretory pathway; further, they lack an ER-retention signal, trans-membrane spanning domains (the latter include the Golgi-retention signal), and targeting motifs to the nucleus, mitochondrion, peroxisomes, and vacuoles (e.g., (GOGLEVA *et al.* 2018) and references therein).

Experimental secretome determination has recently made major technological advancements by introducing 'spatial proteomics', reducing contaminations with non-secreted proteins, e.g., due to cell injury or apoptosis (reviewed in IMAI AND NAKAI 2020). However, the limitation remains that the experimental results depend on the culture conditions and physiological and developmental stage of the organism, and thus give only a partial picture. The alternative to experimental determination is *in silico* secretome predictions. Available tools use not only the presence/absence of targeting or retention motifs detected by weight matrices or Hidden Markov models (HMMs), but also sequence alignments and general physicochemical properties for a variety of machine-learning algorithms (IMAI AND NAKAI 2020). Yet, predicted secretomes should be also taken with a grain of salt. While most tools have comparable accuracy (CHOO *et al.* 2009), the particular subset of a given proteome predicted by the different tools differ considerably, because the various machine learning algorithms have been trained with different datasets.

In the analysis presented here, the prediction of signal peptide (SP)-carrying proteins was performed with *Phobius* (SONNHAMMER *et al.* 1998). This tool has the advantage over other software such as *SignalP* (PETERSEN *et al.* 2011) and *TargetP* (EMANUELSSON *et al.* 2000) that it infers simultaneously SPs and transmembrane domains (TMDs), thus reducing the common problem of confusing certain TMDs at the N-terminus of proteins with SPs (SONNHAMMER *et al.* 1998).

RESULTS and DISCUSSION

About 10% of the total inferred proteome from *D. papillatum* was predicted to be secreted outside of the cell (**Supplementary Table S21**), which makes the size of the theoretical *Diplonema* secretome comparable with that from other free-living, heterotrophic microeukaryotes: the kinetoplastid *Bodo saltans* (6%; 1,187 out of 18,963 proteins (POWELL *et al.* 2016)); and, as a phylogenetically distant comparison point, fungi (2–8% (LUM AND MIN 2011)). It should be noted that the secretome sizes listed above are only tentatively comparable because different software tools and parameters have been used in the predictions.

To obtain insight into the feeding strategy of *D. papillatum*, we examined the number of inferred proteins with functions in lipid, protein, and carbohydrate degradation, and the percentage of these enzymes predicted to be secreted outside the cell (**Supplementary Table S22**). Lipases and lipoxygenases were identified based on the EC number assigned to them by the automated function-annotation pipeline (see **Supplementary Information**: Section 2. Assembly and annotation of the nuclear genome and transcriptome of *Diplonema papillatum*), whereas proteolytic enzymes were identified by searches against the peptidase database MEROPS, and carbohydrate-degrading enzymes were determined by the procedure used for the daily update of the CAZyme database (CANTAREL *et al.* 2009; LOMBARD *et al.* 2014; DRULA *et al.* 2022), followed by expert validation (see also Supplementary Information on CAZymes). The corresponding enzyme counts in the **Supplementary Table S22** are most likely underestimated because it is sometimes difficult to infer, based on sequence information alone, the exact substrates.

Since the three enzyme classes were identified in very different ways, the class sizes are not comparable to each other, but what is interpretable and relevant is the portion of non-secreted versus secreted proteins. In this latter comparison,

CAZymes stand out, with about 7% being part of the secretome. Among the secreted CAZymes, enzymes that degrade carbohydrates make up as much as 77%. The largest group within these secreted catabolic CAZymes are enzymes that break down pectin. Interestingly, secreted and non-secreted pectinases include the same CAZyme families, which indicates that *D. papillatum* is capable not only of intracellular but also of extracellular pectin degradation. As discussed more extensively in the section on feeding strategy, this finding corroborates the view that by secreting cell-wall and plasma-membrane-degrading enzymes, *Diplonema* is capable of feeding not only on microeukaryotes but also multicellular eukaryotes, i.e., prey that is much larger than it can engulf in the cytopharynx.

Supplementary Table S21. Summary of secretome prediction

Category	Nr. of proteins [a]	% of proteome)
Proteome	42,423	/
Signal-peptide bearing	5,160	12.2%
Secretome [b]	4,467	10.5%

[a] Count includes isoforms.
[b] Signal-peptide bearing proteins minus those with ER-retention signaled.

Supplementary Table S22. Secreted lipid, protein, and carbohydrate-degrading enzymes in *D. papillatum*

Enzyme class	Total nr. of proteins	% of proteome [a]	Secreted proteins in enzyme class		
			Nr.	% of secretome [a]	% of secreted CAZome
Lipases and lipoxygenases [b]	98	0.2%	21	0.5%	/
Proteases [c]	618	1.5%	178	4.0%	/
CAZymes [d]	527	1.2%	306	6.9%	/
Degraders[e]	369	0.9%	234	5.2%	76.5%
Pectin de-graders[f]	121	0.3%	78	1.7%	25.5%

[a] For the total number of proteins in the inferred proteome and secretome, see **Supplementary Table S21**.
[b] Enzymes were identified by the EC number assigned to the proteins by the function-annotation pipeline (see **Supplementary Information**: Section 2. Assembly and annotation of the nuclear genome and transcriptome of *Diplonema papillatum*). Lipases (EC 3.1.1.1, 2, 4, 7, 32, 47, 24) include carboxylesterases, triacylglycerol lipases, phospholipases A(1) and (2), acetylcholinesterases, and 1-alkyl-2-acetylglycerophosphocholine esterases. Lipoxygenases (EC 1.13.11.12, 75; seven proteins) include linoleate 13S-lipoxygenases and all-trans-8'-apo-beta-carotenal 15,15'-oxygenases.
[c] Enzymes were identified by a search against peptidase subsequences of the Merops peptidase database. From the corresponding peptunit.lib library file, we removed those units that correspond to peptidase inhibitors, transposable-element ORFs or proteasome components; see Methods.
[d] For CAZyme identification, see **Supplementary Information**: Section 12. CAZyme-coding genes in *Diplonema papillatum*.
[e] Proteins belong to the CAZyme classes glycosid hydrolases (GH), carboxyhydrate esterases (CE), and polysaccharide lyases (PL); see **Supplementary Information**: Section 12. CAZyme-coding genes in *Diplonema papillatum*.
[f] Proteins belonging to the CAZyme classes GH28, GH53, GH54, GH78, GH145, CE8, CE13, PL1, and PL4, see **Supplementary Information**: Section 12. CAZyme-coding genes in *Diplonema papillatum*.

METHODS

Secretome prediction

The prediction of signal peptide (SP)-carrying proteins was performed with *Phobius* 1.01 (SONNHAMMER *et al.* 1998). From the predicted SP-carrying proteins, we removed those carrying an ER-retention signal that we recognized by the presence of the typical motif `'[KRHQSA][DENQ]EL'`. The resulting protein collections represents the predicted secretome.

Prediction of lipid, protein and carbohydrate-degrading enzymes

Proteins involved in **lipid degradation** were identified in the inferred *Diplonema* proteome based on their EC number (https://enzyme.expasy.org) assigned by the automated annotation pipeline. For carboxylic-ester hydrolases (CEHs), we selected EC 3.1.1.1 (carboxyesterases), 3.1.1.3 (triacylglycerol lipases), 3.1.1.4 (phospholipases A(2)), 3.1.1.7 (acetylcholinesterases), 3.1.1.32 (phospholipases A(1)), and 3.1.1.47 (1-alkyl-2-acetylglycerophosphocholine esterases), combined with lipoxygenases EC 1.13.11.12 (linoleate 13S-lipoxygenases) and EC 1.13.75 (all-trans-8'-apo-beta-carotenal

15,15'-oxygenases). **Proteases** were predicted by a *BLASTP* search of the *D. papillatum* proteome (reporting threshold 1.0e-05) against peptidase units (pepunit.lib) downloaded from the MEROPS database release 12.4 (RAWLINGS *et al.* 2017). To obtain the list of proteases involved in nutrient digestions, we removed from pepunit.lib peptidase inhibitors, proteasome subunits, and peptidases associated with transposable elements. These sequences were recognized by the following terms in the header line: 'inhibitor unit', 'retrotransposon', 'virus', and 'ubiquitin'. By that, the search space of pepunit.lib was reduced by 15%, from 1,221,971 to 1,046,870 sequences. We also 'cleaned' the blastp output, by removing hits against *Diplonema* annotated as retrotransposon ORFs or containing a reverse-transcriptase domain, reducing the number of matches by 4%. Enzymes involved in the overall carbohydrate metabolism (**CAZymes**) were predicted by integrating gapped *BLASTP* and HMM searches against enzymes with experimentally confirmed function compiled in a 'high-quality library' at the Carbohydrate-Active Enzymes database (CAZy) (CANTAREL *et al.* 2009), followed by expert validation (see also **Supplementary Information**: Section 12. CAZyme-coding genes in *Diplonema papillatum*). The *carbohydrate-degrading* CAZyme subset ('Degraders' in the **Supplementary Table S22**) are those belonging to classes GH (glycosyl hydrolases), carboxyhydrate esterases (CE), or polysaccharide lyases (PL). 'Pectin degraders' are an enzyme subset composed of the classes GH28, GH53, GH54, GH78, GH145, CE8, CE13, PL1, and PL4; see main text and **Additional File 4** cazymeList.

AUTHOR CONTRIBUTIONS

Conceptualization, Data curation, Investigation, Formal analysis, Writing, original draft – G.B.; **Writing, review & editing –** all co-authors.

REFERENCES

Cantarel, B. L., P. M. Coutinho, C. Rancurel, T. Bernard, V. Lombard *et al.*, 2009 The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Res 37**:** D233-238.

Choo, K. H., T. W. Tan and S. Ranganathan, 2009 A comprehensive assessment of N-terminal signal peptides prediction methods. BMC Bioinformatics 10 Suppl 15**:** S2.

Drula, E., M. L. Garron, S. Dogan, V. Lombard, B. Henrissat *et al.*, 2022 The carbohydrate-active enzyme database: functions and literature. Nucleic Acids Res 50**:** D571-d577.

Emanuelsson, O., H. Nielsen, S. Brunak and G. von Heijne, 2000 Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 300**:** 1005-1016.

Gogleva, A., H. G. Drost and S. Schornack, 2018 SecretSanta: flexible pipelines for functional secretome prediction. Bioinformatics 34**:** 2295-2296.

Imai, K., and K. Nakai, 2020 Tools for the recognition of sorting signals and the prediction of subcellular localization of proteins from their amino acid sequences. Front Genet 11**:** 607812.

Lombard, V., H. Golaconda Ramulu, E. Drula, P. M. Coutinho and B. Henrissat, 2014 The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res 42**:** D490-495.

Lum, G., and X. J. Min, 2011 FunSecKB: the Fungal Secretome KnowledgeBase. Database (Oxford) 2011**:** bar001.

Montegut-Felkner, A. E., and R. E. Triemer, 1996 Phylogeny of *Diplonema ambulator* (Larsen and Patterson): 2. Homologies of the feeding apparatus. European Journal of Protistology 32**:** 64-76.

Petersen, T. N., S. Brunak, G. von Heijne and H. Nielsen, 2011 SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8**:** 785-786.

Powell, B., V. Amerishetty, J. Meinken, G. Knott, F. Yu *et al.*, 2016 ProtSecKB, pp.

Rawlings, N. D., A. J. Barrett, P. D. Thomas, X. Huang, A. Bateman *et al.*, 2017 The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. Nucleic Acids Res 46**:** D624-D632.

Sonnhammer, E. L., G. von Heijne and A. Krogh, 1998 A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol 6**:** 175-182.

Tashyreva, D., G. Prokopchuk, J. Votýpka, A. Yabuki, A. Horák *et al.*, 2018 Life cycle, ultrastructure, and phylogeny of new diplonemids and their endosymbiotic bacteria. MBio 9**:** e02447-02417.

Triemer, R. E., 1992 Ultrastructure of mitosis in *Diplonema ambulato*r Larsen and Patterson (Euglenozoa). Eur J Protistol 28**:** 398-404.

Triemer, R. E., and D. W. Ott, 1990 Ultrastructure of *Diplonema ambulator* Larsen & Patterson (Euglenozoa) and its relationship to *Isonema*. Eur J Protistol 25**:** 316-320.

15. Genes horizontally transferred from bacteria to *Diplonema papillatum*

INTRODUCTION

Horizontal (or lateral) gene transfer (HGT) between species, as opposed to the vertical descent of genes from parent to offspring, is a well-known and frequent phenomenon in prokaryotes. Only relatively recent methodological advances and careful experimental design provide convincing support of HGT and its critical contribution to the evolution of eukaryotes (ALSMARK *et al.* 2013; HUSNIK AND MCCUTCHEON 2018; VAN ETTEN AND BHATTACHARYA 2020). For instance, HGT from bacteria to eukaryotes has considerably expanded the capacities of certain eukaryotes to survive in hypoxic or anoxic environments (STAIRS *et al.* 2018; STAIRS *et al.* 2020). Several mutually non-exclusive mechanisms have been proposed for the horizontal transfer of genetic material across domains of life and from bacteria to eukaryotes in particular (SIBBALD *et al.* 2020). For diplonemids, which are free-living marine organisms, at least three pathways for acquiring bacterial genes are plausible; first *via* endosymbionts, since various diplonemids are known to harbour intracellular bacteria (TASHYREVA *et al.* 2018; PROKOPCHUK *et al.* 2019; GEORGE *et al.* 2020), second *via* food, because certain diplonemid species prey on bacteria (PROKOPCHUK *et al.* 2022), and third *via* marine viruses that have been shown to act as predominant HGT vectors in algae (e.g., NELSON *et al.* 2021) and to infect kinetoplastids, the sister clade of diplonemids (KOSTYGOV *et al.* 2021; IRWIN *et al.* 2022).

RESULTS

The identification of protein-coding genes potentially transferred horizontally from bacteria to *D. papillatum* consisted of three steps. First, we collected *D. papillatum* proteins of potential bacterial origin by reciprocal BLAST against NCBI's nr database. Then, we constructed phylogenetic trees for each candidate and its best hits against a taxonomically representative reference proteome database (see Methods) and selected *Diplonema* proteins nested with high support within bacterial clades. Expert validation included visual tree inspection and verification that the HGT candidate gene resides on a contig together with unambiguously nuclear genes.

The *D. papillatum* genome assembly contains 96 genes that most likely originate from bacteria *via* HGT. Phylogenetic analyses (not shown) did not point to any particular bacterial group as a preferential source but indicated a large number of donor clades. At least 90% of HGT genes in *D. papillatum* appear to be functional because they are transcribed, and their pre-mRNAs contain a 5′ spliced leader (SL), as do transcripts from ordinary genes. All genes discussed in more detail below adhere to these criteria.

HGT genes of *D. papillatum* can be grouped into 56 gene families (**Additional File 5** horizTransfer). Ten families comprise multiple members with ~15%–100% protein-sequence identity, while the remaining families consist of singletons. Most multi-member HGT families, especially those with >70% identity across all members, likely arose by amplification of the founder gene after its transfer to the new host. Half of the acquisitions took place specifically in the *D. papillatum* lineage and a quarter in the last common diplonemid ancestor, while the timing of the remaining events is uncertain.

Three-quarters of HGT families in *D. papillatum* have been assigned functions. Most (26 families, 41 proteins) are involved in various metabolic pathways, followed by transport (4 families, 5 proteins), detoxification of reactive oxygen species (3 families, 21 proteins), nucleic acids processing (3 families, 8 proteins) and regulatory functions and signalling (2 families, 2 proteins).

The largest HGT family in the *D. papillatum* genome assembly comprises 14 members of a gene encoding a PAP2-like superfamily protein, with top BLAST hits in the GenBank-nr database annotated as vanadium-dependent haloperoxidases but no significant matches among UniProtKB reviewed sequences. Most of the PAP2-family members are tandemly arranged in the *Diplonema* genome, with up to five genes clustering together (all expert-validated). Another large HGT family, the catalase group, forms a tandem repeat of six almost identical copies (98.6–100% protein-sequence identity; assembly validated by long-read mapping). Phylogenetic analyses of catalase sequences across euglenozoans suggest an alpha-proteobacteria ancestry of this gene and transfer specifically to *D. papillatum* since all other diplonemids possess a eukaryotic-type enzyme (ŠKODOVÁ-SVERÁKOVÁ *et al.* 2020). Catalase genes are known to be frequently transferred across the domains of life (FAGUY AND DOOLITTLE 2000; KRAEVA *et al.* 2017).

Since *D. papillatum* has the potential to degrade diverse polysaccharides, we analyzed in detail the four HGT families (17 genes in total) that are involved in carbohydrate breakdown (see also **Supplementary Information**: Section 12. CAZyme-coding genes in *Diplonema papillatum* and Section 14. Secretome prediction). The corresponding genes appear to be specific to *D. papillatum* to the exclusion of the other ten diplonemid taxa (although only transcriptome data are available for these latter species, the absence of detectable expression strongly indicates the lack of these genes).

With 12 members, the largest of the four HGT-CAZyme families encodes a xylan α-glucuronidase of the glycoside hydrolase family GH115. Ten proteins are full-length (DIPPA_15749 being a prototypical member of this family), whereas DIPPA_07185 and DIPPA_08958 are truncated versions representing an N-terminal and C-terminal moiety, respectively. Half of the genes are arranged in tandem pairs, and both pairs and singles are dispersed across the contigs (all verified by long-read mapping). Protein-sequence identity between family members is above 80%, and DIPPA_15749 and DIPPA_15740 (a tandem pair) are entirely identical. Phylogeny places *Diplonema*'s HGT-GH115-family proteins as a sister clade to sequences from Bacteroidetes, Planctomycetes, Gamma-proteobacteria and Verrucomicrobia (**Supplementary Figure S22**).

The *D. papillatum* genome assembly includes an additional GH115 domain-containing gene, DIPPA_19609, which is not affiliated with bacteria (and therefore was not retrieved during the HGT-detection procedure employed here), but rather with a clade that includes fungi and other eukaryotes (**Supplementary Figure S22**). Whatever the ancestral origin of the corresponding genes may be, they appear to have a long history of residence in eukaryotes. GH115 domain-bearing proteins were also detected in five additional diplonemids to the exclusion of *D. papillatum*; these form yet another clade together with bacterial sequences (**Supplementary Figure S22**).

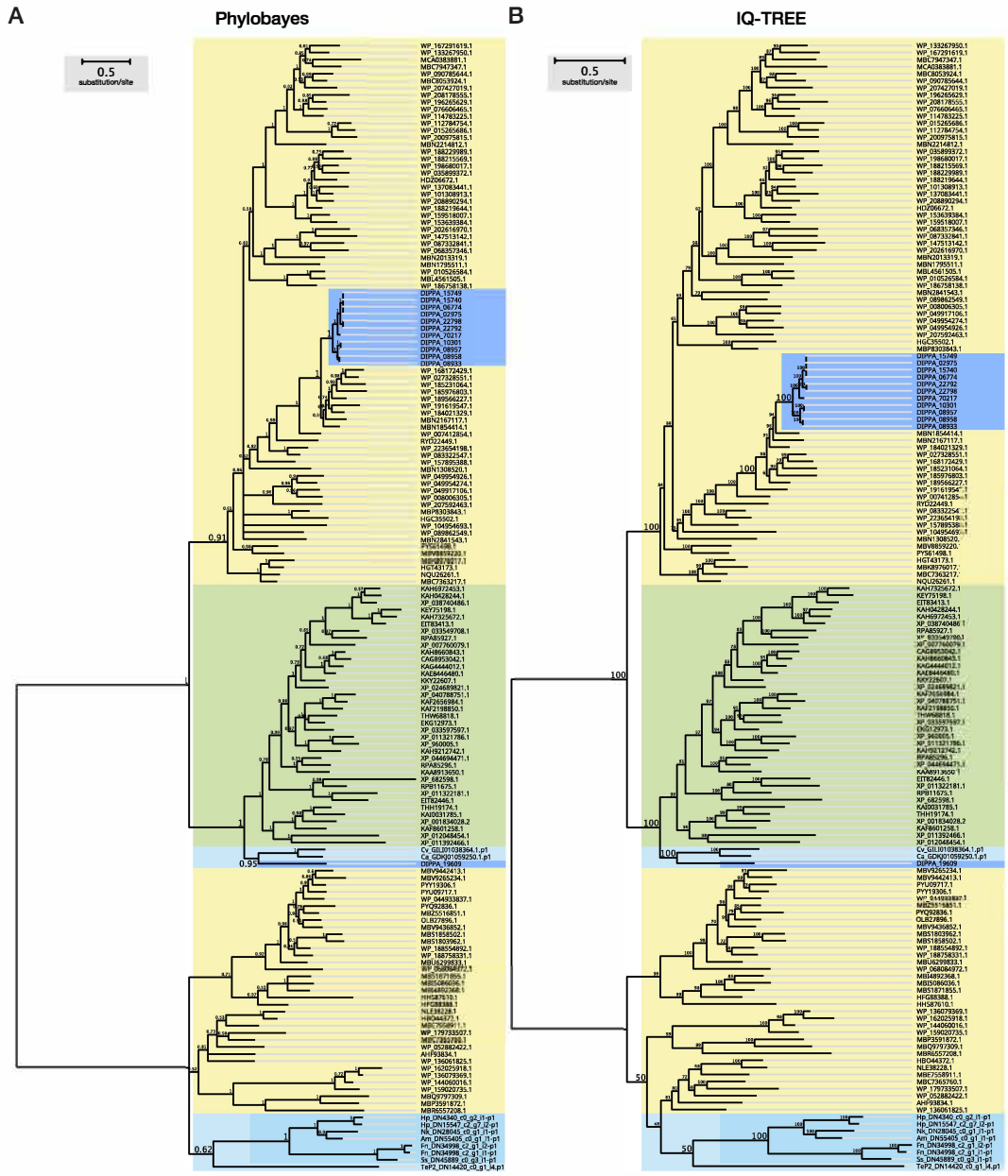
A second HGT family specific to *D. papillatum* and involved in glycan metabolism comprises proteins characterized by the carbohydrate-binding module CBM6. The three family members share ~99% protein-sequence identity and are arranged in tandem in the genome (assembly verified by long-read mapping). Top BLAST hits in the GenBank-nr database are annotated as endo-1,4-β-xylanase from Delta- and Alpha-proteobacteria. However, the query coverage is low, including only two ~100 amino acid-long regions in the N-terminal and C-terminal CBM6 domains of the *Diplonema* proteins; the ~300 amino acids-long central region has only spurious matches. Therefore, the protein is probably not an enzymatically active xylanase. It should be noted that CBM6 domains occur also in other *D. papillatum* proteins, either as tandem repeats or attached to GH128 domains of β-1,3-glucanases (see also **Supplementary Information**: Section 12. CAZyme-coding genes in *Diplonema papillatum*).

A third type of carbohydrate-degrading genes that *D. papillatum* acquired apparently by HGT is DIPPA_18298, which contains CBM10 and GH26 domains. BLAST searches against a database of experimentally characterized CAZymes and CAZyme domains showed that the GH26 catalytic domain of this protein is >60% identical to the β-1,4-mannanase from *Cellvibrio japonicus* (HOGG *et al.* 2003) and the CBM10 domain resembled that of validated β-mannanases (YOON 2016). This strongly suggests that the *Diplonema* protein most likely targets β-1,4-mannans, which—in addition to glucose and optionally xylose and mannose units—are building blocks of hemi-celluloses.

Lastly, the HGT candidate DIPPA_19532 is predicted to be a D-arabinitol 4-dehydrogenase, an enzyme involved in xylulose reduction (which is part of mannose and fructose metabolic pathways and pentose interconversions). This *D. papillatum* protein is closely related to sequences from several Alpha-proteobacteria.

DISCUSSION

D. papillatum readily expresses at least a hundred genes that originate from bacteria and were acquired recently by either the ancestor of diplonemids or *D. papillatum* itself. The number of HGT genes is similar to that observed across diverse eukaryotic genomes, and the same applies to the nature of the genes: most bacteria-to-eukaryote transferred genes expand or rewire the metabolic capabilities of the recipient (reviewed in (HUSNIK AND MCCUTCHEON 2018; VAN ETTEN AND BHATTACHARYA 2020)). Genes encoding CAZymes have been observed among the most frequently acquired and expanded functional categories across eukaryotes from ciliates to nematodes and oomycetes (RICARD *et al.* 2006; DANCHIN *et al.* 2010; RICHARDS *et al.* 2011; ALSMARK *et al.* 2013). The four CAZyme families specific to *D. papillatum* and obtained from bacteria are most likely involved in the degradation of various hemi-celluloses, mannans, and their derivatives. The acquisition of such enzymes has, in all likelihood, broadened the capacity of *D. papillatum* to feed on the most diverse algae and plants in its natural marine habitat (PORTER 1973).



Supplementary Figure S22. Phylogeny of GH115 domain-containing proteins. The phylogenetic tree was constructed with Bayesian (**A**) and maximum likelihood (**B**) methods using the sequences of GH115 domains from a wide collection of proteins (see Methods). Posterior probabilities (Phylobayes) and UF bootstrap support values (IQ-TREE) are indicated next to each branch. Both approaches resolve the tree into several clusters containing bacterial (yellow) or eukaryotic sequences (fungi in green, other eukaryotes in light blue, diplonemids in blue). Sequences specifically from *D. papillatum* are highlighted in darker background shades. Note that DIPPA_07185 was not included in the tree because it lacks the GH115 domain (see text for details).

METHODS

The procedure to detect genes potentially transferred horizontally from bacteria to *D. papillatum* included the following steps. First, HGT candidates were predicted with the taxonomic-annotation algorithm of Diamond v2.0 with default parameters (BUCHFINK *et al.* 2015) using the GenBank-nr database as a reference dataset. This step resulted in 1,317 *D. papillatum* proteins, for which the best hit was a bacterial sequence. These candidates were then queried, using Diamond with default parameters, against the GenBank-nr database and a custom reference dataset comprising proteomes of 408 organisms, notably 148 bacteria, 146 archaea, and 114 eukaryotes, including *D. papillatum* and 17 other euglenozoans (**Additional File 5** horizTransfer). For each of the 1,317 *D. papillatum* proteins, we selected all the hits above the default e-value threshold (0.0001) from the reference database and up to 50 of the top hits in GenBank-nr. All *D. papillatum* sequences that produced reciprocal Diamond hits (e=0.0001) against a *D. papillatum* database were combined, resulting in 370 distinct protein groups. The *D. papillatum* protein groups together with their corresponding hits in the GenBank-nr database and the reference dataset, were aligned with MAFFT v7.4 (KATOH AND STANDLEY 2013) using default parameters. Each multiple alignment was trimmed using BMGE v1.12 with default parameters (CRISCUOLO AND GRIBALDO 2010) and used to build a phylogenetic tree using IQ-TREE v1.6.12 (NGUYEN *et al.* 2015) with the LG+G+F model, -bnni option and 1,000 bootstrap replicates. To detect a HGT signal in the phylogenetic trees, we applied the following algorithm implemented as a custom Python script (available at https://github.com/AnnaNenarokova/ngs/blob/master/projects/hgt/dpapi/check_hgt_trees.py). Branches with <70% bootstrap support were collapsed. If a branch consisted exclusively of one or more diplonemid sequences with two closest neighbouring branches comprising solely bacterial sequences, the diplonemid sequences were considered a result of a singular HGT from bacteria. The 64 trees that displayed the above-described pattern and the underlying multiple protein alignments were inspected in detail to eliminate errors. For example, five *D. papillatum* protein-coding genes, located on two small contigs, were identified as bacterial contamination (see **Supplementary Information**: Section 2. Assembly and annotation of the nuclear genome and transcriptome of *Diplonema papillatum*). Further, proteins in three other trees were considered false positives because the underlying alignment to bacterial sequences relied on marginal similarity. The remaining proteins were further queried against our database of transcriptome-inferred diplonemid proteomes (VALACH *et al.* 2017; BUTENKO *et al.* 2020; KAUR *et al.* 2020) to assess whether the HGT event occurred before or after the emergence of *D. papillatum*.

To construct a phylogeny tree of GH115 domain-containing proteins, we performed BLAST searches using DIPPA_15749 and DIPPA_19609 protein sequences against the full GenBank-nr database, and Diamond searches against a local copy of the EukProt compilation (https://github.com/beaplab/EukProt), as well as against our collection of diplonemid proteomes. The 100 top Blast hits against GenBank-nr were clustered via CD-HIT (FU *et al.* 2012) at a 70% identity threshold, and only the cluster-representative sequences were kept. Sequences were pre-aligned with Muscle v3.8.1551 (EDGAR 2004) and realigned with an HMM search (the profile HMM was built based on the initial Muscle alignment) using hmmsearch (HMMER v3.3) (EDDY 2011). Only those columns of the multiple protein alignment that aligned with a posterior probability of 1.0 were retained for the phylogenetic analysis. For tree construction, we used PhyloBayes v4.1b (LARTILLOT *et al.* 2013) by running four independent chains, six gamma categories and the CAT-GTR model. In addition, IQ-TREE v2.1.3 was used with default parameters and the option to calculate 1,000 ultrafast bootstrap replicates (MINH *et al.* 2020).

AUTHOR CONTRIBUTIONS

Conceptualization, Formal analysis – A.N.; **Data curation, Investigation, Writing, original draft** – A.N., M.V.; **Visualization** – B.F.L., M.V.; **Writing, review & editing** – all co-authors.

REFERENCES

Alsmark, C., P. G. Foster, T. Sicheritz-Ponten, S. Nakjang, T. Martin Embley *et al.*, 2013 Patterns of prokaryotic lateral gene transfers affecting parasitic microbial eukaryotes. Genome Biol 14**:** R19.

Buchfink, B., C. Xie and D. H. Huson, 2015 Fast and sensitive protein alignment using DIAMOND. Nat Methods 12**:** 59-60.

Butenko, A., F. R. Opperdoes, O. Flegontova, A. Horák, V. Hampl *et al.*, 2020 Evolution of metabolic capabilities and molecular features of diplonemids, kinetoplastids, and euglenids. BMC Biol 18**:** 23.

Criscuolo, A., and S. Gribaldo, 2010 BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol 10: 210.
Danchin, E. G., M. N. Rosso, P. Vieira, J. de Almeida-Engler, P. M. Coutinho *et al.*, 2010 Multiple lateral gene transfers and duplications have promoted plant parasitism ability in nematodes. Proc Natl Acad Sci U S A 107: 17651-17656.
Eddy, S. R., 2011 Accelerated profile HMM searches. PLoS Comput Biol 7: e1002195.
Edgar, R. C., 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792-1797.
Eme, L., E. Gentekaki, B. Curtis, J. M. Archibald and A. J. Roger, 2017 Lateral gene transfer in the adaptation of the anaerobic parasite *Blastocystis* to the gut. Curr Biol 27: 807-820.
Faguy, D. M., and W. F. Doolittle, 2000 Horizontal transfer of catalase-peroxidase genes between archaea and pathogenic bacteria. Trends Genet 16: 196-197.
Fu, L., B. Niu, Z. Zhu, S. Wu and W. Li, 2012 CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28: 3150-3152.
George, E. E., F. Husnik, D. Tashyreva, G. Prokopchuk, A. Horák *et al.*, 2020 Highly reduced genomes of protist endosymbionts show evolutionary convergence. Curr Biol 30: 925-933.e923.
Hogg, D., G. Pell, P. Dupree, F. Goubet, S. M. Martín-Orúe *et al.*, 2003 The modular architecture of *Cellvibrio japonicus* mannanases in glycoside hydrolase families 5 and 26 points to differences in their role in mannan degradation. Biochem J 371: 1027-1043.
Husnik, F., and J. P. McCutcheon, 2018 Functional horizontal gene transfer from bacteria to eukaryotes. Nat Rev Microbiol 16: 67-79.
Irwin, N. A. T., A. A. Pittis, T. A. Richards and P. J. Keeling, 2022 Systematic evaluation of horizontal gene transfer between eukaryotes and viruses. Nat Microbiol 7: 327-336.
Katoh, K., and D. M. Standley, 2013 MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30: 772-780.
Kaur, B., K. Záhonová, M. Valach, D. Faktorová, G. Prokopchuk *et al.*, 2020 Gene fragmentation and RNA editing without borders: eccentric mitochondrial genomes of diplonemids. Nucleic Acids Res 48: 2694-2708.
Kostygov, A. Y., A. Karnkowska, J. Votýpka, D. Tashyreva, K. Maciszewski *et al.*, 2021 Euglenozoa: taxonomy, diversity and ecology, symbioses and viruses. Open Biol 11: 200407.
Kraeva, N., E. Horáková, A. Y. Kostygov, L. Kořený, A. Butenko *et al.*, 2017 Catalase in *Leishmaniinae*: with me or against me? Infect Genet Evol 50: 121-127.
Lartillot, N., N. Rodrigue, D. Stubbs and J. Richer, 2013 PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. Syst Biol 62: 611-615.
Minh, B. Q., H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams *et al.*, 2020 IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol 37: 1530-1534.
Nelson, D. R., K. M. Hazzouri, K. J. Lauersen, A. Jaiswal, A. Chaiboonchoe *et al.*, 2021 Large-scale genome sequencing reveals the driving forces of viruses in microalgal evolution. Cell Host Microbe 29: 250-266.e258.
Nguyen, L. T., H. A. Schmidt, A. von Haeseler and B. Q. Minh, 2015 IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32: 268-274.
Porter, D., 1973 *Isonema papillatum* sp. n., a new colorless marine flagellate: a light- and electronmicroscopic study. The Journal of Protozoology 20: 351-356.
Prokopchuk, G., T. Korytář, V. Juricová, J. Majstorović, A. Horák *et al.*, 2022 Trophic flexibility of marine diplonemids - switching from osmotrophy to bacterivory. Isme j.
Prokopchuk, G., D. Tashyreva, A. Yabuki, A. Horák, P. Masařová *et al.*, 2019 Morphological, ultrastructural, motility and evolutionary characterization of two new Hemistasiidae species. Protist 170: 259-282.
Ricard, G., N. R. McEwan, B. E. Dutilh, J. P. Jouany, D. Macheboeuf *et al.*, 2006 Horizontal gene transfer from bacteria to rumen ciliates indicates adaptation to their anaerobic, carbohydrates-rich environment. BMC Genomics 7: 22.
Richards, T. A., D. M. Soanes, M. D. Jones, O. Vasieva, G. Leonard *et al.*, 2011 Horizontal gene transfer facilitated the evolution of plant parasitic mechanisms in the oomycetes. Proc Natl Acad Sci U S A 108: 15258-15263.
Sibbald, S. J., L. Eme, J. M. Archibald and A. J. Roger, 2020 Lateral gene transfer mechanisms and pan-genomes in eukaryotes. Trends Parasitol 36: 927-941.
Škodová-Sveráková, I., K. Záhonová, B. Bučková, Z. Füssy, V. Yurchenko *et al.*, 2020 Catalase and ascorbate peroxidase in euglenozoan protists. Pathogens 9.
Stairs, C. W., J. E. Dharamshi, D. Tamarit, L. Eme, S. L. Jørgensen *et al.*, 2020 Chlamydial contribution to anaerobic metabolism during eukaryotic evolution. Sci Adv 6: eabb7258.

Stairs, C. W., L. Eme, S. A. Muñoz-Gómez, A. Cohen, G. Dellaire *et al.*, 2018 Microbial eukaryotes have adapted to hypoxia by horizontal acquisitions of a gene involved in rhodoquinone biosynthesis. Elife 7.

Tashyreva, D., G. Prokopchuk, J. Votypka, A. Yabuki, A. Horak *et al.*, 2018 Life cycle, ultrastructure, and phylogeny of new diplonemids and their endosymbiotic bacteria. MBio 9**:** e02447-02417.

Valach, M., S. Moreira, S. Hoffmann, P. F. Stadler and G. Burger, 2017 Keeping it complicated: mitochondrial genome plasticity across diplonemids. Sci Rep 7**:** 14166.

Van Etten, J., and D. Bhattacharya, 2020 Horizontal gene transfer in eukaryotes: not if, but how much? Trends Genet 36**:** 915-925.

Yoon, K.-H., 2016 Molecular cloning and characterization of β-mannanase B from *Cellulosimicrobium* sp. YB-43. Korean Journal of Microbiology 52**:** 336-343.

16. Evolution of gene families

INTRODUCTION

Gene families are groups of genes in a given organism that have evolved by duplication from a common ancestor. Family members (paralogs) diversify over time by 'dividing labor' (subfunctionalization) or adopting new functions (neofunctionalization). Gene families may continue to grow or shrink if there is no selective force to maintain all copies. New gene families can have different origins. They may emerge either from initially non-coding regions or by horizontal gene transfer. Alternatively, new families can originate from pseudogenes whose ultimate descent is unrecognizable.

A key factor in the genetic diversity of organisms is the evolution of gene families. Birth and death, expansion and contraction of families, provide a window on the particular adaptations of species to their environment. The aim of analyzing gene family evolution in diplonemids and related taxa was to delineate genetic factors potentially contributing to the success of diplonemids in marine ecosystems.

RESULTS and DISCUSSION

For comparative gene family analysis, we established a phylogenetically equilibrated reference dataset from 30 eukaryotic species (reference dataset; **Supplementary Table S23**). The genome-inferred proteomes from diplonemids and the reference species were clustered into ~30,000 families, each of which contained at least two proteins—more than a third included sequences from *D. papillatum*. To trace the evolution of those gene families that are widely represented across eukaryotes, we first constructed a phylogenetic tree using the 167 orthologous proteins that are present in 25 out of 30 species in the dataset. The resulting tree is well-resolved, with all the internal nodes of Euglenozoa having maximum support values (**Supplementary Figure S23**).

As the next step, we determined the evolution of protein families by fitting the data to a birth-death model (CSURÖS 2010), which involved the mapping the family counts onto the tree topology (**Supplementary Figure S23**). Interestingly, in Euglenozoa, the gains of new families and the expansions of existing families are much more frequent than losses and contractions. The by far highest count of gains (~7,000) in the entire tree is the ancestral diplonemid node (N10), demonstrating a substantial diversification of the gene repertoire in the last common ancestor of diplonemids. Similarly prolific is the expansion of gene families at that node.

The euglenozoan clade experienced predominantly family gains and expansions, but exceptions were observed in Kinetoplastea, including the ancestral node (N21) and free-living kinetoplastids. At this node, family contractions and member losses occurred predominantly in genes involved in metabolism, which is corroborated by previous reports of reduced metabolic capabilities in this group (OPPERDOES *et al.* 2016; BUTENKO *et al.* 2020).

In the following, we discuss gains and expansions of the diplonemids; details of the analysis are compiled in the **Additional File 6** geneFam.

Families gains in diplonemids

Out of the ~5,500 gains traced to the diplonemid ancestral branch, about 22% have a functional annotation, inferred from the closest homolog found in the Uniprot database (see **Supplementary Information**: Section 2. Assembly and annotation of the nuclear genome and transcriptome of *Diplonema papillatum*). Only ~200 protein families could be confidently assigned a KEGG ID and consequently classified into higher-order KEGG categories (e.g., 'metabolism', 'signal transduction', 'folding, sorting and degradation', etc.). The low number is presumably due to sequence divergence and sparsity of the KEGG database coverage among unconventional model organisms. A non-exhaustive inspection of the annotations of functionally annotated gene families, which could not be automatically assigned a KEGG ID, indicated that these were distributed across KEGG-type categories in similar proportions as the assigned ones. Thus, the distributions discussed hereafter, while limited to the KEGG-assigned dataset, likely mirror the distribution of the entire dataset in relative proportions (though not in absolute numbers).

Proteins involved in metabolic processes represent the largest group of KEGG-annotated families gained by the diplonemid ancestor. Families in the KEGG category '**carbohydrate metabolism**' have the largest share. Predicted glucoside hydrolases, and particularly glucanases, point to specialization in algal or plant food sources. Diverse gene families for carbohydrate-active enzymes have been specifically gained in *D. papillatum.* The remarkable repertoire of the type species has been analyzed in detail in another section (**Supplementary Information**: Section 12. CAZyme-coding genes in *Diplonema papillatum*).

Families of metabolic proteins falling in the KEGG category '**amino-acid metabolism**' are also numerous. In contrast to other Euglenozoa in our dataset, diplonemids appear to be capable of creatine biosynthesis because the genes for both enzymes, glycine amidinotransferase and guanidinoacetate N-methyltransferase, are present. Therefore, diplonemids likely represent an important source of creatine in the ocean, in addition to metazoans and diatoms (WAWRIK *et al.* 2017). All Euglenozoa, except the parasitic trypanosomatids (which use arginine phosphate as energy buffer (PEREIRA *et al.* 2000)), possess creatine kinase and thus can convert creatine to energy-rich creatine phosphate. It remains to be elucidated which compound serves as a creatine source for euglenozoans other than diplonemids or, alternatively, whether they can survive without it.

At least 40 KEGG-annotated families gained by the ancestral diplonemid branch are involved in **environmental-information processing** and **signal transductions**, with putative kinases, phosphatases and peptidases accounting together for ~12% of all functionally annotated protein families. Notably, two adenylate cyclase families (one with up to 80 members in diplonemids) were gained by the last common diplonemid ancestor. An analysis of kinase-domain-containing proteins (using Pfam PF00069 as a query) across diplonemids, euglenids and free-living prokinetoplastids, revealed between 500 and 1,000 kinase homologs in each analyzed species, numbers comparable to the situation in human (518) and plants (600–2,500) (MANNING *et al.* 2002; LEHTI-SHIU AND SHIU 2012).This observation points to an elaborate kinase repertoire and complex signaling pathways in free-living (unicellular) Euglenozoa, comparable to those in multicellular organisms.

Families expanded in diplonemids

In addition to family gains, expansions appear to have played a comparably important role in the evolution of diplonemids. Similarly to gene gains, expanded families often contribute to **metabolic processes and signaling** (126 and 35 families). In addition, 71 expanded families act in **translation and transcription**, including initiation factors and ribosomal proteins (RPs; see also **Supplementary Information**: Section 10. The cytosolic ribosome of *Diplonema papillatum*). The incorporation of various RP paralogs into ribosomes may allow translation regulation under diverse and fluctuating environmental conditions (HUMMEL *et al.* 2012).

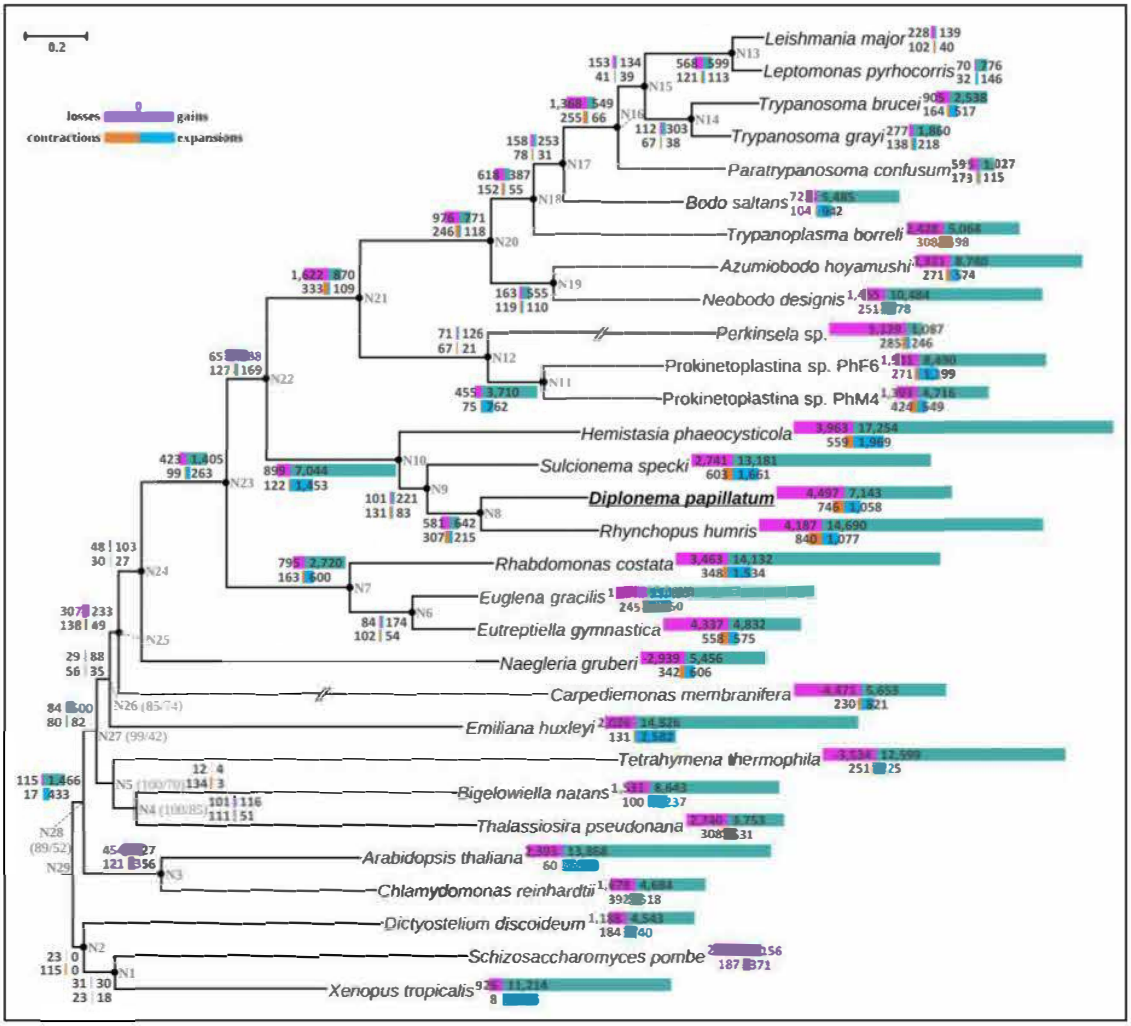
The majority of families expanded specifically in *D. papillatum* are characterized by a small number of members (mainly two or three; **Supplementary Figure S24A**) with highly similar protein sequences (>90%; **Supplementary Figure S24B**), which indicates recent duplication events. In contrast, most families expanded in the diplonemid ancestor seem to have undergone an early diversification (**Supplementary Figure S25**). An interesting group of families expanded in *D. papillatum* but not in the other diplonemids are involved in **oxidative stress protection** (e.g., glutathione S-transferase, cytochrome *c* peroxidase, tryparedoxin, glutaredoxin). Therefore, the type species might be well adapted to life in the surface seawater layer penetrated by solar radiation or in coastal waters polluted with metals, polychlorinated biphenyls, and radioactive waste —conditions triggering the production of cytotoxic reactive oxygen species (LUSHCHAK 2011).

Limitations of the presented analysis

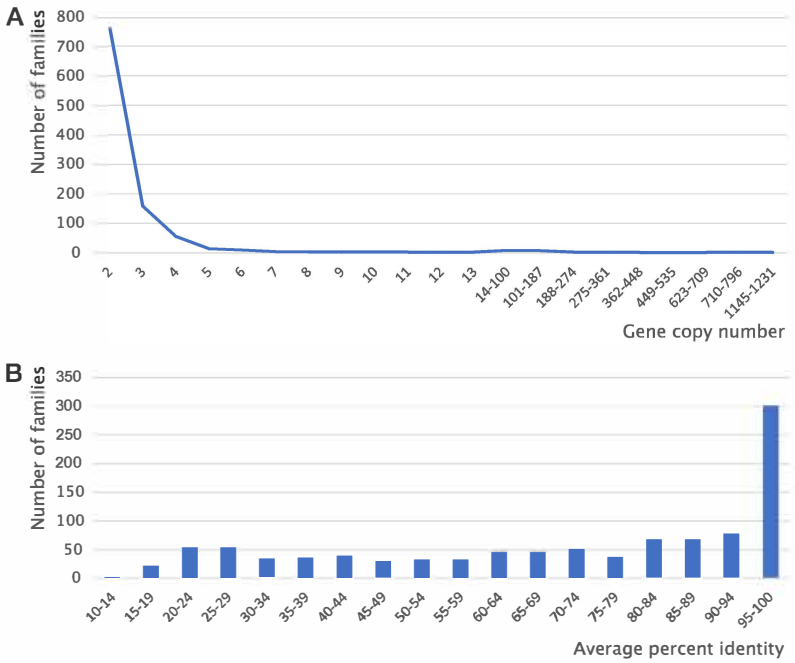
Several shortcomings have complicated our analysis and the consequent interpretation of its results. First, a high proportion of proteins from the understudied diplonemids is not annotated, which applies to 81% and 48% of gained and expanded families, respectively, on the ancestral diplonemid branch. Second, sequencing depth varies across diplonemids, with the effect that certain gene families appear absent in a given species only because the available dataset is less complete than another. Third, cytosolic rRNA phylogenies (TASHYREVA *et al.* 2018) and analyses of the mitochondrial genome (YABUKI *et al.* 2016; KAUR *et al.* 2020) point to an exceptionally accelerated evolution of the *Hemistasia* nuclear genome. Highly divergent protein sequences hamper recognizing homologs and thus artificially increase the count of family gains and lower the count of expansions. Finally, a major difficulty is that only transcriptome-inferred proteomes are available for diplonemids other than *D. papillatum* (**Supplementary Table S23**). In particular, the genome-inferred proteome of *D. papillatum* includes >250 transposon-related hypothetical proteins (see also **Supplementary Information**: Section 2. Assembly and annotation of the nuclear genome and transcriptome of *Diplonema papillatum* and **Supplementary Information**: Section 6. Repetitive sequences in the nuclear genome of *Diplonema papillatum* (assembly v_1.0)), but the corresponding genes seem not to be expressed under usual cultivation conditions because their mRNAs are absent in our RNA-Seq data. Therefore, the considerable family gains and expansions of the type species seen in **Supplementary Figure S23** should be taken with a grain of salt as some of these families are made up of non-expressed genes. Once genome information is available from other diplonemids, the evolution of gene families originating from transposable elements will be worth revisiting.

In sum, almost 80% gained and ~60% expanded gene families in diplonemids lack any functional annotation (i.e., are annotated as 'hypothetical protein'). Out of these, >60% occur in at least three of the four analyzed diplonemids (*D. papillatum*, *Sulcionema specki*, *Rhynchopus humris*, *Hemistasia phaeocysticola*). Most of these conserved genes have a

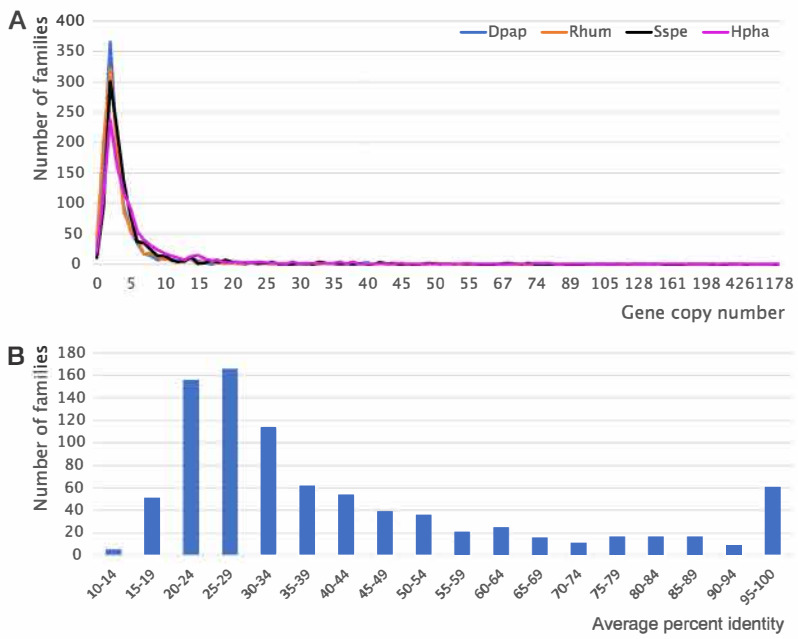
well-expressed representative in *D. papillatum*, which underscores their functional importance. Functional genetics studies will be critical in determining the biological roles of the genetic 'dark matter' comprised in diplonemid genomes.



Supplementary Figure S23. Evolution of protein families. A maximum-likelihood phylogenetic tree based on the concatenated alignment of 167 proteins containing 57,565 amino acid positions. Nodes with maximal statistical support are indicated with black circles; for the remaining nodes the supports are in grey in the following format: (bootstrap support/SH-aLRT value). Double-crossed branches were reduced to half of their original length. Node numbers are shown in grey. Black horizontal bar indicates the number of substitutions per site. The number of protein families gained, lost, expanded, or contracted at certain nodes (based on the sum of probabilities of the respective events at each node/tip) is depicted with green, magenta, blue, and orange bars, respectively. Note the substantial gain of 7,044 families on the ancestral diplonemid branch, which is the largest gene gain on an internal node. See the discussion on *Limitations*.



Supplementary Figure S24. Statistics for protein families expanded in *D. papillatum*. A) The distribution of families ranked by the number of family members (copies) they contain. **B)** The distribution of families ranked by the average percent of protein identity within the group.



Supplementary Figure S25. Statistics for protein families expanded in the reconstructed diplonemid ancestor. A) The distribution of families according to the number of proteins they contain. **B)** The distribution of families according to the average *D. papillatum* protein percent identity within the groups. Species abbreviations are as follows: Dpap, *D. papillatum*; Rhum, *R. humris*; Sspe, *S. specki*; Hpha, *H. phaeocysticola*.

Supplementary Table S23. Data sources.

Species	Strain	Abbreviation	Lineage	Data source[a]	Proteome inferred
Arabidopsis thaliana	-	Atha	Streptophyta	NCBI Genomes	Genome
Azumiobodo hoyamushi	Hirose, Nozawa,	Ahoya	Euglenozoa	(YAZAKI *et al.* 2017)	Transcriptome
Bigelowiella natans		Bnat	Rhizaria		Genome
Bodo saltans	Konstanz	Bsal	Euglenozoa	Wellcome Trust	Genome
Carpediemonas	-	Cmem	CLO/Fornicata	NCBI Genomes	Genome
Chlamydomonas reinhardtii	-	Crei	Chlorophyta	NCBI Genomes	Genome
Dictyostelium discoideum	AX4	Ddis	Amoebozoa	NCBI Genomes	Genome
Diplonema papillatum	ATCC 50162	Dpap	Euglenozoa	this study	Genome
Emiliania huxleyi	CCMP1516	Ehux	Haptophyta	NCBI Genomes	Genome
Euglena gracilis	Z	Egra	Euglenozoa	(EBENEZER *et al.*	Transcriptome
Eutreptiella gymnastica	NIES-381	Egym	Euglenozoa	MMETSP	Transcriptome
Hemistasia phaeocysticola	YPF1303	Hpha	Euglenozoa	(BUTENKO *et al.*	Transcriptome
Leishmania major	Friedlin	Lmaj	Euglenozoa	TriTryp v.9.0	Genome
Leptomonas pyrrhocoris	H10	Lpyr	Euglenozoa	TriTryp v.9.0	Genome
Naegleria gruberi	NEG-M	Ngru	Heterolobosea	NCBI Genomes	Genome
Neobodo designis	CCAP 1951/1	Ndes	Euglenozoa	MMETSP	Transcriptome
Paratrypanosoma confusum	CUL13-MS	Pcon	Euglenozoa	TriTryp v.9.0	Genome
Perkinsela sp.	CCAP 1560/4	Perk	Euglenozoa	NCBI Genomes	Genome
Prokinetoplastina sp.	PhM-4	PhM-4	Euglenozoa	Patrick Keeling	Transcriptome
Prokinetoplastina sp.	PhF-6	PhF-6	Euglenozoa	Patrick Keeling	Transcriptome
Rhabdomonas costata	PANT2	Rcos	Euglenozoa	Vladimír Hampl	Transcriptome
Rhynchopus humris	YPF1608	Rhum	Euglenozoa	(BUTENKO *et al.*	Transcriptome
Schizosaccharomyces pombe	assembly ASM294v2	Spomb	Fungi	NCBI Genomes	Genome
Sulcionema specki	YPF1618	Sspe	Euglenozoa	(BUTENKO *et al.*	Transcriptome
Tetrahymena thermophila	SB210 (assembly JCVI-	Ttherm	Alveolata	NCBI Genomes	Genome
Thalassiosira pseudonana	CCMP1335	Tpse	Stramenopiles	NCBI Genomes	Genome
Trypanoplasma borreli	Tt-JH	Tbor	Euglenozoa	(BUTENKO *et al.*	Transcriptome
Trypanosoma brucei	TREU927	Tbru	Euglenozoa	TriTryp v.9.0	Genome
Trypanosoma grayi	ANR4	Tgra	Euglenozoa	TriTryp v.9.0	Genome
Xenopus tropicalis	-	Xtro	Metazoa	NCBI Genomes	Genome

[a] Download date of all datasets, except when specified otherwise: 06/2017-02/2018.

MATERIALS AND METHODS

Inference of protein families and phylogenomic tree construction

Clustering of proteins into families was performed using OrthoFinder v2.1.2 (EMMS AND KELLY 2019) with the default settings on a dataset of 30 species, including *D. papillatum*, 18 other euglenozoans and several additional species representing major eukaryotic outgroup lineages (**Supplementary Table S23**). In the case of proteins predicted from the euglenozoan transcriptomes, CD-HIT-EST (LI AND GODZIK 2006) with the 90% identity threshold was used to reduce protein redundancy caused by the presence of transcript isoforms. For phylogenomic tree construction, proteins present in two or more copies were removed from the OrthoFinder output, and 167 families containing proteins encoded by single-copy genes in at least 25 species out of 30 were used in further analysis. The respective sequences were aligned using MAFFT v7.402 with 1,000 iterations and the "localpair" option (KATOH AND STANDLEY 2013), trimmed with TrimAl v1.2 (CAPELLA-GUTIÉRREZ *et al.* 2009) to remove poorly-aligned positions and concatenated. The phylogenomic tree was inferred using IQ-Tree v1.6.12 (NGUYEN *et al.* 2015) with the LG+C60+F substitution model, which was the best-fitting model according to the BIC criterion; 1000 ultrafast bootstrap replicates were used to assess branch support (HOANG *et al.* 2018). Gene family evolution was modelled on the phylogenomic tree using the phylogenetic birth-and-death model implemented in Count (CSURÖS 2010). We fit a birth-death model with three rate categories for each, gene gain, loss, duplication, and family-specific rate multipliers. The family-wise posterior probabilities of gene presence, expansion, and contraction at each internal node of the phylogeny were used to compute the number of events at each node and the list of genes present with probability ≥ 0.5; the latter was used for ancestral metabolic reconstructions. Since we experienced problems fitting a birth-death model for the largest gene families, we employed a Wagner parsimony analysis with a gain penalty of 3 for the families containing more than 150 members (332 families in total).

Sequence analysis

For the analysis of general metabolism, BLASTp searches with an E-value cut-off of 10e-20 were conducted against the Euglenozoa species in the reference dataset (**Supplementary Table S23**) (CAMACHO *et al.* 2009). When necessary, additional homology searches were performed using the HMMER package v.3.1 with an E-value cut-off of 10e-5 (EDDY 2011). The average percent identity within OGs of interest was calculated using the alistat script from the HMMER package. Signal peptides were predicted using SignalP v. 5.0 server (ALMAGRO ARMENTEROS *et al.* 2019).

AUTHOR CONTRIBUTIONS

Conceptualization – A.B., T.W.; **Data curation, Investigation, Formal analysis** – A.B., T.W., C.P.; **Visualization** – A.B.; **Writing, original draft** – A.B., T.W., C.P., J.L.; **Writing, review & editing** – all authors.

REFERENCES

Almagro Armenteros, J. J., K. D. Tsirigos, C. K. Sønderby, T. N. Petersen, O. Winther *et al.*, 2019 SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol 37**:** 420-423.

Butenko, A., F. R. Opperdoes, O. Flegontova, A. Horák, V. Hampl *et al.*, 2020 Evolution of metabolic capabilities and molecular features of diplonemids, kinetoplastids, and euglenids. BMC Biol 18**:** 23.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos *et al.*, 2009 BLAST+: architecture and applications. BMC Bioinformatics 10**:** 421.

Capella-Gutiérrez, S., J. M. Silla-Martínez and T. Gabaldón, 2009 trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25**:** 1972-1973.

Csurös, M., 2010 Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics 26**:** 1910-1912.

Ebenezer, T. E., M. Zoltner, A. Burrell, A. Nenarokova, A. M. G. Novák Vanclová *et al.*, 2019 Transcriptome, proteome and draft genome of *Euglena gracilis.* BMC Biol 17**:** 11.

Eddy, S. R., 2011 Accelerated profile HMM searches. PLoS Comput Biol 7**:** e1002195.

Emms, D. M., and S. Kelly, 2019 OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol 20**:** 238.

Hoang, D. T., O. Chernomor, A. von Haeseler, B. Q. Minh and L. S. Vinh, 2018 UFBoot2: improving the ultrafast bootstrap approximation. Mol Biol Evol 35**:** 518-522.

Hummel, M., J. H. Cordewener, J. C. de Groot, S. Smeekens, A. H. America *et al.*, 2012 Dynamic protein composition of *Arabidopsis thaliana* cytosolic ribosomes in response to sucrose feeding as revealed by label free MSE proteomics. Proteomics 12**:** 1024-1038.

Katoh, K., and D. M. Standley, 2013 MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30**:** 772-780.

Kaur, B., K. Záhonová, M. Valach, D. Faktorová, G. Prokopchuk *et al.*, 2020 Gene fragmentation and RNA editing without borders: eccentric mitochondrial genomes of diplonemids. Nucleic Acids Res 48**:** 2694-2708.

Lehti-Shiu, M. D., and S. H. Shiu, 2012 Diversity, classification and function of the plant protein kinase superfamily. Philos Trans R Soc Lond B Biol Sci 367**:** 2619-2639.

Li, W., and A. Godzik, 2006 Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22**:** 1658-1659.

Lushchak, V. I., 2011 Environmentally induced oxidative stress in aquatic animals. Aquat Toxicol 101**:** 13-30.

Manning, G., D. B. Whyte, R. Martinez, T. Hunter and S. Sudarsanam, 2002 The protein kinase complement of the human genome. Science 298**:** 1912-1934.

Nguyen, L. T., H. A. Schmidt, A. von Haeseler and B. Q. Minh, 2015 IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32**:** 268-274.

Opperdoes, F. R., A. Butenko, P. Flegontov, V. Yurchenko and J. Lukeš, 2016 Comparative metabolism of free-living *Bodo saltans* and parasitic trypanosomatids. J Eukaryot Microbiol 63**:** 657-678.

Pereira, C. A., G. D. Alonso, M. C. Paveto, A. Iribarren, M. L. Cabanas *et al.*, 2000 *Trypanosoma cruzi* arginine kinase characterization and cloning. A novel energetic pathway in protozoan parasites. J Biol Chem 275**:** 1495-1501.

Tashyreva, D., G. Prokopchuk, A. Yabuki, B. Kaur, D. Faktorová *et al.*, 2018 Phylogeny and morphology of new diplonemids from Japan. Protist 169**:** 158-179.

Wawrik, B., D. A. Bronk, S. E. Baer, L. Chi, M. Sun *et al.*, 2017 Bacterial utilization of creatine in seawater. Aquat Microb Ecol 80**:** 153-165.

Yabuki, A., G. Tanifuji, C. Kusaka, K. Takishita and K. Fujikura, 2016 Hyper-eccentric structural genes in the mitochondrial genome of the algal parasite *Hemistasia phaeocysticola*. Genome Biol Evol 8**:** 2870-2878.

Yazaki, E., S. A. Ishikawa, K. Kume, A. Kumagai, T. Kamaishi *et al.*, 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 Genet Syst 92**:** 35-42.

17. Feeding strategy and food of *Diplonema papillatum*

LITERATURE REVIEW and DISCUSSION

The feeding strategy of diplonemids has been controversial for several decades, ranging from parasitic to predatory and saprophytic/scavenging (e.g., SCHUSTER *et al.* 1968; TRIEMER AND OTT 1990; VICKERMAN 2000; VON DER HEYDEN *et al.* 2004). Documented eukaryvors are, for example, *Hemistasia phaeocysticola* (ELBRÄCHTER *et al.* 1996; YABUKI AND TAME 2015) and *Rhynchopus coscinodiscivorus* (SCHNEPF 1994), both of which have been observed in the act of predating diatoms, whereas *R. euleeides*, for instance, consumes bacteria (ROY *et al.* 2007). In addition, a given species may feed on the food that is available, as observed for *D. japonicum* and *R. humris*, both of which are osmotrophs in a rich medium containing serum, but gradually switch to bacterivory under nutrient-poor conditions (PROKOPCHUK *et al.* 2022).

D. papillatum almost certainly has similar dietary plasticity as the two latter diplonemids. Under laboratory conditions, the type species feeds osmotrophically, but equipped, as are all diplonemids, with a cytostome and cytopharynx (PORTER 1973), it likely engulfs small microbes and particles in the wild. However, *D. papillatum* seems not to consume bacteria, because (i) enzymes for degrading peptidoglycan (murein), the main cell wall constituent of most bacteria, have not been detected in its genome (see **Supplementary Information**: Section 12. CAZyme-coding genes in *Diplonema papillatum*), and (ii) feeding experiments offering diverse live bacteria that otherwise allow profuse growth of, e.g., *D. japonicum* and *R. humris*, do not support measurable proliferation of *D. papillatum* (PROKOPCHUK *et al.* 2022). Interestingly, it has been suggested that the flagella of *Diplonema* (and *Rhynchopus*) serve for the recognition of suitable solid food rather than for locomotion because their axoneme may be disordered and because the flagella are too short for support and bending (SCHNEPF 1994).

It is generally assumed that *D. papillatum* scavenges on the debris of dead microeukaryotes or plants. This view is in line with the particular morphology of this protist, characterized by a narrow (~0.5–1 μm; see **Fig. 1**, main text) and rigid, microtubule-reinforced cytostome (PORTER 1973), limiting the type of microeukaryotic species that can be ingested intact by *D. papillatum.* More specifically, the upper size limit of eukaryotic prey that can be devoured whole by *D. papillatum* would be in the range of the smallest microalga described, *Ostreococcus tauri* (~0.7 μm wide and 1 μm long; COURTIES *et al.* 1994; CHRÉTIENNOT-DINET *et al.* 1995). In contrast to *D. papillatum*, the eukaryvorous heterotrophic euglenids *Peranema* and *Heteronema* (BREGLIA *et al.* 2013) and raptorial ciliates (VERNI AND GUALTIERI 1997) have highly expandable cytostomes, allowing ingestion of prey nearly as wide as the predator itself.

The extensive arsenal of CAZyme-encoding genes that we detected in the *D. papillatum* nuclear genome strongly suggests that in its natural habitat, this protist feeds on live eukaryotes after piercing and rupturing prey cells enzymatically. Cell particles can then be engulfed in the cytopharynx through the cytostome and broken down extracellularly into oligomeric compounds. The subsequent steps most likely follow classical endocytosis, by which the material is taken up inside the cell via food vacuoles, which in turn fuse with lysosomes in which the oligomers are broken down to amino acids, fatty acids, and other monomeric molecules, for use as an energy source and building blocks of the cell.

The feeding strategy mentioned above would allow *D. papillatum* to forage on eukaryotes from an extensive taxonomic range and of any physical size. Such a feeding behaviour, together with an adequate swimming capacity of trophic cells, could explain the ecological distribution of *D. papillatum* in eukaryote-rich coastal regions.

AUTHOR CONTRIBUTIONS

Writing, original draft – G.B.; **Writing, review & editing** – all co-authors.

REFERENCES

Breglia, S. A., N. Yubuki and B. S. Leander, 2013 Ultrastructure and molecular phylogenetic position of *Heteronema scaphurum*: a eukaryovorous euglenid with a cytoproct. J Eukaryot Microbiol 60**:** 107-120.

Chrétiennot-Dinet, M. J., C. Courties, A. Vaquer, J. Neveux, H. Claustre *et al.*, 1995 A new marine picoeucaryote: *Ostreococcus tauri* gen. et sp. nov. (Chlorophyta, Prasinophyceae). Phycologia 34**:** 285-292.

Courties, C., A. Vaquer, M. Troussellier, J. Lautier, M. J. Chrétiennot-Dinet *et al.*, 1994 Smallest eukaryotic organism. Nature 370**:** 255-255.

Elbrächter, M., E. Schnepf and I. Balzer, 1996 *Hemistasia phaeocysticola* (Scherffel) comb. nov., redescription of a free-living, marine, phagotrophic kinetoplastid flagellate. Arch. Protistenkd. 147**:** 125-136.

Porter, D., 1973 *Isonema papillatum* sp. n., a new colorless marine flagellate: a light- and electronmicroscopic study. The Journal of Protozoology 20: 351-356.
Prokopchuk, G., T. Korytář, V. Juricová, J. Majstorović, A. Horák *et al.*, 2022 Trophic flexibility of marine diplonemids - switching from osmotrophy to bacterivory. ISME J.
Roy, J., D. Faktorova, O. Benada, J. Lukes and G. Burger, 2007 Description of *Rhynchopus euleeides* n. sp. (Diplonemea), a free-living marine euglenozoan. J Eukaryot Microbiol 54: 137-145.
Schnepf, E., 1994 Light and electron microscopical observations in *Rhynchopus coscinodiscivorus* spec. nov., a colorless, phagotrophic euglenozoon with concealed flagella. Arch. Protistenkd. 144: 63-74.
Schuster, F. L., S. Goldstein and B. Hershenov, 1968 Ultrastructure of a flagellate, *Isonema nigricans* nov. gen. nov. sp., from a polluted marine habitat. Protistologica 4: 141-149.
Triemer, R. E., and D. W. Ott, 1990 Ultrastructure of *Diplonema ambulator* Larsen & Patterson (Euglenozoa) and its relationship to *Isonema*. Eur J Protistol 25: 316-320.
Verni, F., and P. Gualtieri, 1997 Feeding behaviour in ciliated protists. Micron 28: 487-504.
Vickerman, K., 2000 Diplonemids (Class: Diplonemea Cavalier Smith, 1993). pp. 1157-1159 in *An Illustrated Guide to the Protozoa*, edited by J. J. Lee, G. F. Leedale and P. Bradbury. Society of Protozoologists, Lawrence, Kansas, U.S.A.
von der Heyden, S., E. E. Chao, K. Vickerman and T. Cavalier-Smith, 2004 Ribosomal RNA phylogeny of bodonid and diplonemid flagellates and the evolution of Euglenozoa. J Eukaryot Microbiol 51: 402-416.
Yabuki, A., and A. Tame, 2015 Phylogeny and reclassification of *Hemistasia phaeocysticola* (Scherffel) Elbrächter & Schnepf, 1996. J Eukaryot Microbiol 62: 426-429.

18. Environmental distribution of *Diplonema papillatum*

INTRODUCTION

Diplonemids are ubiquitous members of microbial communities in marine environments, found in the top layer down to abyssal/hadal zones (>6,000 m below the surface), in both the pelagic (planktonic) and benthic (sediment) habitats, and in all geographic regions from the tropics to the poles.

Estimates of the **relative abundance** of diplonemids vary greatly depending on sampling methods, depth, and geographic location. For example, an *in situ* fluorescence hybridization survey of the tropical Atlantic Ocean estimated that diplonemids make up 1–3% of eukaryotes in the water column at 100–7,000 m below the surface (MORGAN-SMITH *et al.* 2013). Other studies used metabarcoding, a large-scale bio-assessment method that involves PCR amplification of hypervariable regions of 18S rDNA or rRNA in environmental samples. The V9 hypervariable region of the 18S rDNA is best-suited for targeting diplonemids. In contrast, the widely used V4 metabarcoding (e.g., MASSANA *et al.* 2015) often fails to detect members of this group because their V4 region exceeds 500 bp, which is above the typical insert length of high-throughput amplicon sequencing pipelines (FLEGONTOVA *et al.* 2016). Analysis of a V9 metabarcoding dataset obtained from 0.8–2,000 µm marine-plankton size fractions from the temperate and tropical oceans (DE VARGAS *et al.* 2015) revealed that diplonemids represented ~1% of all microeukaryotes in the photic zone (0–200 m depth), and reached a peak of as much as 14% in the upper portion of the aphotic zone, at a depth of ~750–1,000 m (FLEGONTOVA *et al.* 2020). A similar V9 metabarcoding survey led to the estimate that diplonemids make up, on average, 5% of microeukaryotes in the bathyal-hadal marine sediment (SCHOENLE *et al.* 2021).

Microbial diversity is best represented by the number of operational taxonomic units (**OTUs**) that groups organisms into bins, which are traditionally based on sequence similarity, but more recently on sequence clustering (MAHÉ *et al.* 2015). According to several (but not all) global V9 metabarcoding studies, the total number of diplonemid OTUs detected in the plankton of the world ocean amounts to tens of thousands, surpassing the numbers of OTUs for planktonic metazoans, stramenopiles, dinoflagellates, and rhizarians (FLEGONTOVA *et al.* 2020; CORDIER *et al.* 2022) (see also **Supplementary Table S24**).

The large majority of the biomass of diatoms, pelagophytes, dinoflagellates and some other protists is constituted by a relatively small number of OTUs (KEELING AND CAMPO 2017). In diplonemids, the bias in the OTU abundance profile is even more pronounced. For example, in the Tara Oceans dataset, the hundred most abundant OTUs (out of 45,197) represent more than 92% of all diplonemid reads (FLEGONTOVA *et al.* 2016). Planktonic diplonemids of the Eupelagonemidae clade account for >97% of all diplonemid OTUs and reads, whereas the other three clades—Diplonemidae, Hemistasiidae, and DSPD II—each account for only ~1% or even less (DE VARGAS *et al.* 2015; FLEGONTOVA *et al.* 2016; FLEGONTOVA *et al.* 2020). It should be noted that Eupelagonemidae and DSPD II OTUs occur predominantly in the mesopelagic zone (at 200–1,000 m depth). In contrast, Diplonemidae and Hemistasiidae OTUs are primarily found in the surface zone (FLEGONTOVA *et al.* 2016).

Diplonemids living in freshwater are an exception. They have been detected in the plankton of lakes in Sweden (SKUJA 1948), Japan, the Czech Republic, and Switzerland (MUKHERJEE *et al.* 2020), and in the plankton and sediments of Lake Baikal (YI *et al.* 2017; DAVID *et al.* 2021; REBOUL *et al.* 2021). Freshwater diplonemids occur at a very low abundance (<1% of microeukaryotes), and their diversity is limited. Samples collected from the geographically most distant sites all fall within the Diplonemidae clade, consistent with Skuja's hypothesis of a recent habitat transition from a marine to a freshwater environment (MUKHERJEE *et al.* 2020; DAVID *et al.* 2021).

The true diversity of diplonemids in the environment has been a matter of controversy due to genomic variability of the 18S rRNA gene sequences found in single-cell genomes of ten distinct marine planktonic diplonemids (MUKHERJEE *et al.* 2020). However, such intra-genomic variability is not unique to diplonemids and probably affects nearly all estimates of diversity relying on high-throughput sequencing of marker-gene amplicons. For instance, not only multiple distinct sequences, but multiple clusters of sequences (OTUs of the V9 18S rRNA region) were found in mono-species ciliate cultures (FORSTER *et al.* 2019). The same is true for single radiolarian cells, where multiple OTUs of the V9 and V4 18S rRNA regions were found (DECELLE *et al.* 2014). Metabarcoding approaches can overestimate the diversity due to various causes: sequencing errors and PCR artefacts such as chimera formation, as well as intra-genomic and intra-population sequence variability (BÁLINT *et al.* 2016; SANTOFERRARA *et al.* 2020). MUKHERJEE *et al.* 2020 have not demonstrated that diplonemids are more prone to overestimation than other eukaryotes, and we believe that interpreting *relative* diversity of protist groups is possible, even though absolute diversity estimates are currently unreliable.

While we have a good understanding of the environmental and geographical distribution of the major diplonemid groups, the same parameters for the type species *Diplonema papillatum* are not known. Therefore, we searched available

barcoding resources for the presence of sequences from *D. papillatum*. To leverage information not only from V9, but also from the more frequent V4 datasets, we analyzed forward and reverse sequencing reads from the latter separately, allowing incomplete coverage of the V4 region.

Supplementary Table S24. Abundance of OTUs from selected taxa in the Tara project datasets [a].

	OTU abundance bins (reads per million) [b]								
Taxon	≥10,000	≥1,000–<10,000	≥100–<1,000	≥10–<100	≥1–<10	≥0.1–<1	≥0.01–<0.1	<0.01	Total OTUs
Diplonemea–Diplonemidae	0	0	0	1	9	16	81	94	201
Diplonemea–Hemistasiidae	0	0	0	5	7	61	83	58	214
Diplonemea–Eupelagonemidae	0	8	23	14	16	82	3,074	**8,087**	11,304
Diplonemea–DSPDII	0	0	0	3	0	1	14	**38**	56
Diplonemea	0	8	23	23	32	160	3,266	**8,277**	11,775
Kinetoplastea	0	0	4	7	10	36	50	60	167
Euglenida	0	0	1	0	7	11	49	59	127
Heterolobosea	0	0	0	0	5	5	33	24	67
Stramenopiles	0	20	107	208	293	567	1,745	2,870	5,810
Dinoflagellata	1	22	221	724	1,158	1,696	5,254	**10,865**	19,941
Ciliophora	0	4	23	90	164	263	563	781	1,888
Rhizaria	3	20	94	208	354	998	3,435	**5,641**	10,753
Haptophyta	0	1	15	34	46	62	128	228	514

[a] For this analysis, the Tara Oceans and Tara Arctic datasets were combined (IBARBALZ *et al.* 2019). In the Tara Oceans dataset, diplonemids emerged as the most OUT-rich taxon, whereas in the combined Tara datasets, diplonemids rank at second place, behind dinoflagellates.

[b] OTUs were defined by the SWARM algorithm (MAHÉ *et al.* 2015). Sequences sharing >90% identity to a particular reference sequence were merged for consistency with the 90% threshold used for *D. papillatum* sequences that form a single OTU. OTUs of <0.01 abundance that collectively includes>50% of all OTUs of a given taxon are highlighted in blue bold.

RESULTS and DISCUSSION

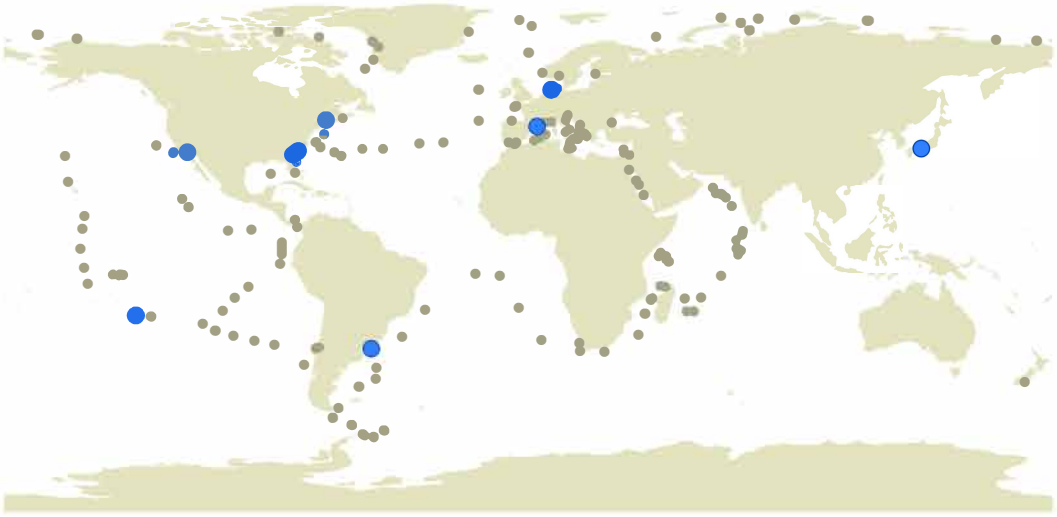
To gain a picture of the global and ecosystemic distribution of *D. papillatum*, we searched for the corresponding sequences in the global V9 metabarcoding datasets ('Tara Oceans' and 'Tara Arctic') generated by the Tara project (DE VARGAS *et al.* 2015; IBARBALZ *et al.* 2019). *D. papillatum* sequences were found in four out of 970 planktonic samples. All four samples originated from surface water: one collected off the Mediterranean Spanish coast (four *D. papillatum* reads; latitude 41.6686, longitude 2.7996) and three from coastal waters in the reef surrounding the Gambier Islands in the Pacific Ocean (29 reads; latitude cca. −23, longitude cca. −135; **Supplementary Figure S26**). The collection sites had a relatively high water temperature (23–25 ºC), a salinity between 36.46 and 36.52, an oxygen concentration between 201 and 207 µmol/L, and a low chlorophyll A concentration (0.01–0.07), a measure of the amount of suspended phytoplankton. With only 1 to 25 reads from *D. papillatum* per 1–2 million total reads, the abundance in these samples was very low. Since no *D. papillatum*-representing sequence was found in Tara samples from the open ocean, we prioritized, for further analyses, the datasets that include coastal samples.

Our search in the V4 and V9 metabarcoding datasets 'Ocean Sampling Day 2014' (KOPF *et al.* 2015) and 'Helgoland Roads 2016'(KÄSE *et al.* 2020), which comprise coastal samples, returned a small number of *D. papillatum* sequences. Specifically, among 31 metabarcoding samples of the Ocean Sampling Day project, nine contained 1–13 V4 and/or V9 reads per sample (4 reads on average, per 0.3-0.4 million total reads) from *D. papillatum*. The corresponding samples were collected at a depth of 0–10 m under the surface at temperatures ranging from 11 °C to 30 °C and originated from geographically distant locations off the coast of Japan, the East and West coasts of North America, the East coast of South America, and the North Sea coast (**Supplementary Figure S26**). Among 287 V4-metabarcoding samples collected during various seasons at the Helgoland Roads Long-Term Ecological Research site (Helgoland Islands coast, North Sea, depth of 1 m; latitude 54.1838333, longitude 7.9; temperature 6–13 °C) (KÄSE *et al.* 2020), *D. papillatum* sequences were identified in five samples (1–10 *D. papillatum* reads per sample, per 0.1-3 million total reads).

D. papillatum sequences were absent from the two other V4 metabarcoding datasets collected from coastal waters of Europe (MASSANA *et al.* 2015), Brittany (France) and Senegal (RAMOND *et al.* 2019). The failure to detect *D. papillatum* sequences in these latter data sets might be due to the relatively low number of total reads produced by these two studies (127–30,296 and 16,611–185,889 reads per sample, respectively).

We compared the relative abundance of V9 reads from *D. papillatum* with that of other diplonemids in Tara Oceans and Tara Arctic datasets (**Supplementary Table S25**). The table shows that *Hemistasia phaeocysticola* and the Diplonemidae isolate YPF1523 reached a total abundance of ~70, and more reads per million. Otherwise, the abundance was generally below 1 read per million, with *D. papillatum* being among the least abundant diplonemid species (~1–10 reads per million in samples in which it was detected and ~0.04 reads per million over all samples; **Supplementary Table S25, Additional File 7** envirDistr). To summarize, while apparently quite rare, *D. papillatum* seems to be broadly distributed in temperate coastal surface waters of the world ocean (**Supplementary Figure S26**).

We note, however, that the barcode-derived quantification of diplonemids remains rather approximative. In eukaryotes, the rDNA loci copy number can vary substantially based on environmental or cultivation cues within a species, but also between species irrespective of their degree of relatedness, thus introducing biases into abundance estimates (LAVRINIENKO *et al*. 2021). We currently lack information on the expansion of the rDNA loci in diplonemids other than *D. papillatum*. Taken together, the conclusion about the relative rarity of *D. papillatum* compared to other diplonemids rests upon the assumption that the rDNA copy number is both similar and relatively constant across diplonemids.



Supplementary Figure S26. Oceanic distribution of *D. papillatum*. The world map shows the distribution of sampling locations from the three datasets in which *D. papillatum* was detected, namely the Tara project, 'Ocean Sampling Day 2014', and 'Helgoland Roads 2016' (beige dots). Sites, where *D. papillatum*-representing OTUs were detected, are highlighted in blue (see the **Additional File 7** envirDistr for details).

Supplementary Table S25. Abundance of selected diplonemids in the combined Tara project datasets.

Family	Species	Abundance (V9 reads per million)[a]
Diplonemidae	*Diplonema papillatum*	0.04
	Diplonema aggregatum	0.04
	Diplonema japonicum	0.06
	Diplonema sp. ATCC 50232	0.10
	Rhynchopus humris	0.40
	Rhynchopus serpens	1.17
	Rhynchopus sp. SH-2004-I	1.26
	Flectonema neradi	<0.01
	Lacrimia lanifica	2.79
Hemistasiidae	*Hemistasia phaeocysticola*	69.13
Unknown	Diplonemidae sp. YPF1523	78.08

[a] Number of reads across all samples. In the subset of samples that contain reads from *D. papillatum*, the number of these reads per sample is 1–10.

METHODS

We searched for *D. papillatum* sequences in the datasets from six studies that reported 18S rDNA V4 or V9 metabarcoding reads from marine plankton (**Additional File 7** envirDistr; (DE VARGAS *et al.* 2015; KOPF *et al.* 2015; MASSANA *et al.* 2015; IBARBALZ *et al.* 2019; RAMOND *et al.* 2019; KÄSE *et al.* 2020). Amplification primers were removed from raw reads using *cutadapt* v1.16 (MARTIN 2001). The V9 and V4 regions of *D. papillatum* are 128 and 530 bp long, respectively. To simplify sequence analysis, we merged paired V9 reads using *bbmerge* (BBMap package v38.9) (BUSHNELL *et al.* 2017). The V4 reads could only be merged for a single study because in this case they were long enough to cover the whole V4 region (KÄSE *et al.* 2020). For the remaining studies, forward and reverse V4 reads were analyzed separately. Taxonomic annotation of the processed reads was performed using a global alignment tool *ggsearch* (FASTA package v3.5) (PEARSON 2000). As a reference database we used all 525 diplonemid 18S rRNA sequences from the EukRef-excavates collection (KOLISKO *et al.* 2020), including two *D. papillatum* sequences. To account for slightly different amplification primers used across the studies, we processed the database sequences so that they specifically corresponded to the amplified region in each study. Lastly, since the V9 and V4 regions are highly variable across species and can even vary within a single genome, we established a percent-identity cutoff for *ggsearch* hits. We compared V4 and V9 regions of *D. papillatum* and EukRef diplonemid sequences. We also compared all V4 and V9 sequence variants found in the *D. papillatum* genome assembly to the sequence deposited in NCBI (GenBank accession number KF633466). Within the *D. papillatum* genome assembly multiple copies of the V4 and V9 regions exist (as delimited by the amplification primers), but the identity of sequences from other diplonemids never exceeded 90%; therefore, we chose 90% as the identity cutoff. Relative abundance in reads per million (**Supplementary Tables S24, S25**) was calculated using a combined dataset of eukaryotic metabarcodes from Tara Oceans and Tara Arctic (DE VARGAS *et al.* 2015; IBARBALZ *et al.* 2019).

AUTHOR CONTRIBUTIONS

Conceptualization, Data curation, Formal analysis, Investigation, Writing, original draft – O.F.; **Visualization** – M.V.; **Writing, review & editing** – all co-authors.

REFERENCES

Bálint, M., M. Bahram, A. M. Eren, K. Faust, J. A. Fuhrman *et al.*, 2016 Millions of reads, thousands of taxa: microbial community structure and associations analyzed via marker genes. FEMS Microbiol Rev. 40**:** 686-700.

Bushnell, B., J. Rood and E. Singer, 2017 BBMerge - Accurate paired shotgun read merging via overlap. PLoS One 12**:** e0185056.

Cordier, T., I. B. Angeles, N. Henry, F. Lejzerowicz, C. Berney *et al.*, 2022 Patterns of eukaryotic diversity from the surface to the deep-ocean sediment. Sci Adv 8**:** eabj9309.

David, G. M., D. Moreira, G. Reboul, N. V. Annenkova, L. J. Galindo *et al.*, 2021 Environmental drivers of plankton protist communities along latitudinal and vertical gradients in the oldest and deepest freshwater lake. Environ Microbiol 23**:** 1436-1451.

de Vargas, C., S. Audic, N. Henry, J. Decelle, F. Mahe *et al.*, 2015 Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. Science 348**:** 1261605.

Decelle, J., S. Romac, E. Sasaki, F. Not and F. Mahé, 2014 Intracellular diversity of the V4 and V9 regions of the 18S rRNA in marine protists (radiolarians) assessed by high-throughput sequencing, PLoS One 9**:** e104297.

Flegontova, O., P. Flegontov, P. A. C. Londoño, W. Walczowski, D. Šantić *et al.*, 2020 Environmental determinants of the distribution of planktonic diplonemids and kinetoplastids in the oceans. Environ Microbiol 22**:** 4014-4031.

Flegontova, O., P. Flegontov, S. Malviya, S. Audic, P. Wincker *et al.*, 2016 Extreme diversity of diplonemid eukaryotes in the ocean. Curr Biol 26**:** 3060-3065.

Forster, D., G. Lentendu, S. Filker, E. Dubois, T.A. Wilding *et al.*, 2019 Improving eDNA-based protist diversity assessments using networks of amplicon sequence variants, Environ Microbiol 21**:** 4109-4124.

Ibarbalz, F. M., N. Henry, M. C. Brandão, S. Martini, G. Busseni *et al.*, 2019 Global trends in marine plankton diversity across kingdoms of life. Cell 179**:** 1084-1097.e1021.

Käse, L., A. C. Kraberg, K. Metfies, S. Neuhaus, P. A. A. Sprong *et al.*, 2020 Rapid succession drives spring community dynamics of small protists at Helgoland Roads, North Sea. J Plankton Res 42**:** 305-319.

Keeling, P. J., and J. D. Campo, 2017 Marine protists are not just big bacteria. Curr Biol 27**:** R541-r549.

Kolisko, M., O. Flegontova, A. Karnkowska, G. Lax, J. M. Maritz *et al.*, 2020 EukRef-excavates: seven curated SSU ribosomal RNA gene databases. Database (Oxford) 2020.

Kopf, A., M. Bicak, R. Kottmann, J. Schnetzer, I. Kostadinov *et al.*, 2015 The ocean sampling day consortium. Gigascience **4**: 27.
Lavrinienko, A., T. Jernfors, J. J. Koskimäki, A. M. Pirttilä and P. C. Watts, 2021 Does intraspecific variation in rDNA copy number affect analysis of microbial communities? Trends Microbiol. **29**: 19-27.
Mahé, F., T. Rognes, C. Quince, C. de Vargas and M. Dunthorn, 2015 Swarm v2: highly-scalable and high-resolution amplicon clustering. PeerJ **3**: e1420.
Martin, M., 2001 Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17.
Massana, R., A. Gobet, S. Audic, D. Bass, L. Bittner *et al.*, 2015 Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing. Environ Microbiol **17**: 4035-4049.
Morgan-Smith, D., M. A. Clouse, G. J. Herndl and A. B. Bochdansky, 2013 Diversity and distribution of microbial eukaryotes in the deep tropical and subtropical North Atlantic Ocean. Deep Sea Research Part I: Oceanographic Research Papers **78**: 58-69.
Mukherjee, I., M. M. Salcher, A. Andrei, V. S. Kavagutti, T. Shabarova *et al.*, 2020 A freshwater radiation of diplonemids. Environ Microbiol **22**: 4658-4668.
Pearson, W. R., 2000 Flexible sequence similarity searching with the FASTA3 program package. Methods Mol Biol **132**: 185-219.
Ramond, P., M. Sourisseau, N. Simon, S. Romac, S. Schmitt *et al.*, 2019 Coupling between taxonomic and functional diversity in protistan coastal communities. Environ Microbiol **21**: 730-749.
Reboul, G., D. Moreira, N. V. Annenkova, P. Bertolino, K. E. Vershinin *et al.*, 2021 Marine signature taxa and core microbial community stability along latitudinal and vertical gradients in sediments of the deepest freshwater lake. Isme j **15**: 3412-3417.
Santoferrara, L., F. Burki, S. Filker, R. Logares, M. Dunthorn *et al.*, 2020 Perspectives from ten years of protist studies by high-throughput metabarcoding. J Eukaryot Microbiol. **67**: 612-622.
Schoenle, A., M. Hohlfeld, K. Hermanns, F. Mahé, C. de Vargas *et al.*, 2021 High and specific diversity of protists in the deep-sea basins dominated by diplonemids, kinetoplastids, ciliates and foraminiferans. Commun Biol **4**: 501.
Yi, Z., C. Berney, H. Hartikainen, S. Mahamdallie, M. Gardner *et al.*, 2017 High-throughput sequencing of microbial eukaryotes in Lake Baikal reveals ecologically differentiated communities and novel evolutionary radiations. FEMS Microbiol Ecol 93.

19. DNA and RNA preparation for high-throughput sequencing

MATERIALS AND METHODS

Strains and culture conditions

Diplonema papillatum (ATCC 50162) was initially obtained from the American Type Culture Collection (ATCC) and cultivated axenically without shaking at 15–22 °C in liquid saline medium (OS) containing 33 g/L Instant Ocean sea salt and supplemented with 1% (v/v) horse serum as described earlier (VALACH *et al.* 2014). For extended cultivations, chloramphenicol (40 mg/L) was added to prevent bacterial contamination.

For nucleic acid extraction and sequencing at Genome Quebec (Illumina DNA- and RNA-Seq, PacBio DNA-Seq), the protist was cultured for ~1 week, yielding a total of 3×10^8 exponentially dividing cells, which were harvested by centrifugation of the culture at 3000×g for 10 min, then washed with ice-cold buffer (0.65 M sorbitol, 20 mM Tris, pH 7.5, 5 mM EDTA), and disrupted by nitrogen decompression at 600 psi (Parr Instruments Company). Alternatively, for nucleic acids extractions for their sequencing at Eurofins (Illumina DNA-Seq), Takara Bio (PacBio DNA-Seq), and Novogene (Illumina RNA-Seq), the organism was cultivated at 25 °C in an ATCC 1532 medium supplemented with 1% (v/v) horse serum and 0.1% (w/v) tryptone (MORALES *et al.* 2016).

Extraction of nucleic acids

Total cellular DNA was isolated from the disintegrated cells using the columns Genomic-tip 100/G (Qiagen). For long-read DNA sequencing, DNA was further size-selected to avoid copious mitochondrial DNA (6–7 kbp circular molecules) by employing the BluePippin instrument (Sage Science) with a 20-kbp cut-off. Total cellular RNA was extracted using a homemade Trizol substitute (RODRIGUEZ-EZPELETA *et al.* 2009), and residual DNA was removed by digestion with an RNase-free DNase followed by another round of Trizol-substitute extraction. After denaturation at 72 °C for 2 min and subsequent chilling on ice, poly(A) RNA was enriched by a passage through oligo(dT)-cellulose (Amersham). DNA and RNA samples were submitted for sequencing to the technology platforms Genome Quebec Innovation Center, Montreal. Alternatively, a commercial TRIZOL reagent (Invitrogen) and GenElute columns were used, and the resulting total RNA was submitted for mRNA enrichment and sequencing to Novogene.

A total of 10 libraries were constructed according to manufacturers' recommendations and sequenced with the Illumina or PacBio technologies. Information on library preparation and genome and transcriptome sequencing methodologies are described in the **Supplementary Information**: Section 2. Assembly and annotation of the nuclear genome and transcriptome of *Diplonema papillatum*.

REFERENCES

Morales, J., M. Hashimoto, T. A. Williams, H. Hirawake-Mogi, T. Makiuchi *et al.*, 2016 Differential remodelling of peroxisome function underpins the environmental and metabolic adaptability of diplonemids and kinetoplastids. Proc Biol Sci 283**:** 20160520.

Rodriguez-Ezpeleta, N., S. Teijeiro, L. Forget, G. Burger and B. F. Lang, 2009 Construction of cDNA libraries: focus on protists and fungi. Methods Mol Biol 533**:** 33-47.

Valach, M., S. Moreira, G. N. Kiethega and G. Burger, 2014 Trans-splicing and RNA editing of LSU rRNA in *Diplonema* mitochondria. Nucleic Acids Res 42**:** 2660-2672.

Chapter 2

Endogenous tagging of subunit composition for mitochondrial dehydrogenase complexes in marine diplonemids

1 **Subunit composition of mitochondrial dehydrogenase complexes in diplonemid flagellates**

2

3 Kristína Záhonová[1,2,3,4,#], Matus Valach[5,#], Pragya Tripathi[1,6,#], Corinna Benz[1], Fred R.
4 Opperdoes[7], Peter Barath[8,9], Veronika Lukáčová[9], Maksym Danchenko[8], Drahomíra
5 Faktorová[1,6], Anton Horváth[10], Gertraud Burger[5], Julius Lukeš[1,6,*] and Ingrid Škodová-
6 Sveráková[1,2,10,*]

7

8 [1] Institute of Parasitology, Biology Centre, Czech Academy of Sciences, České Budějovice
9 (Budweis), Czech Republic
10 [2] Life Science Research Centre, Faculty of Science, University of Ostrava, Ostrava, Czech
11 Republic
12 [3] Department of Parasitology, Faculty of Science, Charles University, BIOCEV, Vestec, Czech
13 Republic
14 [4] Division of Infectious Diseases, Department of Medicine, University of Alberta, Edmonton,
15 Canada
16 [5] Department of Biochemistry and Robert-Cedergren Centre for Bioinformatics and Genomics,
17 Université de Montréal, Montreal, Canada
18 [6] Faculty of Sciences, University of South Bohemia, České Budějovice (Budweis), Czech
19 Republic
20 [7] de Duve Institute, Université Catholique de Louvain, Brussels, Belgium
21 [8] Institute of Chemistry, Slovak Academy of Sciences, Bratislava, Slovakia
22 [9] Medirex Group Academy, Nitra, Slovakia
23 [10] Faculty of Natural Sciences, Comenius University, Bratislava, Slovakia

24

25

26 [#] These authors contributed equally to this work
27 [*] Corresponding authors: skodovaister@gmail.com; jula@paru.cas.cz
28

Abstract

In eukaryotes, pyruvate, a key metabolite produced by glycolysis, is converted by a tripartite mitochondrial pyruvate dehydrogenase (PDH) complex to acetyl-coenzyme A, which is fed into the tricarboxylic acid cycle. Two additional enzyme complexes with analogous composition catalyze similar oxidative decarboxylation reactions albeit using different substrates, the branched-chain ketoacid dehydrogenase (BCKDH) complex and the 2-oxoglutarate dehydrogenase (OGDH) complex. Comparative transcriptome analyses of diplonemids, one of the most abundant and diverse groups of oceanic protists, indicate that the conventional E1, E2, and E3 subunits of the PDH complex are lacking. E1 was apparently replaced in the euglenozoan ancestor of diplonemids by an AceE protein of archaeal type, a substitution that we also document in dinoflagellates. Here we demonstrate that the mitochondrion of the model diplonemid *Paradiplonema papillatum* displays pyruvate and 2-oxoglutarate dehydrogenase activities. Protein mass spectrometry of mitochondria reveal that the AceE protein is as abundant as the E1 subunit of BCKDH. This corroborates the view that the AceE subunit is a functional component of the PDH complex. We hypothesize that by acquiring AceE, the diplonemid ancestor not only lost the eukaryotic-type E1, but also the E2 and E3 subunits of the PDH complex, which are present in other euglenozoans. We posit that the PDH activity in diplonemids seems to be carried out by a complex, in which the AceE protein partners with the E2 and E3 subunits from BCKDH and/or OGDH.

Key words: dehydrogenase complexes; evolution; *Diplonema papillatum*; diplonemids; protist; mitochondrion.

1. Introduction

Pyruvate is made from two main energy sources, carbohydrates and amino acids. Pyruvate dehydrogenase (PDH) is a multi-enzyme complex that controls the entry of pyruvate into the tricarboxylic acid (TCA) cycle [1]. PDH not only decarboxylates pyruvate as its name suggests but transforms it into acetyl-coenzyme A (CoA), CO_2, and NADH. Two other dehydrogenase complexes, the 2-oxoglutarate dehydrogenase (OGDH) and branched-chain ketoacid dehydrogenase (BCKDH) complexes, share several structural and enzymatic properties with

PDH, and catalyze an analogous reaction, by which a 2-oxoacid is oxidatively decarboxylated [2]. The three complexes are referred to collectively as DH complexes.

All three complexes, PDH, OGDH, and BCKDH, consist of three distinct subunits referred to as the Enzyme 1, 2, and 3 (E1, E2, and E3), each having a particular catalytic function. The E1 subunit is a thiamine diphosphate-dependent enzyme with a 2-oxoacid dehydrogenase activity (also known as pyruvate, 2-oxoglutarate or ketoacid dehydrogenase) that catalyzes in a two-step reaction the irreversible decarboxylation of the substrate, and the reductive acylation of lipoyl groups covalently attached to E2. The E2 subunit, a dihydrolipoamide acyltransferase, catalyzes the transfer of an acyl moiety to CoA producing acyl-CoA. Finally, the E3 subunit, a flavoprotein with dihydrolipoamide dehydrogenase activity, transfers electrons from dihydrolipoyl groups of E2 to FAD and then to NAD^+ generating NADH + H^+ [3]. This E3 subunit is a member of the disulfide oxidoreductases family [4], which also includes glutathione reductase [5], thioredoxin reductase [6], trypanothione reductase [7], and mercuric reductase [8]. In eukaryotes, E3-binding protein (E3BP; also called Protein X) is a non-catalytic auxiliary protein of the PDH complex. This protein is structurally and functionally related to the E2 component and anchors E2 in the complex [9].

Across the tree of life, the **E1 subunit** of PDH typically consists of either a single polypeptide encoded by an *aceE* gene, as e.g., in Gram-negative bacteria, or alternatively, as encountered in eukaryotes and Gram-positive bacteria [10], of a heterotetramer composed of two proteins, 2xE1-α and 2xE1-β, encoded by *pdhA* and *pdhB* genes, respectively [11,12]. (Note that for simplicity, we follow the gene nomenclature used in bacteria.) While functionally equivalent, the E1 components of the PDH, OGDH, and BCKDH complexes can have a different composition and the homologous proteins differ in sequence. For clarity, the homologs in the various complexes will be referred to in the following as E1p-α, E1o-α, and E1b-α, etc.

The **E2 subunit** consists of a single protein encoded by *aceF* (also designated *pdhC*) and forms the core of the PDH complex, physically interacting with both the E1 and E3 subunits [13]. Again, the E2 proteins from the three DH complexes differ in sequence and will be referred to as E2p, E2o, and E2b. Lastly, the homodimeric **E3 subunit** comprises two identical polypeptide chains, designated the 'E3 proteins' (encoded by *lpdA*/*pdhD*). In contrast to E1 and E2, the nature and distribution of the E3 subunit can be more complicated, depending on the organism: in some cases, the PDH, OGDH, and BCKDH complex each contains a DH-specific

E3 [14], in others, all three DH complexes contain an identical E3 [15], while still in others, DH complexes are composed of E3 isoenzymes in various ratios [16].

Several exceptions to the conventional DH-complex structures exist. For example, Actinobacteria harbor an enzyme that carries out both E1 and E2 activities of the OGDH [17,18]. Moreover, the PDH and OGDH complexes are present in a mixed supercomplex [19]. Due to their functional similarity [20], mutual substitution of DH complexes occasionally occurs. For instance, in several apicomplexan parasites, BCKDH has taken over the function of PDH [21,22]. Furthermore, apicomplexan OGDH is capable of *in vitro* of decarboxylating both pyruvate and branched-chain ketoacids, yet with a lower preference for pyruvate [23]. In addition to the canonical DH complexes, a 2-oxoadipate complex exist that is involved in the final degradation of lysine, hydroxylysine, and tryptophan. Notably, this complex contains E2o instead of a specific E2; hence, a cross-talk between the 2-oxoadipate and OGDH complexes was suggested [24].

We recently examined the gene complement involved in basic metabolic pathways in diplonemids, an abundant, ecologically important group of marine protists [25]. It appeared that diplonemids including the well-studied representative *Paradiplonema papillatum* (originally designated *Diplonema papillatum* [26]) are lacking the genes encoding subunits of the PDH complex. To identify the subunit composition of all DH complexes in diplonemids, we mined the genome of *P. papillatum*, and the transcriptomes of ten other diplonemids, for the expected DH subunits and compared them to those present in other euglenozoans [27]. In *P. papillatum,* we also determined the subcellular localization of the candidate proteins by shotgun proteomics and measured the enzymatic activities. Lastly, we used the functional genetics approach recently established for *P. papillatum* [28] for epitope-tagging of two putative DH subunits and pull down of their interacting partners. The obtained results allowed us to propose a scenario for the evolution of the three DH complexes in diplonemids and euglenozoans as a whole.

2. Materials and methods

2.1. Sequence searches and phylogenetic analyses

Subunits of *Trypanosoma brucei* [29] and *Euglena gracilis* [30] dehydrogenase complexes were used as queries in BLAST+ v2.8.1 [31] against the genome of *Paradiplonema papillatum* [32],

transcriptomes of diplonemids *Diplonema japonicum*, *Diplonema ambulator*, *Rhynchopus humris*, *Rhynchopus euleeides*, *Lacrimia lanifica*, *Sulcionema specki*, *Flectonema neradi*, *Artemidia motanka*, *Hemistasia phaeocysticola*, and *Namystynia karyoxenos* [33,34] and euglenids *Euglena longa* [35] and *Eutreptiella gymnastica* ([36]; reassembly available at https://doi.org/10.5281/zenodo.257410), single-cell amplified transcriptomes of euglenids [37] and single-cell amplified genomes of euglenozoans [38]. For more sensitive searches, HMMER v3.3 [39] was employed. Protein domains were predicted using InterProScan [40] implemented in Geneious Prime v2020.2.5 [41]. Subcellular targeting was predicted by MitoprotII [42], MitoFates [43], and TargetP2 [44] in default settings.

Euglenozoan protein sequences were used in BLASTP searches against the NCBI non-redundant database to identify homologs from the main eukaryotic groups [45] and prokaryotes. Retrieved hits were used to build multiple sequence alignments using MAFFT v.7458 under L-INS-i strategy [46]. Poorly aligned positions were removed by trimAl v1.4 (-gt 0.8) [47] and sequences shorter than 50% of the length of the trimmed alignment were removed by an in-house python script (https://github.com/kikinocka/ngs/blob/master/py_scripts/filter_alignment.py). Resulting alignments were subjected to maximum-likelihood phylogenetic analysis by IQ-TREE v1.6.12 [48] under the LG+I+G4 (E1α, E1β, and E3) and LG+F+I+G4 (E2) models, which were determined as best-fitting according to Bayesian information criterion, and under the LG+C20+F+G model (AceE) using posterior mean site frequency method [49] with the guide tree inferred under the LG+F+G model. Branch supports were obtained by the ultrafast approximation [50] with 1,000 replicates.

2.2. Gene tagging

We attempted to create an *in situ* C-terminally-tagged version of the E3 enzyme of *P. papillatum* (DIPPA_01402), but no stable, viable clones could be produced. Hence, the E3-Protein A fusion was ectopically expressed after random integration into the genome. The open reading frame (ORF) of DIPPA_01402 was amplified from *P. papillatum* genomic DNA using primers E3_Fw and E3_Rv and cloned into pDP002 vector [28] using BglII/BamHI and NheI restriction sites. The entire 5′UTR-E3-protein A-Neo^R-3′UTR cassette was amplified from the resulting plasmid using primers pDP002_cassette_Fw and pDP002_cassette_Rv, purified, and used for transfecting *P. papillatum*. To endogenously tag AceE (DIPPA_31725) with a C-terminal Protein A-tag,

parts of the ORF (AceE_ORF_Fw and AceE_ORF_Rv) and UTR (AceE_3UTR_Fw and AceE_3UTR_Rv) were amplified from *P. papillatum* genomic DNA using primers containing sequences overlapping with the Protein A-Neo^R cassette of pDP002 [51]. Another PCR was done to amplify the protein A-Neo^R cassette from pDP002 [51] using primers Fw_protein A-Neo_cassette and Rv_protein A-Neo_cassette. To ligate of all three fragments together, they were used as templates in a nested PCR approach using primers AceE_nested_Fw and AceE_nested_Rv and Phusion polymerase (NEB) and the resulting product was A-tailed and cloned into pcr® 2.1-TOPO® (ThermoFisher). About 10 µg of the resulting plasmid was cut with EcoRI, ethanol precipitated, resuspended in 10 µl of water and used for transfecting *P. papillatum*. Primer sequences are listed in Table S1.

2.3. Strain, cultivation, and electroporation of P. papillatum

Paradiplonema papillatum (ATCC 50162; recently renamed from *Diplonema papillatum* [26]) was cultivated axenically in 36 g/l sea salt, 1 g/l tryptone and 1% (v/v) fetal bovine serum as described previously [51]. About 5x10^7 cells were transfected using an Amaxa Nucleofector II as described earlier [28,51]. The transfectants were selected at different concentrations of G418 (25 to 80 µg/ml) in 24-well plates at 27 °C. After two weeks, successful transfectants could be observed. Each clone was transferred to 20 ml medium and was cultured for 3-4 weeks prior to testing the expression of the tagged protein by immunoblotting.

2.4. Immunofluorescence assay

Roughly 20–30 ml of an exponential phase culture was centrifuged at 1,000 g for 5 minutes. Cells were resuspended in 500 µl of 4% paraformaldehyde (dissolved in sea water) and fixed for 20 minutes on Superfrost plus slides (Thermo Scientific, J1800AMNZ) at room temperature. The fixative was washed out from cells with 1x PBS. For antibody staining, cells were permeabilized in ice-cold methanol for 20 minutes. The slides were kept in a humid chamber throughout the procedure. Afterwards, the slides were washed with 1x PBS, and blocked for 45 minutes in 5.5% (w/v) fetal bovine serum in PBS-T (0.05% (v/v) Tween in 1x PBS). The blocking solution was removed, and cells were washed with 1x PBS. The rabbit anti-Protein A primary antibody (1:2,000; Sigma, P3775) diluted in 3% (w/v) BSA (Bovine serum albumin, Sigma, A4503) in PBS-T was added on slides and incubated either for 2 hours at room temperature or at 4 °C

182 overnight covered with parafilm. Next, the primary antibody was removed, and slides were
183 washed three times with PBS-T and twice with 1x PBS. AlexaFluor555-labeled goat anti-rabbit
184 secondary antibody (1:1,000; Invitrogen, A32732) was added and incubated at room temperature
185 for 1 hour in the dark, covered with parafilm. All slides were then rinsed three times with PBS-T
186 and twice with 1x PBS and coated with 4',6-diamidino-2-phenylindole (DAPI) containing the
187 antifade reagent ProlongGold (Life Technologies). Images were acquired using an Olympus
188 BX63 automated fluorescence microscope equipped with an Olympus DP74 digital camera and
189 evaluated with the cellSens Dimension software (Olympus). Since MitoTracker Red and
190 MitoTracker Green do not stain *P. papillatum* mitochondria (our observations), we employed
191 rabbit antibody against β chain of mitochondrial ATP synthase (1: 200, kindly provided by Alena
192 Zíková's laboratory; [52]) to visualize the organelle. Because both anti-ATP synthase β and anti-
193 Protein A antibodies are of rabbit origin, simultaneous imaging of the same slide was not
194 possible.
195
196 *2.5. Immunoprecipitation*
197 Rapid single-step purification of Protein A and its conjugates was used for tagged AceE and E3
198 protein purification. Two different cell lysates were prepared with two non-ionic detergents:
199 0.1% (v/v) IGEPAL CA-630 and 2% (w/v) dodecyl maltoside (DDM). Approximately $5x10^8$
200 cells expressing Protein A-tagged E3, as well as wild-type control cells were grown axenically in
201 vented flasks at 27 °C in seawater-based tryptone-rich medium (36 g/l sea salt, 1 g/l tryptone and
202 1% (v/v) fetal bovine serum) with the appropriate selection antibiotic G418 (75 µg/ml). Cells
203 were harvested by centrifugation at 1,000 g for 10 minutes. After that, cells were resuspended in
204 5 ml of ice cold 1x PBS, centrifuged again at 1,000 g for 10 minutes, and the supernatant was
205 discarded. The pellet was resuspended in ice cold 1x PBS and irradiated with 600 mJ/cm2 UV
206 light (254 nm wavelength) to covalently crosslink proteins (UV Stratalink 1800, Stratagene).
207 UV-crosslinked cells were lysed using lysis buffer (10 mM Tris (pH 6.8), 150 mM NaCl, 0.1%
208 IGEPAL CA-630, 1% (v/v) glycerol, 1x cOmplete EDTA-free protease inhibitors; Roche) and
209 passed through a 30-gauge needle several times. The cell lysate was cleared twice by
210 centrifugation at 10,000 g for 20 minutes. 75 µl IgG Sepharose 6 Fast Flow beads (Sigma) were
211 added to the cleared cell lysate and rotated at 4 °C for 2 to 3 hours. The beads were washed three
212 times with the washing buffer (10 mM Tris (pH 6.8), 250 mM NaCl, 1% (v/v) glycerol)

supplemented with 0.1% (v/v) IGEPAL CA-630, and then twice with the washing buffer without the detergent. Bound proteins were eluted from the beads with 100 µl of 0.1 M glycine (pH 3.0) by rotating for 5 minutes at room temperature and immediately mixed with 10 µl of 1 M Tris-HCl (pH 9.0). Small aliquots of input, flow through, and elution fraction were used for immunoblotting. The elution fraction was subsequently sent for mass spectrometry analysis. The same procedure was performed for Protein A-tagged AceE, except for crosslinking. DDM lysates were processed similarly, except for the crosslinking step and rotation.

2.6. Subcellular fractionation

Cultivation and fractionation were performed essentially as described earlier [53]. Briefly, cells were cultivated axenically without shaking at ~20 °C in ocean salt medium containing 33 g/l Instant Ocean Sea Salt (Instant Ocean) supplemented with 1% (v/v) horse serum (Wisent) and 0.04% (w/v) yeast extract (BioBasic). Cells were grown until the late exponential phase, harvested by centrifugation (2,000 g, 4 °C, 5 minutes), resuspended in a buffer containing 1.2 M sorbitol, 20 mM HEPES pH 7.5, 2 mM EDTA pH 8.0, and 1x cOmplete EDTA-free protease inhibitors (Roche), and then lysed in a nitrogen cavitation chamber (Parr Instrument Company) under 30-bar nitrogen pressure. The cell lysate was separated by ultracentrifugation on a two-step sucrose gradient (36% and 60%; 134,000 g, 4 °C, 60 min). The top fraction (above 36% sucrose) corresponded to the cytosol, while the fraction enriched in mitochondria was collected from the 36/60% sucrose interface. A detailed protocol is available at https://doi.org/10.17504/protocols.io.pkydkxw.

2.7. Mass spectrometry and data analysis of immunoprecipitated samples

Trypsin digestion of the eluted Protein A-tagged E3 and wildtype control samples was performed prior to liquid chromatography-tandem mass spectroscopy (LC-MS/MS) as previously described [54]. Data were processed using MaxQuant v1.6.14 [55], which incorporates the Andromeda search engine [56]. Proteins were identified by searching a custom protein sequence database of *P. papillatum* (43,871 sequences) supplemented with frequently observed contaminants. Default search parameters were employed by MaxQuant for Orbitrap analyzers with full trypsin specificity, allowing for up to two missed cleavages. Carbamidomethylation of cysteine was set as a fixed modification and oxidation of methionine and N-terminal protein acetylation were

allowed as variable modifications. The experimental design included matching between runs for biological replicates. Peptides were required to be at least seven amino acids long, with false discovery rates (FDRs) of 0.01 calculated at the levels of peptides, proteins, and modification sites based on the number of hits against the reversed sequence database. Protein quantification was done using iBAQ indices (raw intensities divided by the number of theoretical peptides) allowing comparison of protein abundances both within samples and between them. After filtering to remove any protein with less than two unique peptides and an Andromeda score of less than 20, the obtained data were processed in Perseus v1.6.14 as described previously [57].

2.8. Proteomic analysis of subcellular fractions

Proteomic analysis of *P. papillatum* subcellular fractions was performed as described earlier [58]. Briefly, aliquots of 50 µg of purified mitochondria, cytosol, and cells were lysed with 6 M urea in 50 mM triethylammonium bicarbonate pH 8 (both Sigma-Aldrich). Subsequently, four biological replicates of each sample were reduced, alkylated, and digested with trypsin (Promega; 1:50 enzyme to protein ratio). Peptides were purified on custom-made microtips with LiChroprep RP-18 25–40-µm particles (Merck-Millipore). Next, their concentration was measured by the Pierce quantitative fluorometric peptide assay (Thermo Fisher Scientific).

For liquid chromatography-coupled mass spectrometry, peptides were loaded onto a trap column (PepMap100 C18, 300 µm × 5 mm, 5-µm particle size; Dionex) and separated with an EASY-Spray C18 analytical column (75 µm × 500 mm, 5-µm particle size; Thermo Fisher Scientific) on Ultimate 3000 RSLCnano system (Dionex). The gradient of 2.4–34.4% acetonitrile was applied for 2 hours at a flow rate 250 nl/minute. Spectra were collected by Orbitrap Elite (Thermo Fisher Scientific) in the data-dependent Top15 mode. Precursors were measured in the mass range 300–1700 m/z with a resolution 120,000, fragmented by the HCD mechanism with normalized collision energy 25 and acquired at a resolution 15,000.

Datasets were processed by MaxQuant v1.6.17.0 using the Andromeda search engine as described in the previous section; however, precursor tolerance in the first search was set to 20 ppm, and 4.5 ppm in the main search upon recalibration, fragment tolerance was 20 ppm, and N-terminal protein acetylation was disallowed. The label-free quantification of proteins relied on LFQ intensities (essentially the sum of peptide intensities normalized for median peptide ratios between samples).

The statistical analysis was performed using Perseus v1.6.15.0. The Output proteinGroup table from MaxQuant was filtered and LFQ intensities $\log_2$-transformed. Proteins with less than one missing value in at least one experimental group were retained and data imputed, assuming a normal distribution. The principal component analysis confirmed excellent analytical reproducibility and ANOVA corrected by permutation test was used with a $Q \leq 0.001$. Pairwise differences were assessed by the Tukey's test with $P \leq 0.001$. Hierarchical clustering was performed on Z-score-normalized averages of LFQ intensities.

To quantitatively compare DH complex components relative to each other, LFQ intensities were normalized to the number of theoretically detectable peptides, i.e., as in the iBAQ approach [59]. The number of trypsin peptides for each DH complex protein was determined using the MS-Digest tool from the ProteinProspector v6.3.1 tool suite (http://prospector.ucsf.edu/prospector/cgi-bin/msform.cgi?form=msdigest). We used the following parameters: trypsin digest; no missed cleavage; carbamidomethyl at Cys residues as fixed modification; Met oxidation as a variable modification; minimal length of 7 amino acids; and peptide mass range from 900 to 5,000 Da; the selected mass range covered >95% of all identified peptides.

2.9. Immunoblotting

Protein samples obtained from immunoprecipitation were boiled for 5 minutes in 2x NuPAGE LDS sample buffer (Invitrogen), run on 12% SDS-PAGE, and transferred to a PVDF membrane (Amersham). After blocking with 5% (w/v) milk in PBS-T (0.05% (v/v) Tween in PBS) for at least 30 minutes at room temperature, the membrane was incubated with rabbit anti-Protein A primary antibody (1:10,000; Sigma-Aldrich, P3775) at 4 °C overnight. After three washes in PBS-T, the membrane was incubated with HRP-coupled goat anti-rabbit secondary antibody (1:1,000; Sigma-Aldrich, A21428) at room temperature for 1 hour. The membrane was then washed three times in PBS-T, and the signal was developed using Clarity Western ECL Substrate (Bio-Rad). An analogous procedure was used to test the expression of tagged proteins in whole cells. The mouse anti-α-tubulin antibody (1:10,000; Sigma-Aldrich, T9026) was used as a loading control.

The effect of the Protein A-tag on the integrity of DH complexes was monitored by their separation using Clear-Native (CN) PAGE followed by western blotting. Briefly, mitochondria

were isolated using solubilization by DDM as described previously [60], and 120 µg of total mitochondrial proteins was loaded into each well. Half of the gel was stained with Coomassie Brilliant Blue G-250 (CBB) to control for equal protein loading. The other half was transferred onto a nitrocellulose membrane overnight at 20 mA. Immunodetection with rabbit anti-Protein A primary antibody was performed as described above.

2.10. Activity measurements

Aliquots of 100 µg of mitochondrial proteins were used to assess the PDH and OGDH activity. The wild-type (WT) *P. papillatum* ATCC 50162 (or its genetically altered derivate) was grown axenically in vented flasks at 15 °C in seawater-based tryptone-rich medium (36 g/l sea salt, 1 g/l tryptone and 1% (v/v) fetal bovine serum) or tryptone-poor medium (36 g/l sea salt, 0.01 g/l tryptone, and 1% (v/v) horse serum). The pyruvate dehydrogenase activity was determined using a commercial kit (Sigma-Aldrich, MAK103) for coupled enzymatic reaction, which resulted in a colorimetric (450 nm) product proportional to the enzymatic activity.

The OGDH activity was monitored in 1 ml assay buffer (50 mM Kpi pH 7.4; 2 mM $MgSO_4$; 6 mM NAD^+; 2 mM thiamine pyrophosphate; 4 mM coenzyme A). The reaction was started by addition of 2-oxoglutarate in 20 mM concentration. Increase in NADH was monitored at 340 nm for 5 minutes. One unit of activity (U) is the amount of enzyme that generates 1 nmol of NADH per minute.

The BCKDH activity was tested in various conditions as reported in published protocols [61–63]; however, none of these conditions worked in the protist.

3. Results

3.1. Predicted composition of dehydrogenase complexes in diplonemids

To identify the components of the DH complexes in diplonemids, we searched in the genome- and transcriptome-inferred proteomes of 11 species (*P. papillatum*, *D. japonicum*, *D. ambulator*, *R. humris*, *R. euleeides*, *L. lanifica*, *S. specki*, *F. neradi*, *A. motanka*, *H. phaeocysticola*, and *N. karyoxenos*) for homologs of proteins constituting the DH subunits previously described in *Trypanosoma brucei* [29] and *Euglena gracilis* [30]. From each diplonemid species, we retrieved eight distinct E1-subunit proteins, two E2-proteins, and a single E3 protein (Table S2).

To determine the affiliation of these proteins to a particular DH complex, we performed phylogenetic analyses including sequences from organisms spanning the diversity of eukaryotes and prokaryotes (Fig. 1; Fig. S1). The phylogenetic analyses clearly distinguished clades of subunits specific for different DH complexes. All diplonemids examined appear to encode a single E3 (represented by the *P. papillatum* protein DIPPA_01402), while the other euglenozoans have several homologs of this subunit. The diplonemid E3 groups with its E3 counterparts known to be shared by OGDH and BCKDH in its closest relatives, kinetoplastids and euglenids. For E2, we found one homolog each of the OGDH- and BCKDH-specific proteins (represented in *P. papillatum* by DIPPA_17068 (E2o) and DIPPA_02535 (E2b), respectively); however, we could not identify an E2 specific for PDH (E2p). Phylogenetic analyses of E1α and E1β indicates that DIPPA_25191 and DIPPA_04229 are components of the OGDH complex, while DIPPA_70163 and DIPPA_14886 belong to the BCKDH complex. An additional homolog (DIPPA_08195) that did not cluster with any of the assigned DH complexes possibly belongs to another unknown DH complex referred to here as 'unspecified' DH. None of the diplonemid E1 sequences grouped with the PDH (E1p) sequences from other eukaryotes. Instead, we found two orthologs of prokaryotic AceE proteins. One (DIPPA_31725) is related to archaeal *AceE* and is present in all examined diplonemids. The second ortholog, which groups with the bacterial counterpart, is confined to a subgroup of Hemistasiidae (Fig. 1; Fig. S1).

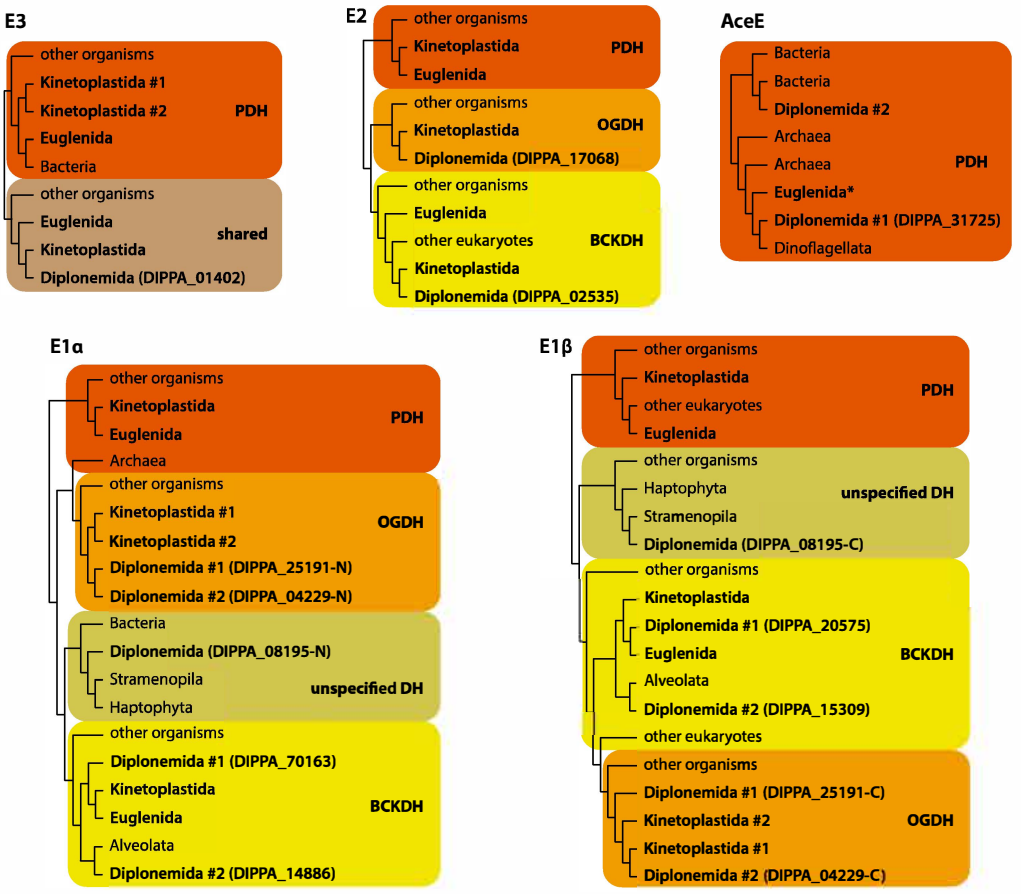


Fig. 1. Schematic phylogenetic trees of subunits of dehydrogenase complex subunits. Each tree represents a summary version of the full tree shown in Fig. S1. All presented branches were supported by >90% ultrafast bootstraps (except for the E2b kinetoplastid/diplonemid split at 45%). Euglenozoan taxa are in bold. The identifier of *Paradiplonema papillatum* proteins representing Diplonemida is indicated in parentheses. Paralogous sequences are labeled '#1' and '#2'. To analyze E1 subunit sequences irrespective of their existence as a single chain or two polypeptides, we split, prior to the construction of multiple sequence alignments, single polypeptides into their 'N' and 'C' moieties that correspond to the E1α and E1β subunits, respectively (indicated by these suffixes for the *P. papillatum* IDs). The asterisk indicates partial sequences from two *Anisonema* strains, whose eukaryotic origin remains to be confirmed by full genome or transcriptome sequencing.

Active regions and residues of E2o were previously identified in the catalytic domain of the *E. coli* protein [64], which allowed us to investigate their counterparts in the euglenozoans. We separately aligned euglenozoan E2 sequences from different phylogenetic clades with the *E. coli* sequence (Fig. S2). The corresponding regions of E2o possess the highest percentage of pairwise identity of all three regions. Moreover, two out of three *E. coli* residues responsible for the E2o's substrate specificity [64] are conserved only in the phylogenetically assigned E2o sequences from euglenozoans (Fig. S2). Collectively, these results support the protein classification obtained by our phylogenetic analyses. We also investigated the distribution of lipoyl-binding domains, the number of which was previously proposed to correlate with the DH complex type [20]. However, we found no informative correlation among the euglenozoan sequences (Table S3), suggesting that the previous conclusion [20] might have been biased by a limited sampling across the diversity of organisms available at that time.

Based on the phylogenetic analyses, we inferred the following makeup of E1 subunits across the three DH complexes from euglenozoans. In kinetoplastids and euglenids, the E1p subunit consists of two proteins, E1p-α and E1p-β, but of a single protein, an AceE homolog, in diplonemids. E1o from kinetoplastids and diplonemids is a single fusion protein composed of the E1-α and β proteins (E1o-α+β), whereas this subunit is absent from euglenids. The situation of the E1b subunit is simpler, as it consists in all euglenozoans of an α and β protein (Fig. 2). Further, the examined diplonemids encode an additional E1-α+β fusion, but the DH complex it belongs to has yet to be identified.

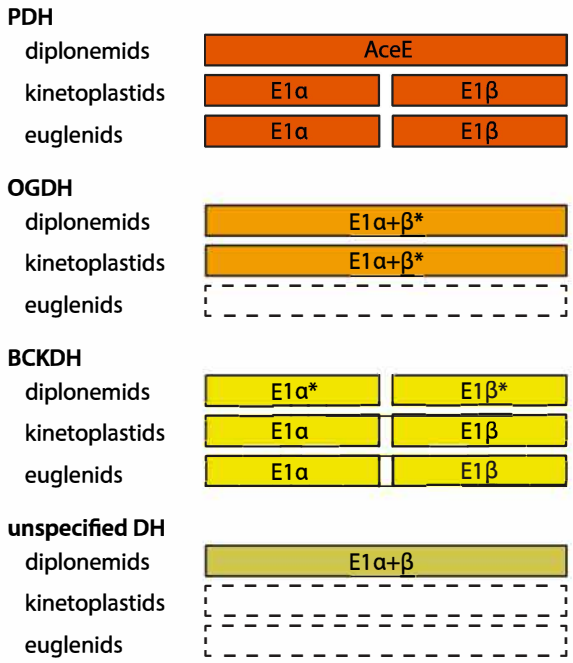


Fig. 2. The structure of E1 subunit genes in euglenozoans. Asterisks indicate the presence of two paralogs. The structure of euglenid E1p is based on that of the model *Euglena gracilis*, however, AceE was identified in *Anisomena acinus* (Fig. S1A).

We have identified in several kinetoplastid and euglenid flagellates orthologs of the non-catalytic E3BP previously described in *T. brucei* [29] (Table S4). Importantly, no E3BP homolog was found in any diplonemid species, even when more sensitive searches with a euglenozoan-specific profile hidden Markov model was employed. Altogether, our results strongly suggest that the canonical PDH complex was lost from the diplonemid lineage.

3.2. Mitochondrial localization of dehydrogenase complexes in P. papillatum

To determine the ratios of the 11 postulated components of the three DH complexes, and to verify the predicted mitochondrial localization (Table S2) of the corresponding proteins, we conducted liquid chromatography-tandem mass spectrometry (MS) of whole-cell lysates and subcellular fractions enriched in either cytosolic or mitochondrial proteins (Fig. 3; Table S5A).

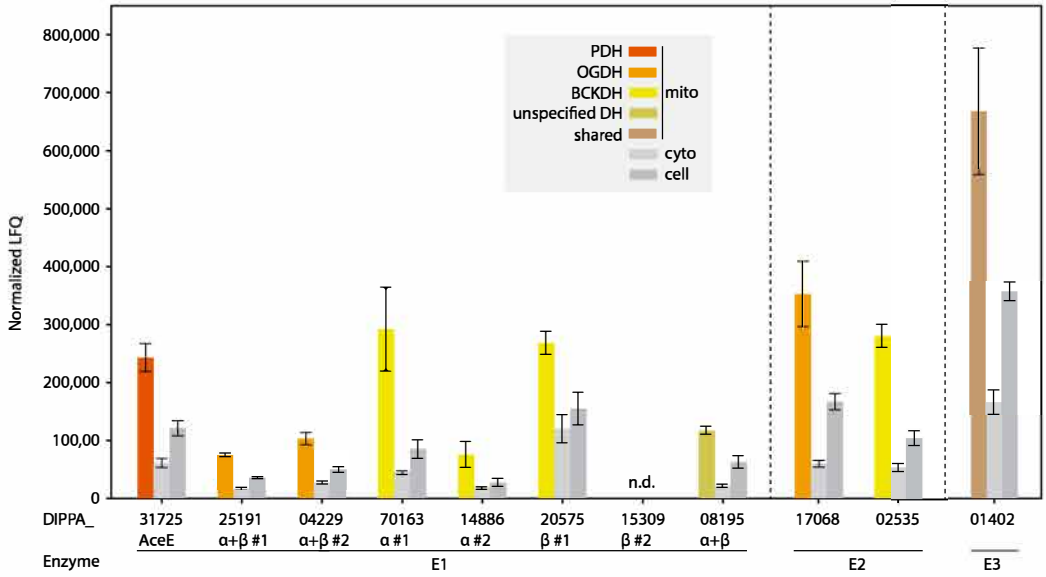


Fig. 3. Subunit enrichment of dehydrogenase complexes in cellular fractions of *P. papillatum*. The assignment of subunits to DH complexes was based on phylogenetic analyses. The label-free quantification of proteins was based on label-free quantitation (LFQ) intensities (see also Table S5). PDH, pyruvate dehydrogenase complex; OGDH, 2-oxoglutarate dehydrogenase complex; BCKDH, branched-chain ketoacid dehydrogenase complex.

Principal component analysis of the MS experiments showed satisfactory distinctiveness and reproducibility (Table S5B). For instance, the cytosolic fraction was enriched in cytosolic ribosomal proteins and translation factors, proteasome subunits, as well as metabolic enzymes involved in glycolysis and gluconeogenesis, while the mitochondrial fraction contained typical organellar proteins, such as respiratory chain complex components, carrier proteins, and prohibitins (Table S5C). All above identified subunits of the *P. papillatum* DH complexes were significantly enriched in the mitochondrial fraction, except for the inferred protein DIPPA_15309 (E1b-β #2), which was not detected in any sample (Fig. 3), even though the gene is transcribed at a level similar to that of other subunits (Table S2).

To further validate experimentally the mitochondrial localization of DH complexes, we created *P. papillatum* cell lines expressing DIPPA_31725 (AceE) and DIPPA_01402 (E3) that were C-terminally tagged with Protein A (Fig. S3). C-terminal tagging was chosen to avoid interference with mitochondrial import signals generally located at the N-terminus. The resident

AceE gene was replaced by the tagged version. However, in the case of E3, knock-in did not succeed, so that we introduced an ectopic copy of the tagged E3 gene into the genome. Immunofluorescence-based detection with anti-Protein A antibodies showed that AceE and E3 localized to the single reticulated mitochondrion (Fig. 4). As a control, the mitochondrion was immunolabeled with an antibody against the β chain of the mitochondrial ATP synthase. The staining pattern observed using the anti-ATP synthase antibody strongly resembled the signals obtained with AceE and E3, confirming the predicted mitochondrial localization of these two DH-complex components.

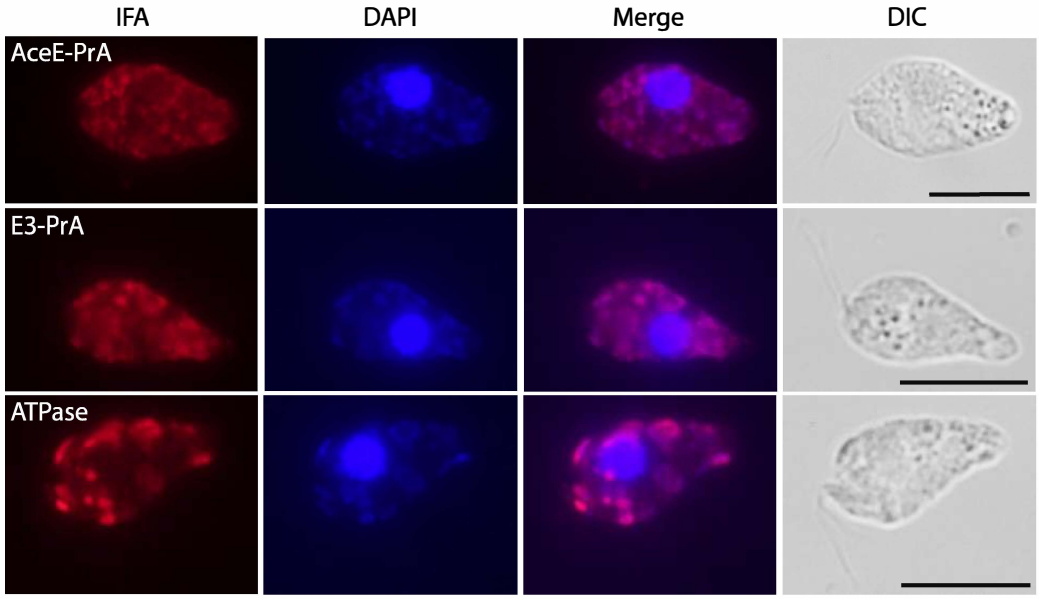


Fig. 4. Subcellular localization of AceE and E3 proteins in *P. papillatum*. Immunofluorescence assay (IFA) of Protein A-tagged AceE (first row) and E3 (second row) proteins using polyclonal anti-Protein A antibodies (red) confirmed their mitochondrial localization. Rabbit-antibodies against mitochondrial ATP synthase (third row) were used for comparison (see Methods). DNA was stained with DAPI (blue). The scale bar is 10 μm. DIC, differential interference contrast.

3.3. Purification of DH complexes from P. papillatum

To evaluate whether the Protein A-tagged subunits E3 and AceE assembled into large complexes, we extracted mitochondrial proteins, separated them *via* CN-PAGE (2–10% gradient gel), and performed immunoblotting. We observed signals at a high molecular weight (upper area of the gel; Fig.5), which indicated the incorporation of the tagged proteins into complexes. However, the vast majority of the Protein A-tagged AceE was detected as a smear of varied molecular weights centered around 600 kDa, which suggested disassembly and/or degradation of a much larger complex (presumably PDH) complex during protein extraction. The Protein A-tagged E3 displayed a less prominent smear of apparent disassembly products but was also detected as a distinct signal in the low molecular-weight region (presumably as a monomer) (Fig. 5).

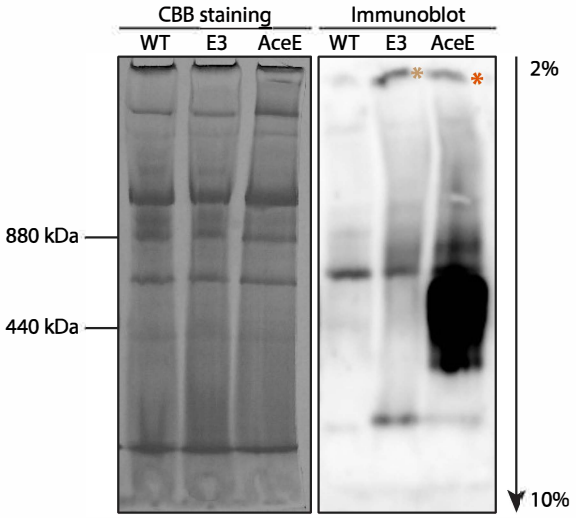


Fig. 5. Immunodetection of complexes *via* Protein A-tagged E3 and AceE on native gel. Coomassie brilliant blue (CBB) staining (left panel) served as a loading control. The right panel shows the immunoblotting of proteins using anti-Protein A antibodies. Asterisks indicate large complexes, while the signals in the lower molecular-weight region represent disassembly and/or degradation products and unspecific epitopes.

To investigate the composition the three DH complexes from *P. papillatum* were pulled down using the C-terminally-tagged AceE and E3 proteins and analyzed by MS. We reasoned that if the E3 protein was part of all three DH complexes, it would pull down all complexes,

whereas AceE, which we expected to be an exclusive component of the PDH complex, would only pull down this single complex. Immunoblotting of the fraction pulled down *via* the E3 protein confirmed that the eluate was enriched in the bait protein (Fig. 6). MS analysis of this fraction revealed three proteins strongly associated with the E3 protein, which were all predicted components of the OGDH complex, namely DIPPA_25191 (E1o-α+β #1), DIPPA_04229 (E1o-α+β #2), and DIPPA_17068 (E2o). E2b (DIPPA_02535) was also detected but at low abundance and below the significance threshold (Fig. 5), while subunits of the PDH complex were absent from the eluate. In a separate set of experiments, we immunoprecipitated the tagged AceE, but MS analysis did not uncover any interactors at significant amount and enrichment. Essentially identical results were obtained when the detergent used for the pull-down experiments, octylphenoxypolyethoxyethanol (IGEPAL), was replaced by n-dodecyl-beta-D-maltoside (DDM) (for details, see Table S6).

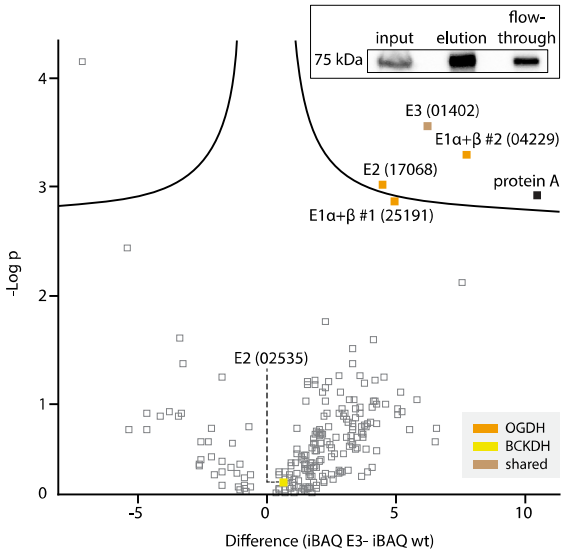


Fig. 6. Immunopurification of E3. Volcano plot of the $-\log_{10}$ p value of the t-test plotted against the intensity-based absolute quantification (iBAQ) difference between the samples of the Protein A-tagged E3 subunit and wildtype (wt) control, based on three biological replicates. Statistically significant hits are placed above the curve in the top right quadrant; their identity is indicated. Hits are color-coded according

to the complex they belong to. The inset shows an immunoblot of a representative immunoprecipitation; the blot was probed with anti-Protein A antibodies to detect tagged E3.

3.4. Activity measurements

To show that PDH and OGDH are indeed present in *P. papillatum*, we measured the activities of these complexes in mitochondria of the WT and tagged cell lines (Table 1). The presence of the tag did not inactivate either the PDH or the OGDH complex. Due to technical difficulties, we could not evaluate the effect of the Protein A-tagged E3 on the activity of the BCKDH complex.

Table 1. Activity measurements of *Paradiplonema papillatum* mitochondrial lysates[a].

	PDH [U/mg]	OGDH [U/mg]	BCKDH [U/mg]
WT	14.7 ± 5.6	3.3 ± 1.1	n.d.
E3	13.5 ± 5.0	5.2 ± 1.6	n.d.
AceE	10.5 ± 3.6	5.9 ± 0.7	n.d.

[a] One unit of activity (U) is the amount of enzyme that generates 1 nmol of NADH per minute. Activities were calculated from 2-3 biological replicates. n.d., not determined

4. Discussion

4.1. The E3 protein is likely a shared component of all three DH complexes in diplonemids

The PDH, OGDH, and BCKDH complexes share mechanistic and structural similarities but carry out distinct enzymatic reactions in the cell. While traditionally viewed as highly conserved across all domains of life, recent studies have shown that in some organisms a given subunit of one complex can mix with and substitute its homologous counterpart from another complex [15], leading to hybrid complexes [65]. The most widespread example of such an architectural plasticity is the E3 protein, which, in many bacteria (e.g., *E. coli*) and eukaryotes (e.g., land plants, animals, euglenozoans) is a constituent of two or even three different DH complexes [16,66].

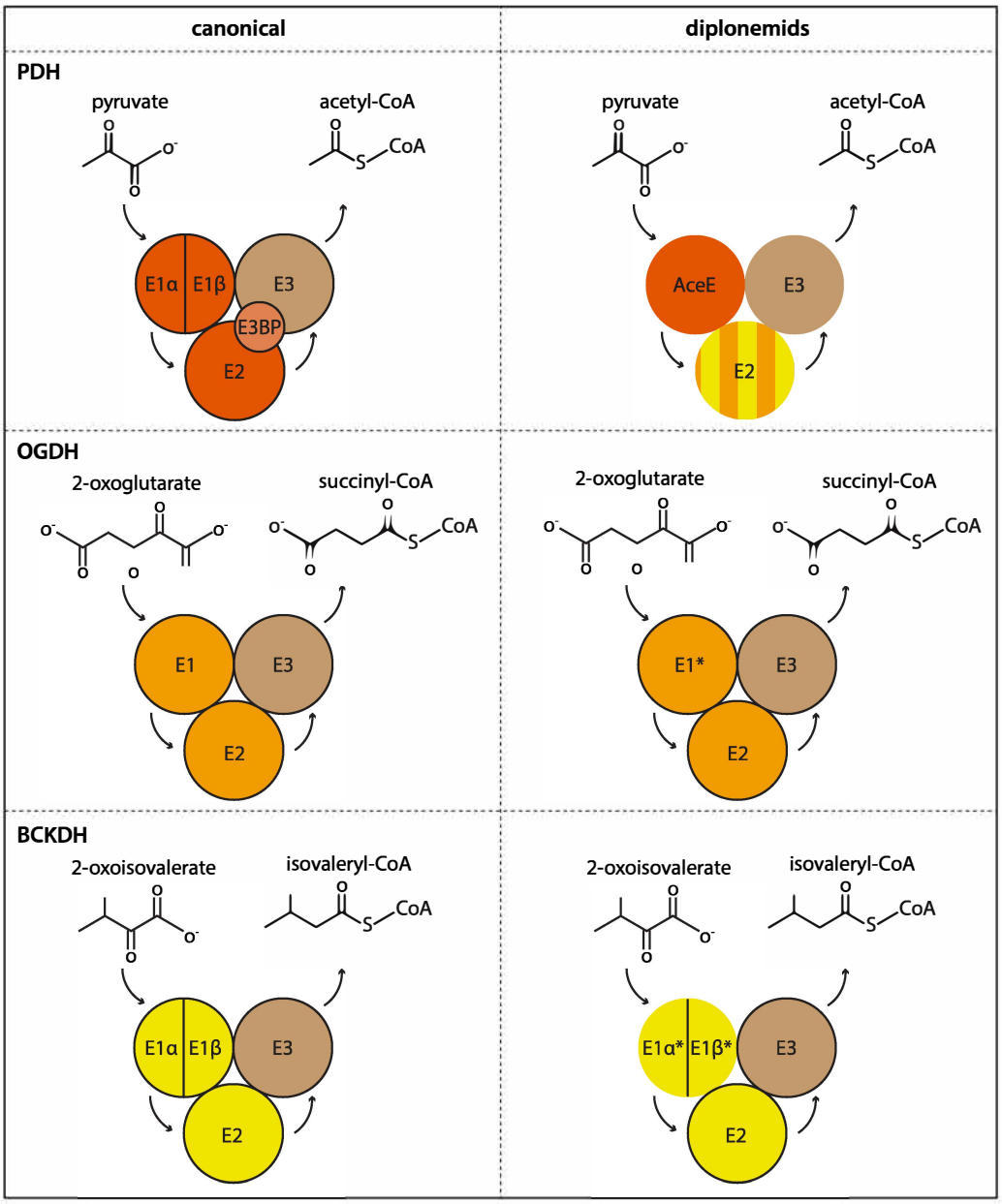
Unexpectedly, the tagged E3 protein of *P. papillatum* pulled down exclusively the OGDH complex. From the BCKDH complex, we detected only insignificant traces, and from the PDH complex not a single protein. Still, several lines of evidence indicate that the E3 protein is not exclusively associated with the OGDH complex. First, in isolated mitochondria of *P.*

513 *papillatum*, the E3 protein is almost twice as abundant as E2o (Fig. 3). Thus, the 'surplus' E3
514 could be part of the elusive PDH complex. This is corroborated by the fact that complexes have
515 generally several times less copies of E3 than of E1 and E2. Still, in an individual DH complex,
516 the ratio between the subunits E1, E2, and E3 can fluctuate *in vivo*, and the architecture and the
517 size of the complexes vary among organisms [19,67,68]. In mammals and fungi, for example,
518 individual PDH complexes have E1:E2:E3 ratios of ~30:60:6 and ~20:60:12, respectively
519 [68,69]. The reason for this variability is unclear.
520 Second, our measurements of the pyruvate dehydrogenase activity from *P. papillatum* in
521 isolated mitochondria suggests the presence of a PDH complex. However, we cannot exclude as
522 discussed later that the BCKDH complex catalyzes pyruvate decarboxylation.
523 Third, our CN-PAGE separations indicated that E3 and AceE-containing complexes of *P.*
524 *papillatum* are unstable. This might be an inherent feature of these complexes because it was
525 shown that the PDH complex from human and yeast easily disintegrate during isolation [70,71].
526 Alternatively, the tags that we appended to the AceE and E3 proteins may have destabilized the
527 complexes of which these proteins are part. Although tagged E3 seems to be readily assembled
528 into the OGDH complex, the tag could obstruct E3 integration into the other two complexes.
529 Detection of all BCKDH proteins (except for the second paralog of E1b-β) and the AceE
530 homolog in the mitochondrial fraction strongly suggests that both the OGDH and BCKDH
531 complexes correctly assemble in wild-type cell lines and do comprise E3, even though direct
532 experimental evidence for the existence of these complexes is yet to be provided.
533
534 *4.2. Structure of the hypothetical diplonemid PDH complex: the E2 subunit has likely been*
535 *replaced*
536 Our extensive sequence searches failed to retrieve an E2p homolog from any diplonemid genome
537 or transcriptome, which suggests that in contrast to the vast majority of organisms, diplonemids
538 genuinely lack a PDH-specific E2. We posit that E2p is substituted by E2 from either OGDH or
539 BCKDH, a situation not without precedents. In humans, for instance, the E2o and E3o subunits
540 associate not only with the E1o protein to form the typical OGDH, but also with the
541 evolutionarily much younger E1a protein that is specific for the 2-oxoadipate DH complex [24].
542 Similarly, in the bacterium *Corynebacterium glutamicum*, the PDH and OGDH complexes have
543 the same E2 and E3 subunits but distinct E1p and E1o subunits [65].

544 It is currently unclear whether the presumed PDH complex in diplonemids comprises E2o
545 or E2b, but hints come from *in vitro* activity tests of the three DH complexes, as well as
546 instances of complete loss of the PDH complex in certain organisms. First, the inability of
547 purified mammalian PDH to oxidize 2-oxo-glutarate *in vitro* suggests that the high substrate
548 specificity prevents this complex from substituting the activity of OGDH [71]. In contrast,
549 BCKDH from *Bacillus subtilis* is capable to decarboxylate pyruvate (but not 2-oxoglutarate) *in*
550 *vitro*, albeit with lower specificity and efficacy than branched-chain ketoacids [72]. Lastly, in the
551 apicomplexans *Toxoplasma gondii* and *Plasmodium falciparum*, the absence of PDH is fully
552 compensated for by BCKDH, but not OGDH [22]. All this suggests that the BCKDH complex is
553 more permissive to utilize pyruvate as a substrate.
554 One possibility is that similarly to apicomplexans [22], it is the BCKDH complex that
555 carries out the mitochondrial PDH activity in *P. papillatum*. However, this scenario is not very
556 likely, because apicomplexans lack all proteins otherwise present in the PDH complex, whereas
557 diplonemids possess a dedicated E1p in the form of AceE, strongly suggesting that a functional
558 PDH complex is indeed present in the latter group. We consider the E2b of the BCKDH complex
559 as the most promising E2p candidate of the PDH complex, because the former complex can
560 facilitate the PDH activity [22]. But an investigation in which E2 subunit associates with the
561 diplonemid PDH complex, requires experimental approaches that, in contrast to protein tagging,
562 fully maintain the native complex structure. One possibility would be the less disruptive
563 purification on AMP-Sepharose, which is, however, complicated by the high background of
564 NAD^+-dependent dehydrogenases and ATP dependent kinases (our unpublished data).
565
566 *4.3. Structure of the hypothetical diplonemid PDH complex: recruitment of an alien E1p*
567 Typically, the E1p subunit of eukaryotes is made up of E1p-α and E1p-β proteins [73]. The
568 corresponding genes are encoded in the genomes of various euglenozoans [29,30] – except for
569 diplonemids. Apparently the diplonemid E1p heterotetramer was replaced with a prokaryotic
570 AceE protein. Although we failed to identify experimentally the protein partners interacting with
571 AceE, we showed that its Protein-A-tagged version assembles into high molecular-weight
572 complexes, although these are very fragile under all tested isolation conditions. Importantly, the
573 relatively high abundance of this component in mitochondria attests to its functional importance.

All diplonemids possess an AceE homolog, and we also discovered the gene in the single-cell transcriptomes of the euglenid *Anisonema*, but not in *E. gracilis* and *E. longa* (Fig. S1A). As euglenids other than *Euglena* spp. are poorly represented in public sequence repositories, it is plausible that *Anisonema* is not an exception in this protist lineage. Still, it needs to be confirmed by genomic and/or transcriptomic data that the partial *Anisonema* AceE sequences do not represent a contamination. If truly of euglenid origin, we can envisage two scenarios. The *AceE* gene was i) acquired either early on by the last common ancestor of euglenozoans with the subsequent loss in the lineage leading to kinetoplastids and certain euglenid lineages; or ii) several times independently from a related sources by the last common ancestor of diplonemids and by an euglenid lineage. Taxonomically broader sampling will be necessary to solve this question.

The picture is further complicated by an AceE homolog present in essentially all examined dinoflagellates [11], a taxon unrelated to euglenozoans. Our phylogenetic analysis indicates that the ancestor of dinoflagellates acquired the archaeal type of *AceE via* horizontal gene transfer from an euglenozoan. Both groups are marine protists inhabiting the same oceanic depths (Aleš Horák, pers. commun.), making this scenario highly plausible. Curiously, dinoflagellates also appear to have lost the E2p subunit, but retained its E2o and E2b counterparts (Fig. S1D) [11]. Such a convergence of gene repertoire in unrelated lineages strongly suggests that the original E2p was incompatible with the newly acquired E1p substitute in the form of AceE and thus another E2 subunit filled in its role (Fig. 7).



596

597 **Fig. 7. Comparison of canonical and inferred diplonemid dehydrogenase complexes.** A PDH-specific
598 E2 subunit was not identified in diplonemids. We assume that it is functionally substituted by E2o and/or
599 E2b. Subunits experimentally confirmed in *P. papillatum* are indicated by black borders. The scheme is
600 only illustrative and does not reflect the subunit ratios found in the mitochondrion. The lighter shade of

E3BP indicates that this subunit is present exclusively in eukaryotes, whereas other components are common to all organisms. Asterisks indicate that two paralogs are present. PDH, pyruvate dehydrogenase complex; E3BP, E3-binding protein; OGDH, 2-oxoglutarate dehydrogenase complex; BCKDH, branched-chain ketoacid dehydrogenase complex.

Surveying the distribution of DH components across poorly studied Euglenozoa will likely shed light on whether the acquisition of *AceE* is invariably accompanied by the loss of E2p and replacement by E2o or E2b, as seen in diplonemids and dinoflagellates. Such studies also promise to provide insight into the prerequisites and consequences of integrating alien components into a critically important enzyme complex.

CRediT authorship contribution statement

Kristína Záhonová: Conceptualization, methodology, investigation, writing – original draft, visualization. **Matus Valach:** Methodology, investigation, writing – original draft, visualization. **Pragya Tripathi:** Methodology, investigation, writing – review & editing. **Corinna Benz:** Methodology, investigation, writing – review & editing, visualization. **Fred R. Opperdoes:** Conceptualization, investigation, writing – review & editing, funding acquisition. **Peter Barath:** Methodology, investigation, writing – review & editing, funding acquisition. **Veronika Lukáčová:** Methodology, investigation, writing – review & editing. **Maksym Danchenko:** Methodology, investigation, writing – review & editing. **Drahomíra Faktorová:** Methodology, investigation, writing – review & editing, visualization. **Anton Horváth:** Writing – review & editing, funding acquisition. **Gertraud Burger:** Investigation, writing – review & editing, funding acquisition. **Julius Lukeš:** Writing – review & editing, project administration, funding acquisition. **Ingrid Škodová-Sveráková:** Conceptualization, methodology, investigation, writing – original draft, visualization, project administration, funding acquisition.

Declaration of competing interest

The authors declare that they have no conflicts of interest with the contents of this article.

Acknowledgements

We thank Aleš Horák (Institute of Parasitology) for sharing unpublished data. This work was supported by grants from the Czech Grant Agency (23-06479X to J.L.), the Gordon and Betty Moore Foundation (#9354 to J.L.), the ERD project (16_019/0000759 and 313011W428 to J.L.), and e-INFRA CZ (ID:90140), the de Duve Institute (to F.R.O.), Scientific Grant Agency of the Slovak Ministry of Education and the Academy of Sciences (VEGA 1/0553/21 to A.H. and 1/0781/19 to I.Š.S.), Slovak Research and Development Agency Contracts (APVV-20-0129 to A.H.), the ACCORD project co-financed by the ERDF (ITMS2014+: 313021X329), the ERDF project Center for Biomedical Research – BIOMEDIRES – II. stage (313011W428 to P.B.), the Natural Sciences and Engineering Research Council of Canada (NSERC; grant RGPIN-2019-04024 to G.B.), and the Fonds de Recherche du Québec—Nature et Technologies (FRQNT; grant 2023-PR-326068 to G.B.).

Data availability

The mass spectrometry proteomics data were deposited to the ProteomeXchange Consortium *via* the PRIDE partner repository [74] with the dataset identifiers PXD035043 (E3 IP) and PXD035104 (subcellular fractions; 10.6019/PXD035104).

References

[1] E. Meléndez-Hevia, T.G. Waddell, M. Cascante, The puzzle of the Krebs citric acid cycle: Assembling the pieces of chemically feasible reactions, and opportunism in the design of metabolic pathways during evolution, J. Mol. Evol. 43 (1996) 293–303. https://doi.org/10.1007/BF02338838.

[2] H. Kagamiyama, H. Hayashi, Branched-chain amino-acid aminotransferase of *Escherichia coli*, Methods Enzymol. 324 (2000) 103–113. https://doi.org/10.1016/s0076-6879(00)24223-7.

[3] J. Škerlová, J. Berndtsson, H. Nolte, M. Ott, P. Stenmark, Structure of the native pyruvate dehydrogenase complex reveals the mechanism of substrate insertion, Nat. Commun. 12 (2021) 5277. https://doi.org/10.1038/s41467-021-25570-y.

[4] S.S. Mande, S. Sarfaty, M.D. Allen, R.N. Perham, W.G.J. Hol, Protein-protein interactions in the pyruvate dehydrogenase multienzyme complex: Dihydrolipoamide dehydrogenase complexed with the binding domain of dihydrolipoamide acetyltransferase, Structure. 4 (1996) 277–286. https://doi.org/10.1016/S0969-2126(96)00032-9.

[5] G.E. Schulz, R.H. Schirmer, W. Sachsenheimer, E.F. Pai, The structure of the flavoenzyme glutathione reductase, Nature. 273 (1978) 120–124. https://doi.org/10.1038/273120a0.

[6] J. Kuriyan, T.S.R. Krishna, L. Wong, B. Guenther, A. Pahler, C.H. Williams, P. Model, Convergent evolution of similar function in two structurally divergent enzymes, Nature. 352 (1991) 172–174. https://doi.org/10.1038/352172a0.

[7] J. Kuriyan, X.P. Kong, T.S.R. Krishna, R.M. Sweet, N.J. Murgolo, H. Field, A. Cerami, G.B. Henderson, X-ray structure of trypanothione reductase from *Crithidia fasciculata* at 2.4-Å resolution, Proc. Natl. Acad. Sci. U. S. A. 88 (1991) 8764–8768. https://doi.org/10.1073/pnas.88.19.8764.

[8] N. Schiering, W. Kabsch, M.J. Moore, M.D. Distefano, C.T. Walsh, E.F. Pai, Structure of the detoxification catalyst mercuric ion reductase from *Bacillus* sp. strain RC607, Nature. 352 (1991) 168–172. https://doi.org/10.1038/352168a0.

[9] S.J. Sanderson, S.S. Khan, R.G. McCartney, C. Miller, J.G. Lindsay, Reconstitution of mammalian pyruvate dehydrogenase and 2-oxoglutarate dehydrogenase complexes: Analysis of protein X involvement and interaction of homologous and heterologous dihydrolipoamide dehydrogenases, Biochem. J. 319 (1996) 109–116. https://doi.org/10.1042/bj3190109.

[10] X.Y. Pei, C.M. Titman, R.A.W. Frank, F.J. Leeper, B.F. Luisi, Snapshots of catalysis in the E1 subunit of the pyruvate dehydrogenase multienzyme complex, Structure. 16 (2008) 1860–1872. https://doi.org/10.1016/j.str.2008.10.009.

[11] E.R. Butterfield, C.J. Howe, R.E.R. Nisbet, An analysis of dinoflagellate metabolism using EST data, Protist. 164 (2013) 218–236. https://doi.org/10.1016/j.protis.2012.09.001.

[12] M.E. Schreiner, D. Fiur, J. Holátko, M. Pátek, B.J. Eikmanns, E1 enzyme of the pyruvate dehydrogenase complex in *Corynebacterium glutamicum*: Molecular analysis of the gene and phylogenetic aspects, J. Bacteriol. 187 (2005) 6005–6018. https://doi.org/10.1128/JB.187.17.6005-6018.2005.

[13] M. Rahmatullah, S. Gopalakrishnan, P.C. Andrews, C.L. Chang, G.A. Radke, T.E. Roche, Subunit associations in the mammalian pyruvate dehydrogenase complex. Structure and role of protein X and the pyruvate dehydrogenase component binding domain of the dihydrolipoyl transacetylase component, J. Biol. Chem. 264 (1989) 2221–2227.

[14] J.R. Sokatch, V. McCully, C.M. Roberts, Purification of a branched-chain keto acid dehydrogenase from *Pseudomonas putida*, J. Bacteriol. 148 (1981) 647–652. https://doi.org/10.1128/jb.148.2.647-652.1981.

[15] N.S. Nemeria, G. Gerfen, P.R. Nareddy, L. Yang, X. Zhang, M. Szostak, F. Jordan, The mitochondrial 2-oxoadipate and 2-oxoglutarate dehydrogenase complexes share their E2 and E3 components for their function and both generate reactive oxygen species, Free Radic. Biol. Med. 115 (2018) 136–145. https://doi.org/10.1016/j.freeradbiomed.2017.11.018.

[16] A.H. Millar, S.A. Hill, C.J. Leaver, Plant mitochondrial 2-oxoglutarate dehydrogenase complex: Purification and characterization in potato, Biochem. J. Pt 2 (1999) 327–334. https://doi.org/10.1042/0264-6021:3430327.

[17] M. Hoffelder, K. Raasch, J. Van Ooyen, L. Eggeling, The E2 domain of OdhA of *Corynebacterium*

glutamicum has succinyltransferase activity dependent on lipoyl residues of the acetyltransferase AceF, J. Bacteriol. 192 (2010) 5203–5211. https://doi.org/10.1128/JB.00597-10.

[18] A. Niebisch, A. Kabus, C. Schultz, B. Weil, M. Bott, Corynebacterial protein kinase G controls 2-oxoglutarate dehydrogenase activity via the phosphorylation status of the OdhI protein, J. Biol. Chem. 281 (2006) 12300–12307. https://doi.org/10.1074/jbc.M512515200.

[19] E.M. Bruch, P. Vilela, L. Yang, A. Boyko, N. Lexa-Sapart, B. Raynal, P.M. Alzari, M. Bellinzoni, Actinobacteria challenge the paradigm: A unique protein architecture for a well-known, central metabolic complex, Proc. Natl. Acad. Sci. U. S. A. 118 (2021) e2112107118. https://doi.org/10.1073/pnas.2112107118.

[20] R.N. Perham, Swinging arms and swinging domains in multifunctional enzymes: Catalytic machines for multistep reactions, Annu. Rev. Biochem. 69 (2000) 961–1004. https://doi.org/10.1146/annurev.biochem.69.1.961.

[21] F. Seeber, J. Limenitakis, D. Soldati-Favre, Apicomplexan mitochondrial metabolism: a story of gains, losses and retentions, Trends Parasitol. 24 (2008) 468–478. https://doi.org/10.1016/j.pt.2008.07.004.

[22] R.D. Oppenheim, D.J. Creek, J.I. Macrae, K.K. Modrzynska, P. Pino, J. Limenitakis, V. Polonais, F. Seeber, M.P. Barrett, O. Billker, M.J. McConville, D. Soldati-Favre, BCKDH: The missing link in apicomplexan mitochondrial metabolism is required for full virulence of *Toxoplasma gondii* and *Plasmodium berghei*, PLoS Pathog. 10 (2014) e1004263. https://doi.org/10.1371/journal.ppat.1004263.

[23] X.W.A. Chan, C. Wrenger, K. Stahl, B. Bergmann, M. Winterberg, I.B. Müller, K.J. Saliba, Chemical and genetic validation of thiamine utilization as an antimalarial drug target, Nat. Commun. 4 (2013) 2060. https://doi.org/10.1038/ncomms3060.

[24] N.S. Nemeria, G. Gerfen, L. Yang, X. Zhang, F. Jordan, Evidence for functional and regulatory cross-talk between the tricarboxylic acid cycle 2-oxoglutarate dehydrogenase complex and 2-oxoadipate dehydrogenase on the L-lysine, L-hydroxylysine and L-tryptophan degradation pathways from studies in vitro, Biochim. Biophys. Acta - Bioenerg. 1859 (2018) 932–939. https://doi.org/10.1016/j.bbabio.2018.05.001.

[25] A. Butenko, F.R. Opperdoes, O. Flegontova, A. Horák, V. Hampl, P. Keeling, R.M.R. Gawryluk, D. Tikhonenkov, P. Flegontov, J. Lukeš, Evolution of metabolic capabilities and molecular features of diplonemids, kinetoplastids, and euglenids, BMC Biol. 18 (2020) 23. https://doi.org/10.1186/s12915-020-0754-1.

[26] D. Tashyreva, A.G.B. Simpson, G. Prokopchuk, I. Škodová-Sveráková, A. Butenko, M. Hammond, E.E. George, O. Flegontova, K. Záhonová, D. Faktorová, A. Yabuki, A. Horák, P.J. Keeling, J. Lukeš, Diplonemids – A review on "new" flagellates on the oceanic block, Protist. 173 (2022) 125868. https://doi.org/10.1016/j.protis.2022.125868.

[27] A.Y. Kostygov, A. Karnkowska, J. Votýpka, D. Tashyreva, K. Maciszewski, V. Yurchenko, J. Lukeš, Euglenozoa: taxonomy, diversity and ecology, symbioses and viruses., Open Biol. 11 (2021) 200407. https://doi.org/10.1098/rsob.200407.

[28] D. Faktorová, B. Kaur, M. Valach, L. Graf, C. Benz, G. Burger, J. Lukeš, Targeted integration by homologous recombination enables *in situ* tagging and replacement of genes in the marine microeukaryote *Diplonema papillatum*, Environ. Microbiol. 22 (2020) 3660–3670. https://doi.org/10.1111/1462-2920.15130.

[29] A.K. Panigrahi, A. Zíková, R.A. Dalley, N. Acestor, Y. Ogata, A. Anupama, P.J. Myler, K.D. Stuart, Mitochondrial complexes in *Trypanosoma brucei*: a novel complex and a unique oxidoreductase complex, Mol. Cell. Proteomics. 7 (2008) 534–545. https://doi.org/10.1074/mcp.m700430-mcp200.

[30] T.E. Ebenezer, M. Zoltner, A. Burrell, A. Nenarokova, A.M.G. Novák Vanclová, B. Prasad, P. Soukal, C. Santana-Molina, E. O'Neill, N.N. Nankissoor, N. Vadakedath, V. Daiker, S. Obado, S. Silva-Pereira, A.P. Jackson, D.P. Devos, J. Lukeš, M. Lebert, S. Vaughan, V. Hampl, M. Carrington, M.L. Ginger, J.B. Dacks, S. Kelly, M.C. Field, Transcriptome, proteome and draft genome of *Euglena gracilis*, BMC Biol. 17 (2019) 11. https://doi.org/10.1186/s12915-019-0626-8.

[31] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410. https://doi.org/https://doi.org/10.1016/S0022-2836(05)80360-2.

[32] M. Valach, S. Moreira, C. Petitjean, C. Benz, A. Butenko, O. Flegontova, A. Nenarokova, G. Prokopchuk, T. Batstone, P. Lepébie, L. Lemogo, M. Sarrasin, P. Stretenowich, P. Tripathi, E. Yazaki, T. Nara, B. Henrissat, B.F. Lang, M.W. Gray, T.A. Williams, J. Lukeš, G. Burger, Recent expansion of metabolic versatility in *Diplonema papillatum*, the model species of a highly speciose group of marine eukaryotes, BMC Biol. 21 (2023) 99. https://doi.org/10.1186/s12915-023-01563-9.

[33] M. Valach, S. Moreira, S. Hoffmann, P.F. Stadler, G. Burger, Keeping it complicated: Mitochondrial genome plasticity across diplonemids, Sci. Rep. 7 (2017) 14166. https://doi.org/10.1038/s41598-017-14286-z.

[34] B. Kaur, K. Záhonová, M. Valach, D. Faktorová, G. Prokopchuk, G. Burger, J. Lukeš, Gene fragmentation and RNA editing without borders: Eccentric mitochondrial genomes of diplonemids, Nucleic Acids Res. 48 (2020) 2694–2708. https://doi.org/10.1093/nar/gkz1215.

[35] K. Záhonová, Z. Füssy, E. Birčák, A.M.G. Novák Vanclová, V. Klimeš, M. Vesteg, J. Krajčovič, M. Oborník, M. Eliáš, Peculiar features of the plastids of the colourless alga *Euglena longa* and photosynthetic euglenophytes unveiled by transcriptome analyses, Sci. Rep. 8 (2018) 17012. https://doi.org/10.1038/s41598-018-35389-1.

[36] P.J. Keeling, F. Burki, H.M. Wilcox, B. Allam, E.E. Allen, L.A. Amaral-Zettler, E. V Armbrust, J.M. Archibald, A.K. Bharti, C.J. Bell, B. Beszteri, K.D. Bidle, C.T. Cameron, L. Campbell, D.A. Caron, R.A. Cattolico, J.L. Collier, K. Coyne, S.K. Davy, P. Deschamps, S.T. Dyhrman, B. Edvardsen, R.D. Gates, C.J. Gobler, S.J. Greenwood, S.M. Guida, J.L. Jacobi, K.S. Jakobsen, E.R. James, B. Jenkins, U. John, M.D. Johnson, A.R. Juhl, A. Kamp, L.A. Katz, R. Kiene, A. Kudryavtsev, B.S. Leander, S. Lin, C. Lovejoy, D. Lynn, A. Marchetti, G. McManus, A.M. Nedelcu, S. Menden-Deuer, C. Miceli, T. Mock, M. Montresor, M.A. Moran, S. Murray, G. Nadathur, S. Nagai, P.B. Ngam, B. Palenik, J. Pawlowski, G. Petroni, G. Piganeau, M.C. Posewitz, K. Rengefors, G. Romano, M.E. Rumpho, T. Rynearson, K.B. Schilling, D.C.

772 Schroeder, A.G. Simpson, C.H. Slamovits, D.R. Smith, G.J. Smith, S.R. Smith, H.M. Sosik, P. Stief, E.
773 Theriot, S.N. Twary, P.E. Umale, D. Vaulot, B. Wawrik, G.L. Wheeler, W.H. Wilson, Y. Xu, A. Zingone,
774 A.Z. Worden, The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating
775 the functional diversity of eukaryotic life in the oceans through transcriptome sequencing, PLoS Biol. 12
776 (2014) e1001889. https://doi.org/10.1371/journal.pbio.1001889.
777 [37] G. Lax, M. Kolisko, Y. Eglit, W.J. Lee, N. Yubuki, A. Karnkowska, B.S. Leander, G. Burger, P.J. Keeling,
778 A.G.B. Simpson, Multigene phylogenetics of euglenids based on single-cell transcriptomics of diverse
779 phagotrophs, Mol. Phylogenet. Evol. 159 (2021) 107088. https://doi.org/10.1016/j.ympev.2021.107088.
780 [38] K. Záhonová, G. Lax, S.D. Sinha, G. Leonard, T.A. Richards, J. Lukeš, J.G. Wideman, Single-cell genomics
781 unveils a canonical origin of the diverse mitochondrial genomes of euglenozoans, BMC Biol. 19 (2021) 103.
782 https://doi.org/10.1186/s12915-021-01035-y.
783 [39] S.R. Eddy, A new generation of homology search tools based on probabilistic inference, Genome Inf. 23
784 (2009) 205–211.
785 [40] P. Jones, D. Binns, H.Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G.
786 Nuka, S. Pesseat, A.F. Quinn, A. Sangrador-Vegas, M. Scheremetjew, S.Y. Yong, R. Lopez, S. Hunter,
787 InterProScan 5: genome-scale protein function classification, Bioinformatics. 30 (2014) 1236–1240.
788 https://doi.org/10.1093/bioinformatics/btu031.
789 [41] M. Kearse, R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, S. Buxton, A. Cooper, S.
790 Markowitz, C. Duran, T. Thierer, B. Ashton, P. Meintjes, A. Drummond, Geneious Basic: an integrated and
791 extendable desktop software platform for the organization and analysis of sequence data, Bioinformatics. 28
792 (2012) 1647–1649. https://doi.org/10.1093/bioinformatics/bts199.
793 [42] M.G. Claros, P. Vincens, Computational method to predict mitochondrially imported proteins and their
794 targeting sequences, Eur. J. Biochem. 241 (1996) 779–786. https://doi.org/10.1111/j.1432-
795 1033.1996.00779.x.
796 [43] Y. Fukasawa, J. Tsuji, S.C. Fu, K. Tomii, P. Horton, K. Imai, MitoFates: Improved prediction of
797 mitochondrial targeting sequences and their cleavage sites, Mol. Cell. Proteomics. 14 (2015) 1113–1126.
798 https://doi.org/10.1074/mcp.M114.043083.
799 [44] J.J. Almagro Armenteros, M. Salvatore, O. Emanuelsson, O. Winther, G. von Heijne, A. Elofsson, H.
800 Nielsen, Detecting sequence signals in targeting peptides using deep learning, Life Sci. Alliance. 2 (2019)
801 e201900429. https://doi.org/10.26508/lsa.201900429.
802 [45] F. Burki, A.J. Roger, M.W. Brown, A.G.B. Simpson, The new tree of eukaryotes, Trends Ecol. Evol. 35
803 (2020) 43–55. https://doi.org/10.1016/j.tree.2019.08.008.
804 [46] K. Katoh, D.M. Standley, MAFFT multiple sequence alignment software version 7: improvements in
805 performance and usability, Mol. Biol. Evol. 30 (2013) 772–780. https://doi.org/10.1093/molbev/mst010.
806 [47] S. Capella-Gutiérrez, J.M. Silla-Martínez, T. Gabaldón, trimAl: a tool for automated alignment trimming in
807 large-scale phylogenetic analyses, Bioinformatics. 25 (2009) 1972–1973.
808 https://doi.org/10.1093/bioinformatics/btp348.

[48] L.T. Nguyen, H.A. Schmidt, A. von Haeseler, B.Q. Minh, IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies, Mol. Biol. Evol. 32 (2015) 268–274. https://doi.org/10.1093/molbev/msu300.

[49] H.-C. Wang, B.Q. Minh, E. Susko, A.J. Roger, Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation, Syst. Biol. 67 (2018) 216–235. https://doi.org/10.1093/sysbio/syx068.

[50] D.T. Hoang, O. Chernomor, A. von Haeseler, B.Q. Minh, L.S. Vinh, UFBoot2: Improving the ultrafast bootstrap approximation, Mol. Biol. Evol. 35 (2017) 518–522. https://doi.org/10.1093/molbev/msx281.

[51] B. Kaur, M. Valach, P. Peña-Diaz, S. Moreira, P.J. Keeling, G. Burger, J. Lukeš, D. Faktorová, Transformation of *Diplonema papillatum*, the type species of the highly diverse and abundant marine microeukaryotes Diplonemida (Euglenozoa), Environ. Microbiol. 20 (2018) 1030–1040. https://doi.org/10.1111/1462-2920.14041.

[52] K. Šubrtová, B. Panicucci, A. Zíková, ATPaseTb2, a unique membrane-bound FoF1-ATPase component, is essential in bloodstream and dyskinetoplastic trypanosomes, PLoS Pathog. 11 (2015) e1004660. https://doi.org/10.1371/journal.ppat.1004660.

[53] M. Valach, A. Léveillé-Kunst, M.W. Gray, G. Burger, Respiratory chain Complex I of unparalleled divergence in diplonemids, J. Biol. Chem. 293 (2018) 16043–16056. https://doi.org/10.1074/jbc.RA118.005326.

[54] J. Pyrih, V. Rašková, I. Škodová-Sveráková, T. Pánek, J. Lukeš, ZapE/Afg1 interacts with Oxa1 and its depletion causes a multifaceted phenotype, PLoS One. 15 (2020) e0234918. https://doi.org/10.1371/journal.pone.0234918.

[55] J. Cox, M. Mann, MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification, Nat. Biotechnol. 26 (2008) 1367–1372. https://doi.org/10.1038/nbt.1511.

[56] J. Cox, N. Neuhauser, A. Michalski, R.A. Scheltema, J. V. Olsen, M. Mann, Andromeda: A peptide search engine integrated into the MaxQuant environment, J. Proteome Res. 10 (2011) 1794–1805. https://doi.org/10.1021/pr101065j.

[57] M. Zoltner, G.D. Campagnaro, G. Taleva, A. Burrell, M. Cerone, K.F. Leung, F. Achcar, D. Horn, S. Vaughan, C. Gadelha, A. Zíková, M.P. Barrett, H.P. de Koning, M.C. Field, Suramin exposure alters cellular metabolism and mitochondrial energy production in African trypanosomes, J. Biol. Chem. 295 (2020) 8331–8347. https://doi.org/10.1074/jbc.RA120.012355.

[58] I. Škodová-Sveráková, K. Záhonová, V. Juricová, M. Danchenko, M. Moos, P. Baráth, G. Prokopchuk, A. Butenko, V. Lukáčová, L. Kohútová, B. Bučková, A. Horák, D. Faktorová, A. Horváth, P. Šimek, J. Lukeš, Highly flexible metabolism of the marine euglenozoan protist *Diplonema papillatum*, BMC Biol. 19 (2021) 251. https://doi.org/10.1186/s12915-021-01186-y.

[59] B. Schwanhüusser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, M. Selbach, Global quantification of mammalian gene expression control, Nature. 473 (2011) 337–342.

https://doi.org/10.1038/nature10098.

[60] P. Čermáková, A. Maďarová, P. Baráth, J. Bellová, V. Yurchenko, A. Horváth, Differences in mitochondrial NADH dehydrogenase activities in trypanosomatids, Parasitology. 148 (2021) 1161–1170. https://doi.org/10.1017/S0031182020002425.

[61] K.P. Block, R. Paul Aftring, M.G. Buse, A.E. Harper, Estimation of branched-chain α-keto acid dehydrogenase activation in mammalian tissues, Methods Enzymol. 166 (1988) 201–213. https://doi.org/10.1016/S0076-6879(88)66026-5.

[62] J.G. McCormack, E.S. Bromidge, N.J. Dawes, Characterization of the effects of Ca2+ on the intramitochondrial Ca2+-sensitive dehydrogenases within intact rat-kidney mitochondria, BBA - Bioenerg. 934 (1988) 282–292. https://doi.org/10.1016/0005-2728(88)90088-6.

[63] D.J. Danner, E.D. Davidson, L.J. Elsas., Thiamine increases the specific activity of human liver branched chain α-ketoacid dehydrogenase, Nature. 254 (1975) 529–530. https://doi.org/10.1038/254529a0.

[64] J.E. Knapp, D.T. Mitchell, M.A. Yazdi, S.R. Ernst, L.J. Reed, M.L. Hackert, Crystal structure of the truncated cubic core component of the *Escherichia coli* 2-oxoglutarate dehydrogenase multienzyme complex, J. Mol. Biol. 280 (1998) 655–668. https://doi.org/10.1006/jmbi.1998.1924.

[65] H. Kinugawa, N. Kondo, A. Komine-Abe, T. Tomita, M. Nishiyama, S. Kosono, In vitro reconstitution and characterization of pyruvate dehydrogenase and 2-oxoglutarate dehydrogenase hybrid complex from *Corynebacterium glutamicum*, Microbiologyopen. 9 (2020) e1113. https://doi.org/10.1002/mbo3.1113.

[66] C. Schnarrenberger, W. Martin, Evolution of the enzymes of the citric acid cycle and the glyoxylate cycle of higher plants, Eur. J. Biochem. 269 (2002) 868–883. https://doi.org/10.1046/j.0014-2956.2001.02722.x.

[67] N.L. Marrott, J.J.T. Marshall, D.I. Svergun, S.J. Crennell, D.W. Hough, M.J. Danson, J.M.H. Van Den Elsen, The catalytic core of an archaeal 2-oxoacid dehydrogenase multienzyme complex is a 42-mer protein assembly, FEBS J. 279 (2012) 713–723. https://doi.org/10.1111/j.1742-4658.2011.08461.x.

[68] A. De Kok, A.F. Hengeveld, A. Martin, A.H. Westphal, The pyruvate dehydrogenase multi-enzyme complex from Gram-negative bacteria, Biochim. Biophys. Acta. 1385 (1998) 353–366. https://doi.org/10.1016/S0167-4838(98)00079-X.

[69] M. Smolle, J.G. Lindsay, Molecular architecture of the pyruvate dehydrogenase complex: Bridging the gap, Biochem. Soc. Trans. 34 (2006) 815–818. https://doi.org/10.1042/BST0340815.

[70] J. Lee, S. Oh, S. Bhattacharya, Y. Zhang, L. Florens, M.P. Washburn, J.L. Workman, The plasticity of the pyruvate dehydrogenase complex confers a labile structure that is associated with its catalytic activity, PLoS One. 15 (2021) e0243489. https://doi.org/10.1371/journal.pone.0243489.

[71] V. Jagannathan, R.S. Schweet, Pyruvic oxidase of pigeon breast muscle. I. Purification and properties of the enzyme, J. Biol. Chem. 196 (1952) 551–562.

[72] H. Oku, T. Kaneda, Biosynthesis of branched-chain fatty acids in *Bacillis subtilis*. A decarboxylase is essential for branched-chain fatty acid synthetase, J. Biol. Chem. 263 (1988) 18386–18396. https://doi.org/10.1016/s0021-9258(19)81371-6.

[73] S. Prajapati, D. Haselbach, S. Wittig, M.S. Patel, A. Chari, C. Schmidt, H. Stark, K. Tittmann, Structural

and functional analyses of the human PDH complex suggest a "division-of-labor" mechanism by local E1 and E3 clusters, Structure. 27 (2019) 1124–1136. https://doi.org/10.1016/j.str.2019.04.009.

[74] Y. Perez-Riverol, A. Csordas, J. Bai, M. Bernal-Llinares, S. Hewapathirana, D.J. Kundu, A. Inuganti, J. Griss, G. Mayer, M. Eisenacher, E. Pérez, J. Uszkoreit, J. Pfeuffer, T. Sachsenberg, Ş. Yilmaz, S. Tiwary, J. Cox, E. Audain, M. Walzer, A.F. Jarnuczak, T. Ternent, A. Brazma, J.A. Vizcaíno, The PRIDE database and related tools and resources in 2019: improving support for quantification data, Nucleic Acids Res. 47 (2019) D442–D450. https://doi.org/10.1093/nar/gky1106.

Supporting Information

Fig. S1. Phylogenetic analysis of A) AceE, B) E1α, C) E1β, D) E2, and E) E3 subunits. Euglenozoan sequences are in bold. Ultrafast bootstrap support values are shown when ≥75%. Poly-A selection during RNA preparation and the presence of AceE in datasets of two different *Anisonema* strains make the possibility of prokaryotic contamination unlikely.

Fig. S2. Bioinformatic analysis of E2 proteins. Euglenozoan E2 sequences from different phylogenetic clades (**Fig. S1**) were separately aligned with the *E. coli* E2o (Uniprot accession: P0AFG6 (ODO2_ECOLI)). The active site regions identified in the catalytic domain of the *E. coli* protein [64] are boxed, and their percentage of pairwise identity is shown above. The residues conferring substrate specificity in *E. coli* [64] are in orange with arrows above them.

Fig. S3. Immunoblot analysis of *P. papillatum* wild type (WT), and cell lines with Protein A-tagged AceE or E3. The blot was probed by anti-Protein A to detect the tagged AceE and E3 proteins. The expected sizes of AceE and E3 are approximately 140 and 75 kDa, respectively.

Table S1. Oligonucleotides used in this study.

Table S2. Identified sequences for subunits of dehydrogenase complexes in diplonemids. Proteomic data of *P. papillatum* were analyzed in a previous study (PRIDE database accession number PXD025411; [58]). Data for identified peptides are represented as $\log_2$-transformed ratios. ANOVA statistical analysis was performed with the Benjamini-Hochberg correction for multiple testing with a p-value threshold at 0.01. For pairwise comparisons, the post hoc Tukey's

914 test was used at $p \leq 0.01$. Peptides that do not express any change in label-free intensities (LFQ)
915 values are marked as "0", those that were not identified in the proteomic data are indicated as
916 "x". Proteins enriched at least 1-fold of the $\log_2$-transformed ratio were filtered by effect size.
917 Negative values indicate a decrease in protein abundance when cells were grown in the medium
918 without tryptone compared to the medium with tryptone. Abbreviations: p, PDH; o, OGDH; b,
919 BCKDH; TPM, transcripts per million.
920
921 **Table S3. Protein domains of E2 subunits.** Protein domains were identified by InterProScan.
922 Their localizations in the sequences are listed. Abbreviations: p, PDH; o, OGDH; b, BCKDH.
923
924 **Table S4. Identified E3-binding proteins in euglenozoans.**
925
926 **Table S5. Proteomics analyses of subcellular fractions from *P. papillatum*. A**) List of
927 identified proteins and associated quantitative values (MaxQuant LFQ). **B**) Principal component
928 analysis of the four biological replicates showing the differences between the protein content of
929 each fraction and replicate concordance. **C**) List of proteins the levels of which changed
930 significantly in one of the fractions (cell, cytosol, mitochondrion). Hierarchical clustering that
931 allowed assignment of analyzed peptides into the respective cellular fractions, was performed
932 based on Z-score-normalized averages of LFQ intensities.
933
934 **Table S6. Immunopurification of AceE (A,C) and E3 (B,D) by IGEPAL (A-B) and DDM**
935 **(C-D).** iBAQ values for proteins pulled down by AceE- and E3-Protein A in the final eluate
936 represent values from one replicate (A), or average values from three biological (B) or technical
937 (C-D) replicates. The ratios AceE iBAQ/ WT iBAQ and E3 iBAQ/ WT iBAQ indicate the
938 degree of protein enrichment in each fraction. Abbreviations: p, PDH; o, OGDH; b, BCKDH.

Chapter 3

Relocation of paraflagellar rod proteins in *Paradiplonema papillatum*

Relocation of paraflagellar rod proteins in *Paradiplonema papillatum*

Pragya Tripathi[1,2], Michael Hammond[1,2] and Julius Lukeš[1,2]

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

[2] Department of Molecular Biology, Faculty of Science, University of South Bohemia, České Budějovice (Budweis), Czech Republic

Abstract

The paraflagellar rod (PFR) represents the premier synapomorphy among Euglenozoa. This lattice-like structure, running parallel to the flagella axoneme, has been observed in every well-studied member of this phylum, either in at least one life stage or in a vestigial manner, excepting the diplonemid *Paradiplonema papillatum*. Preliminary genome analysis of this flagellate revealed high-scoring homologues to core PFR components present in sister-clades of euglenids and kinetoplastids, prompting us to investigate the localization of these proteins. Tagging of five PFR proteins in *P. papillatum* revealed their absence from the flagella, and their relocation into the cytoplasm. In case of PFR1, we have documented its association with microtubules, suggesting its presence in the apical papilla. The papilla is a projecting cell structure positioned between the cytopharynx and flagellar pocket, of presently unknown function. This represents the first localization of any protein to this characteristic diplonemid structure and suggests retargeting events which have taken place from the flagella to this region. We propose that such relocalizations may have contributed to the prominence of the papilla in this diplonemid, from which its species name is derived. This study demonstrates is the first example of a truly PFR-deficient member of the

Euglenozoa, in which several PFR proteins have been relocated from the flagellum.

Introduction

Flagella are whip-like structures employed by certain unicellular organisms to move through fluids, while in multicellular organisms, they can serve to beat in a unified motion to move fluid and objects around the cell (Berg, 2003; Schavemaker and Lynch, 2022). In eukaryotes, flagella exhibit a conserved fundamental architecture (Satir and Christensen, 2007), consisting of the central axoneme, composed of microtubule doublets arranged cylindrically, surrounded by the flagellar membrane. The microtubule doublets consist of 13 protofilaments, which are in turn arranged around a central pair of microtubules, with their sliding motion, induced by attached motor proteins, being responsible for the bending movement of the flagella (Bastin, Matthews and Gull, 1996).

Eukaryotic flagella can additionally serve as non-motile structures primarily employed for cellular signalling or environmental sensing and are distinguished structurally from motile flagella by their lack of a central pair microtubule doublets and functionally by their inability to move (Maric, Epting and Engman, 2010; Wan, 2018; Pinskey *et al.*, 2022). In mammals, motile flagella transport sperm and eggs, whereas non-motile sensory cells in the ear and nose have roles in olfactory and audio sensing and signal transduction (Satir and Christensen, 2007). The functionalities of flagella are not mutually exclusive, with the unicellular alga *Euglena gracilis* employing its flagella for both environmental sensing and movement responses upon detecting changes in light and gravity within its surrounding environment, showcasing the versatile nature of these structures (Butenko *et al.*, 2021).

In the protist phylum of Euglenozoa, the most universally conserved feature among this group is the paraflagellar rod (PFR) within the flagella, first studied in the kinetoplastid subclade (Kohl, Sherwin and Gull, 1999; Rotureau *et al.*, 2014). The PFR is a unique lattice-like structure that runs parallel to the axoneme and is additionally attached to the flagellar membrane (Bastin, Matthews and Gull, 1996; Coceres *et al.*, 2021). Its precise function/s is not well understood but it is believed to play a role in stabilizing the flagellum and directing movement, as well as having roles in energy metabolism and signal transduction (Lacomble *et al.*, 2009). Furthermore, the PFR has been suggested to play a role in cell division and differentiation, from studies showing the PFR is involved in the positioning of the basal body during cell division (Rosati *et al.*, 1991). Although dinoflagellates were once thought to possess a similar structure, subsequent investigations have revealed that these structures differ significantly from the PFR observed in euglenozoans (Cachon *et al.*, 1988; Maharana *et al.*, 2014) PFR1 and PFR2 represent the two most abundant and highly conserved proteins which were initially discovered in *Crithidia fasciculata* and later observed in various other euglenozoans (Kohl, Sherwin and Gull, 1999; Hammond *et al.*, 2020).

Diplonemids are a group of unicellular eukaryotes that represent a neglected subclade of Euglenozoa. The feeding habits and ecology of diplonemids are not well understood, with most species being bacterivorous (Prokopchuk *et al.*, 2022), although a minority may also have acquired parasitic life style (Roy *et al.*, 2007). Although a few diplonemids are found in freshwater lakes (Mukherjee *et al.*, 2020), the majority of their abundance and diversity are observed in deep-sea marine environments (Flegontova *et al.*, 2020).

The most studied representative is *Paradiplonema papillatum* (formerly member of the genus *Diplonema*) which displays short flagella, and notably lack an ultrastructurally discernible PFR, contrasting with most other diplonemids and

Euglenozoa as a whole (Adl et al., 2019). The PFR was observed in several examined diplonemids, either permanently or in a life stage-dependent manner (Tashyreva *et al.*, 2022), similar to what has been observed for kinetoplastids (Portman and Gull, 2010).

As an apparent exception to this euglenozoan synapomorphy, and as a species that is genetically tractable (Kaur *et al.*, 2018b), *P. papillatum* warrants further investigation to determine if the PFR is truly absent from the flagellum. Genome analysis has revealed that *P. papillatum* retains homologs to conserved PFR proteins in Euglenozoa (Valach, Moreira, Petitjean, Benz, Butenko, *et al.*, 2023), prompting us to investigate where these proteins are localized in the cell. We have thus endogenously tagged five PFR proteins in *P. papillatum*. This tagging showed that these PFR proteins are indeed absent from the flagella, predominantly localized within the cytoplasm, with PFR1 also appearing as a putative component of the apical papilla. These unique protein localizations suggest both retargeting events and functional reassignments. We ultimately demonstrate that *P. papillatum* is truly PFR-deficient, lacking core conserved components of this structure even in a vestigial form, and represents the first confirmed exception among the phylum Euglenozoa.

Methodology

Tagging of PFR candidates

The 3' coding sequence of PFR candidates and the 3' untranslated region were amplified using primers that overlapped with the A-Neomycin cassette of vector pDP002. The vector was synthesized by Eurofins Genomics (Ebersberg, Germany). PFR1 (DIPPA_34257), PFR2 (DIPPA_04871), and UPF1 (DIPPA_15741) candidates were tagged with protein-A, while PFR6 (DIPAA_22051) and PAR4 (DIPPA_17837) were tagged with a cassette from a

modified vector pDP002 at the C-terminal end (Faktorová *et al.*, 2020). The modified plasmid was created by replacing the coding sequence of Protein-A in pDP002 through cloning using the restriction sites NheI and NdeI. To assemble all three fragments, a nested PCR was performed using Phusion polymerase (NEB®) with the amplified fragments as templates (**Fig. 1A**, **Table 1**). The PCR product was then cloned into TM 2.1-TOPO® vector (ThermoFisher). The clones were verified by sanger sequencing. For transfection, 10 µg of the final plasmid was digested with EcoRI (NEB®), followed by ethanol precipitation. The precipitated DNA was resuspended in 10 µl of water.

Electroporation of *P. papillatum*

An Amaxa Nucleofector® II was used to transform a total of 5 x 10^7 cells, as previously described (Kaur *et al.*, 2018). Clones were selected in 24 well plates at 27ºC using various concentrations of antibiotic G418 (45 to 83 g/ml). Successful transfectants could be observed after 2 weeks. Before being verified by western blot, each clone was transferred to a volume of 20 ml culture medium and incubated for additional 2 to 3 weeks.

The DNA isolation kit (Qiagen®) was employed for the isolation of genomic DNA. Primer pairs containing a complementary region of the forward open reading frame for PFR1, PFR2 and UPF1 genes, with the reverse primer corresponding to the Protein-A region utilized for integration verification (**Fig. 1B-D**). Primer pair containing PCR amplification was carried out utilizing OneTaq polymerase (NEB®). Cassette integration in the genome of *P. papillatum* was confirmed through visualising of PCR product of expected size.

Verification of expression of tagged proteins

Protein samples were prepared through pelleting of 5 x 10^5 cells and resuspending them in 25 µl of 2 x SDS sample buffer, separated on 4-20% Mini-protein TGX

stain-free gels (Bio-Rad®), and subsequently transferred to a PVDF membrane at 100 V for 1 hour. After blocking with 5% milk in Phosphate Buffered Saline (PBS) with 0.5% Tween (PBS-T) for at least 30 minutes at room temperature, the membrane was incubated with an anti-protein A (Sigma®; used at 1:10,000) and anti-V5 antibodies (Sigma®; used at 1:1,000) in 5% milk in PBS-T overnight at 4ºC. After 3 washes in PBS-T, the membrane was incubated with horse radish peroxidase conjugated 'anti-rabbit' and 'anti-mouse' antibodies (Sigma®; used at 1:1,000) at room temperature for 1 hour. The membrane was then washed three times in PBS-T and the signal developed using Clarity Western ECL Substrate (Bio-Rad®). An 'anti-tubulin' antibody (Sigma®; used at 1:10,000) served as a loading control.

Immunofluorescence assay

5 to 10 ml of log phase culture (containing approximately 2 x 10^6 cells) was centrifuged at 1,000 x g for 5 minutes. Cells were fixed in 4% paraformaldehyde in artificial seawater (Sigma) at room temperature for 30 minutes. The fixative was removed with 1X PBS mixed with artificial seawater (1:1), and the pellet was resuspended in PBS before being spotted on a gelatine-coated slide. Cells were permeabilized for 20 minutes for antibody labelling in 100% ice-cold methanol. Throughout the procedure, slides were stored in a humidified chamber. Slides were washed in PBS after 20 minutes and blocked in 5% foetal bovine serum for 45 minutes with 0.05% Tween. After removing the blocking solution, the cells were washed in PBS for 5 minutes. The primary antibody used anti-protein-A (Sigma®, 1:2,000, produced in rabbit) and V5 Sigma® (1:100, produced in mouse). Additionally, an anti-tubulin (beta, clone, KMX-1) monoclonal antibody (produced in mouse, used at 1:50; Sigma-Aldrich, T9026) was employed. The slides were covered with parafilm and incubated overnight at 4°C after application of the primary antibody. Following removal of the primary antibody, the slides

were washed three times with PBS. Subsequently, the secondary antibody was applied, which consisted of AlexaFluor555-goat-anti-rabbit (Invitrogen®, 1:1,000) or AlexaFluor488-anti-mouse (Invitrogen®, 1:1,000). The slides were then incubated in the dark at room temperature, covered with parafilm, for a duration of one hour was applied and the slide incubated for 1 hour in the dark at room temperature, covered with parafilm. To visualize DNA, slides were washed in PBS and rinsed with 4-6-diamidino-2-phenylindole (DAPI) containing the antifade reagent ProlongGold (Life Technologies®) after the secondary antibody. Images were obtained using camera Olympus DP73 (Axioplan 2 imaging).

Results

Establishment of stable transfected cell lines

Sequences corresponding to PFR1, PFR2, PFR6 and PAR4 (Tb927.8.4290, Tb927.8.4970, Tb927.7.6970, Tb927.11.13500, respectively) in *T. brucei* strain 927 from TriTrypDB (http://tritrypdb.org)) were identified by Blast queries against the genome of *P. papillatum* (Valach, Moreira, Petitjean, Benz, Butenko, *et al.*, 2023). Due to their high hit scores and conservation amongst available diplonemid transcriptomes, unambiguous homologs were identified. The lower scoring homolog of protein UPF1 (Tb927.10.10140) was additionally included due to its validated protein presence in a previous metabolic analysis of *P. papillatum* (Škodová-Sveráková *et al.*, 2021).

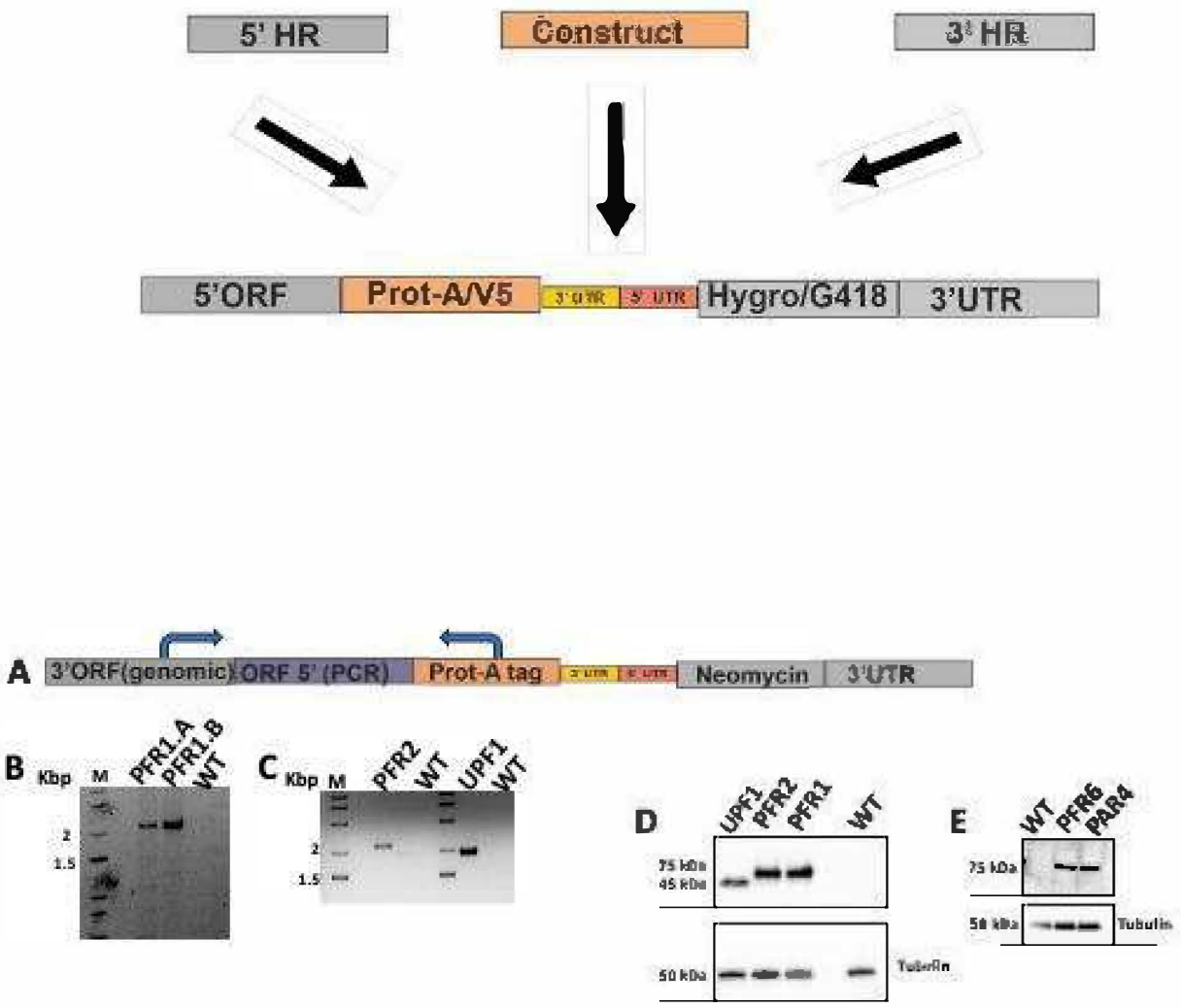


Figure 1. A scheme showing the cloning strategy. **A.** A schematic representation of the primers employed for the verification of integration of PFR1, PFR2 and UPF1 into genomic DNA. **B.** PCR of total DNA of *P. papillatum* wild type (WT) and transformed cell lines (A and B represent two different clones of PFR1). Expected size of the PCR amplicon was 2.5 Kbp. **C.** PCR of total DNA of *P. papillatum* WT and transformed cell lines. The expected size of the PCR amplicons was 2.5 Kbp and 2.0 Kbp for PFR2 and

UPF1, respectively. **D and E.** Western blot analyses of *P. papillatum* WT and transformed cell lines. Tubulin was used as a loading control.

Both the V5 and protein-A tags were selected to test their suitability as markers for traditional flagellar proteins. Tag integration was verified via PCR for PFR1, PFR2 and UPF1 (**Figs. 1B-C**). Next, the generation of tagged proteins was demonstrated through western blots of the total cell lysate (**Figs. 1 D-F**). Combined, this data demonstrates not only a successful integration of the tagged genes in the generated cell lines, but also the production of corresponding proteins.

Cytoplasmic localization of paraflagellar candidates

To visualize the flagella, cytopharynx and other microtubule-associated structures, anti-tubulin antibody was employed for immunofluorescence assays, while DAPI stain was used to illuminate the nuclear and mitochondrial DNA (**Figs. 2 and 3**). Protein signal from the tag epitopes was visible in all five generated cell lines. All cell lines show the protein signal within the cytoplasm, distributed as a series on heterogenous granules **(Figs. 2 and 3)**. Importantly, the signal is not localized in the flagella, which are clearly labeled with the anti-tubulin antibody **(Figs. 2 and 3)**. DAPI staining revealed a prominent nucleus and weak signal in the reticulated mitochondrion, with organellar DNA evenly distributed throughout the lumen **(Figs. 2 and 3)**. In addition to the signal in the cytoplasm, PFR1 is clearly located also in the apical part, underlying the papilla **(Fig. 2A)**.

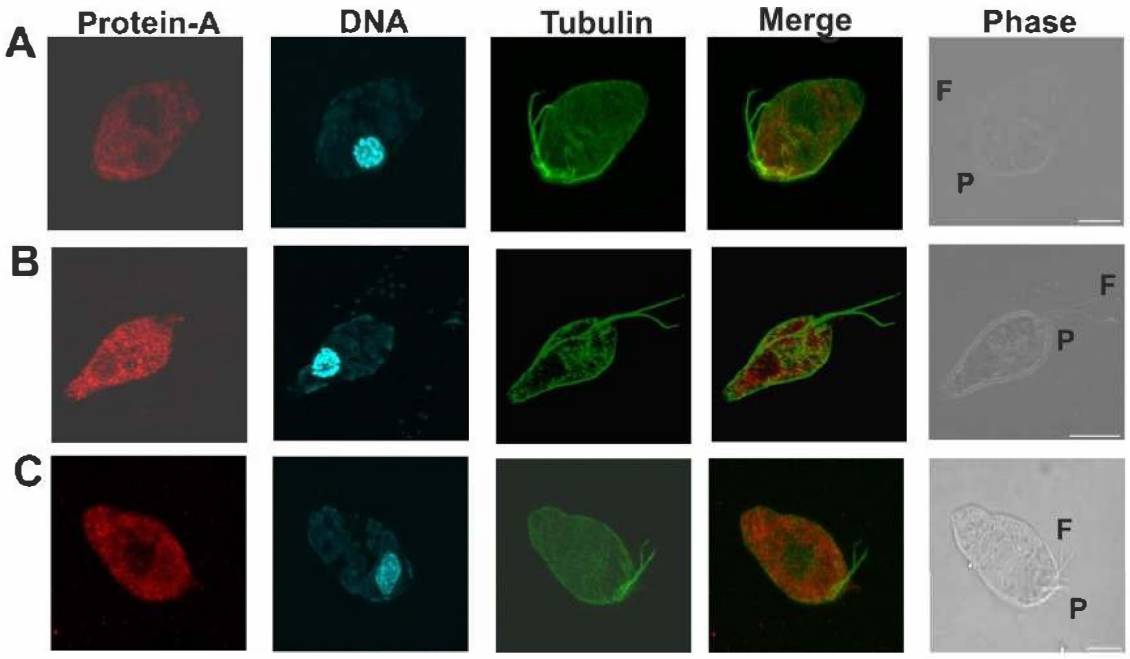


Figure 2. Localization of the tagged proteins by immunofluorescence via the red channel (AlexaFluor 555). Tubulin, which was visualized via the green channel (AlexaFluor 488) is present in the microtubular corset, flagella, the cytopharynx and the apical papilla. DNA was tracked by DAPI via the blue channel. Merge shows the overlay of channels. **A.** PFR1 (DIPPA_34257), **B.** PFR2 (DIPPA_04871), **C.** UPF1(DIPPA_15741). Abbreviations: (F) flagella, (P) emergent papilla. Scale bar represents 10 μm.

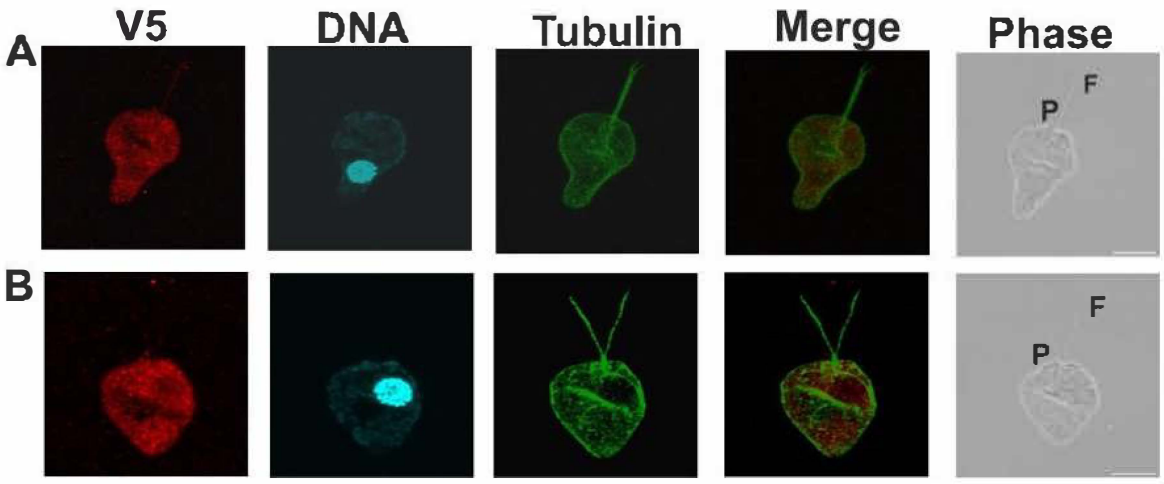


Figure 3. Localization of the tagged proteins, as well as the tubulin and mitochondrial and nuclear DNA as described in Figure 2. **A.** PFR6 (DIPAA_22051), **B.** PAR4 (DIPPA_17837). Abbreviations: (F) flagella, (P) emergent papilla. Scale bar represents 10 μm.

Discussion

Recently surveyed selected diplonemid species manifested the PFR under the starvation-induced 'swimming' and 'sessile' stages (Tashyreva *et al.*, 2022). Thus, a PFR has been observed in all ultrastructurally studied species of Euglenozoa, except the model diplonemid, *P. papillatum*, which does not exhibit distinct life stages, rather, only slightly different morphotypes in response to changing conditions (Tashyreva *et al.*, 2022).

Hence, we decided to investigate this so far unchallenged universal synapomorphy of the phylum Euglenozoa by investigating the structure in *P. papillatum*. While its electron microscopy examination identified no visible PFR

(Tashyreva *et al.*, 2022), it remained a possibility that a cryptic or vestigial version was present within its flagella, in a manner similar to the trypanosmatid *Crithidia deanei* (Gadelha *et al.*, 2005). In this study, we show for the first time the distribution of conserved PFR proteins extraneous to the flagella. Indeed, all five selected homologs of the core components of the Euglenozoan flagellum localize to the cytoplasm of the studied diplonemid, with PFR1 also being concentrated in the microtubule-associated structures near the apical papillum of the cell, qualifying *P. papillatum* as the first surveyed euglenozoan to be truly PFR-deficient. This finding allows at least two alternative scenarios. In the first, the PFR proteins were secondarily retargeted into the cytoplasm, where their function beyond their previously suggested flagellar roles in motility, metabolism signal transduction and cell differentiation (Krüger and Engstler, 2015; Häder and Hemmersbach, 2017; Klena and Pigino, 2022). Alternatively, the cytoplasmic localization is ancestral and the PFR proteins were subject to retargeting and neofunctionalization in virtually all euglenozoans. However, the non-basal position of *P. papillatum* (Flegontova *et al.*, 2016; Tashyreva *et al.*, 2022) makes this scenario rather unlikely.

The mechanisms by which proteins are targeted to the flagella are complex, as they do not employ N-terminal target peptides, but instead rely on small, specific associations with the intraflagellar transport apparatus (Absalon *et al.*, 2008; Mul, Mitra and Peterman, 2022). The discovery of traditional flagellar proteins that have been retargeted holds the promise to provide information on how proteins are targeted and retargeted to the flagella and other regions of the cell (Bastin *et al.*, 1999). In particular, the localization of PFR1 in association with the microtubules in the apical region of the cell suggests structural association underlying the papilla, which until now has remained entirely unknown in terms of its constitution.

Diplonemids uniquely possess the cytoskeletal structure known as the papilla, which is characterized by a protrusion of the cell membrane supported by underlying microtubules and microfilaments (Tashyreva *et al.*, 2022). The complex structure of the papilla includes a basal body, a central filament, and several microtubules extending into it (Elbrächter, Schnepf and Balzer, 1996). Although the precise function of the papilla remains elusive, it has been postulated to play a crucial role in cell motility and feeding (Yubuki, Simpson and Leander, 2013). The papilla has been observed in variable positions, alternately covering the cytopharynx or projecting directly from the cell, though it remains unconfirmed whether this is a directly motile structure (Yubuki, Simpson and Leander, 2013). The papilla of *P. papillatum* is the most prominent one observed in diplonemids, and it remains a possibility that this may be a result of the accumulation of the PFR1 protein.

The presence of other PFR proteins throughout the cytoplasm represents an intriguing and unexpected finding. Further research is needed to understand the roles of these PFR proteins in the cytoplasm and how their functions are integrated with those of other cytoplasmic proteins. Specifically, the use of protein pull-down assays can identify interacting proteins to suggest functional association, and in the case of PFR1, identify other candidate components of the papilla. Knockdown studies additionally have the potential to demonstrate if these PFR proteins perform functions critical to cell survival, as has been amply documented for kinetoplastids (Gadelha *et al.*, 2006; Portman and Gull, 2010; Sunter and Gull, 2016).

On the putative function of the PFR within this subclade, it is worth noting that euglenids exhibit a 'spinning lasso' pattern of their dorsal flagella to enable rapid cell locomotion (Yubuki, Čepička and Leander, 2016b), and that a similar flagellar pattern was recently described also in diplonemids which possess a PFR

(Tashyreva *et al.*, 2022). However, by contrast, the PFR-lacking *P. papillatum* exhibits a more constrained flagella flexing as a means of slower propulsion (our unpubl. data). Amongst the family Diplonemidae, the most closely related genera to the genus *Diplonema*, which accommodates *P. papillatum*, namely the genera *Diplonema*, *Metadiplonema,* and *Rhynchopus* exhibit a 'spinning lasso' of the anterior flagellum solely in life stages that possess a pronounced PFR (Tashyreva *et al.* 2018a, b). This observation prompts us to speculate that the loss of selective pressure for fast movement, facilitated by this specific flagellar behavior triggered the loss of PFR in the flagellum of *P. papillatum.*

The more distantly related diplonemids belonging to the genera *Lacrimia*, *Flectonema,* and *Sulcionema* exhibit limited movement strategies of 'swimming', 'floundering' or 'gliding', which distinctly lack the 'spinning lasso' pattern despite the permanent presence of the PFR (Tashyreva *et al.*, 2022). We suggest that these groups represent intermediate states of the above-proposed process of flagella reduction. More specifically, they exhibit alternate movement strategies while retaining the PFR with a characteristic lattice structure. The ecological valence of diplonemids is virtually limitless, as they can be found in every liter of seawater (de Vargas *et al.*, 2015). This is inevitably associated with different movement strategies, which may or may not rely on the flagella . Therefore, it is plausible to assume that diplonemids "experiment" with their flagella much more extensively than the mostly parasitic kinetoplastids, and that evolutionary pressures they encounter lead to strikingly different outcomes, ranging from short rigid flagella to long and flexible ones. The association of *P. papillatum* with drifting eel grass (*Zostera marina*) (Porter 1973) may predispose it to limited mobility.

In this study, we investigated the presence of the PFR in the model diplonemid *P. papillatum*. We determined that *P. papillatum* is truly PFR-deficient, representing

a unique anomaly of the widely studied euglenozoan protists. Unexpectedly, the PFR proteins were present in the cytoplasm, and in one case were associated with microtubular structures near the specialized projection of the cell called the apical papilla. This is indicative of extensive retargeting events, with plausible neofunctionalization of these proteins, so far firmly associated with the flagellum. We can only speculate that the loss of PFR in *P. papillatum* has occurred due to a lack of selective pressure for efficient flagellar movement. As the first truly PFR-deficient euglenozoan flagellate, further investigation of these PFR proteins within *P. papillatum* offers a unique opportunity to understand both the conserved and divergent functional roles of PFR proteins more conclusively.

References

Absalon, S., Blisnick, T., Kohl, L., Toutirais, G., Doré, G., Julkowska, D., Tavenet, A. and Bastin, P., 2008. Intraflagellar transport and functional analysis of genes required for flagellum formation in trypanosomes. *Molecular Biology of the Cell*, *19*(3), 929-944.

Adl, S.M., Bass, D., Lane, C.E., Lukeš, J., Schoch, C.L., Smirnov, A., Agatha, S., Berney, C., Brown, M.W., Burki, F. and Cárdenas, P., 2019. Revisions to the classification, nomenclature, and diversity of eukaryotes. *Journal of Eukaryotic Microbiology*, *66*(1), 4-119.

Barsanti, L., Vismara, R., Passarelli, V. and Gualtieri, P., 2001. Paramylon (β-1, 3-glucan) content in wild type and WZSL mutant of Euglena gracilis. Effects of growth conditions. *Journal of Applied Phycology*, *13*, 59-65.

Bastin, P., MacRae, T.H., Francis, S.B., Matthews, K.R. and Gull, K., 1999. Flagellar morphogenesis: protein targeting and assembly in the paraflagellar rod of trypanosomes. *Molecular and Cellular Biology*, *19*(12), 8191-8200.

Bastin, P., Matthews, K.R. and Gull, K., 1996. The paraflagellar rod of kinetoplastida: solved and unsolved questions. *Parasitology Today*, *12*(8), 302-307.

Berg, H.C., 2003. The rotary motor of bacterial flagella. *Annual Review of Biochemistry*, *72*(1), 19-54.

Boeuf, D., Edwards, B.R., Eppley, J.M., Hu, S.K., Poff, K.E., Romano, A.E., Caron, D.A., Karl, D.M. and DeLong, E.F., 2019. Biological composition and microbial dynamics of sinking particulate organic matter at abyssal depths in the oligotrophic open ocean. *Proceedings of the National Academy of Sciences*, *116*(24), 11824-11832.

Butenko, A., Opperdoes, F.R., Flegontova, O., Horák, A., Hampl, V., Keeling, P., Gawryluk, R.M., Tikhonenkov, D., Flegontov, P. and Lukeš, J., 2020. Evolution of metabolic capabilities and molecular features of diplonemids, kinetoplastids, and euglenids. *BMC Biology*, *18*(1), 1-28.

Butenko, A., Hammond, M., Field, M.C., Ginger, M.L., Yurchenko, V. and Lukeš, J., 2021. Reductionist pathways for parasitism in euglenozoans? Expanded datasets provide new insights. *Trends in Parasitology*, *37*(2), 100-116.

Cachon, M., Cachon, J., Cosson, J., Greuet, C. and Huitorel, P., 1991. Dinoflagellate flagella adopt various conformations in response to different needs. *Biology of the Cell*, *71*(1-2), 175-182.

Cavalcanti, D.P. and de Souza, W., 2018. The kinetoplast of trypanosomatids: from early studies of electron microscopy to recent advances in atomic force microscopy. *Scanning*, *2018*.

Cavalier-Smith, T., 2010. Kingdoms Protozoa and Chromista and the eozoan root of the eukaryotic tree. *Biology Letters*, *6*(3), 342-345.

Cavalier-Smith, T., 2016. Higher classification and phylogeny of Euglenozoa. *European Journal of Protistology*, *56*, 250-276.

Cavalier-Smith, T., 2017. Euglenoid pellicle morphogenesis and evolution in light of comparative ultrastructure and trypanosomatid biology: Semi-conservative microtubule/strip duplication, strip shaping and transformation. *European Journal of Protistology*, *61*, 137-179.

Cazzulo, J.J., 1992. Aerobic fermentation of glucose by trypanosomatids. *The FASEB Journal*, *6*(13), 3153-3161.

Coceres, V.M., Iriarte, L.S., Miranda-Magalhães, A., Santos de Andrade, T.A., de Miguel, N. and Pereira-Neves, A., 2021. Ultrastructural and functional analysis of a novel extra-axonemal structure in parasitic trichomonads. *Frontiers in Cellular and Infection Microbiology*, 1101.

Danson, M.J., 1988. Dihydrolipoamide dehydrogenase: a 'new'function for an old enzyme?. *Biochemical Society Transactions*, *16*(2), 87-89.

De Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., Lara, E., Berney, C., Le Bescot, N., Probert, I. and Carmichael, M., 2015. Eukaryotic plankton diversity in the sunlit ocean. *Science*, *348*(6237), 1261605.

Elbrächter, M., Schnepf, E. and Balzer, I., 1996. Hemistasia phaeocysticola (Scherffel) comb. nov., redescription of a free-living, marine, phagotrophic kinetoplastid flagellate. *Archiv für Protistenkunde*, *147*(2), 125-136.

Drahomíra, F., Eva, D., Peña-Diaz, P. and Julius, L., 2016. From simple to supercomplex: mitochondrial genomes of euglenozoan protists. *F1000Research*, *5*.

Faktorová, D., Kaur, B., Valach, M., Graf, L., Benz, C., Burger, G. and Lukeš, J., 2020. Targeted integration by homologous recombination enables in situ tagging and replacement of genes in the marine microeukaryote Diplonema papillatum. *Environmental Microbiology*, *22*(9), 3660-3670.

Faktorová, D., Nisbet, R.E.R., Fernández Robledo, J.A., Casacuberta, E., Sudek, L., Allen, A.E., Ares Jr, M., Aresté, C., Balestreri, C., Barbrook, A.C. and Beardslee, P., 2020. Genetic tool development in marine protists: emerging model organisms for experimental cell biology. *Nature Methods*, *17*(5), 481-494.

Flegontova, O., 2017. Diversity and biogeography of diplonemid and kinetoplastid protists in global marine plankton.

Flegontova, O., Flegontov, P., Malviya, S., Audic, S., Wincker, P., de Vargas, C., Bowler, C., Lukeš, J. and Horák, A., 2016. Unexpected diversity and abundance of planktonic diplonemids in the world ocean. *Curr. Biol*, *26*, 3060-3065.

Flegontova, O., Flegontov, P., Malviya, S., Audic, S., Wincker, P., de Vargas, C., Bowler, C., Lukeš, J. and Horák, A., 2016. Extreme diversity of diplonemid eukaryotes in the ocean. *Current Biology*, *26*(22), 3060-3065.

Flegontova, O., Flegontov, P., Londoño, P.A.C., Walczowski, W., Šantić, D., Edgcomb, V.P., Lukeš, J. and Horák, A., 2020. Environmental determinants of the distribution of planktonic diplonemids and kinetoplastids in the oceans. *Environmental Microbiology*, *22*(9), 4014-4031.

Gadelha, C., Wickstead, B., de Souza, W., Gull, K. and Cunha-e-Silva, N., 2005. Cryptic paraflagellar rod in endosymbiont-containing kinetoplastid protozoa. *Eukaryotic Cell*, *4*(3), 516-525.

Gadelha, C., Wickstead, B., McKean, P.G. and Gull, K., 2006. Basal body and flagellum mutants reveal a rotational constraint of the central pair microtubules in the axonemes of trypanosomes. *Journal of Cell Science*, *119*(12), 2405-2413.

Gawryluk, R.M., Del Campo, J., Okamoto, N., Strassert, J.F., Lukeš, J., Richards, T.A., Worden, A.Z., Santoro, A.E. and Keeling, P.J., 2016. Morphological identification and single-cell genomics of marine diplonemids. *Current Biology*, *26*(22), 3053-3059.

Godeanu, S., 2020. Changes in Taxonomy from Linné to Cavalier-Smith; Case Study–Testacean Protists. *Annals Series on Biological Sciences*, *9*(1), 5-19.

Gomaa, F., Garcia, P.A., Delaney, J., Girguis, P.R., Buie, C.R. and Edgcomb, V.P., 2017. Toward establishing model organisms for marine protists: Successful transfection protocols for Parabodo caudatus (Kinetoplastida: Excavata). *Environmental Microbiology*, *19*(9), 3487-3499.

Moreno-Sánchez, R., Rodríguez-Enríquez, S., Jasso-Chávez, R., Saavedra, E. and García-García, J.D., 2017. Biochemistry and physiology of heavy metal resistance and accumulation in Euglena. *Euglena: Biochemistry, Cell and Molecular Biology*, 91-121.

Hammond, M.J., Nenarokova, A., Butenko, A., Zoltner, M., Dobáková, E.L., Field, M.C. and Lukeš, J., 2020. A uniquely complex mitochondrial proteome from Euglena gracilis. *Molecular Biology and Evolution*, *37*(8), 2173-2191.

He, J., Liu, C., Du, M., Zhou, X., Hu, Z., Lei, A. and Wang, J., 2021. Metabolic responses of a model green microalga Euglena gracilis to different environmental stresses. *Frontiers in Bioengineering and Biotechnology*, *9*, 662655.

Hopes, A., Nekrasov, V., Kamoun, S. and Mock, T., 2016. Editing of the urease gene by CRISPR-Cas in the diatom Thalassiosira pseudonana. *Plant Methods*, *12*(1), 1-12.

Ifuku, K., Yan, D., Miyahara, M., Inoue-Kashino, N., Yamamoto, Y.Y. and Kashino, Y., 2015. A stable and efficient nuclear transformation system for the diatom Chaetoceros gracilis. *Photosynthesis Research*, *123*, 203-211.

Kaur, B., Valach, M., Peña-Diaz, P., Moreira, S., Keeling, P.J., Burger, G., Lukeš, J. and Faktorová, D., 2018. Transformation of Diplonema papillatum, the type species of the highly diverse and abundant marine microeukaryotes Diplonemida (Euglenozoa). *Environmental Microbiology*, *20*(3), 1030-1040.

Kaur, B., Záhonová, K., Valach, M., Faktorová, D., Prokopchuk, G., Burger, G. and Lukeš, J., 2020. Gene fragmentation and RNA editing without borders: eccentric mitochondrial genomes of diplonemids. *Nucleic Acids Research*, *48*(5), 2694-2708.
Klena, N. and Pigino, G., 2022. Structural biology of cilia and intraflagellar transport. *Annual Review of Cell and Developmental Biology*, *38*, 103-123.

Kohil, L., Sherwin, T. and Gull, K., 1999. Assembly of the paraflagellar rod and the flagellum attachment zone complex during the Trypanosoma brucei cell cycle. *Journal of Eukaryotic Microbiology*, *46*(2), 105-109.

Kostygov, A.Y., Karnkowska, A., Votýpka, J., Tashyreva, D., Maciszewski, K., Yurchenko, V. and Lukeš, J., 2021. Euglenozoa: taxonomy, diversity and ecology, symbioses and viruses. *Open Biology*, *11*(3),.200407.

Krueger, T. and Engstler, M., 2015, October. Flagellar motility in eukaryotic human parasites. In *Seminars in cell & developmental biology* (Vol. 46,113-127). Academic Press.

Lacomble, S., Vaughan, S., Gadelha, C., Morphew, M.K., Shaw, M.K., McIntosh, J.R. and Gull, K., 2009. Three-dimensional cellular architecture of the flagellar pocket and associated cytoskeleton in trypanosomes revealed by electron microscope tomography. *Journal of Cell Science*, *122*(8), 1081-1090.

Lara, E., Moreira, D., Vereshchaka, A. and López-García, P., 2009. Pan-oceanic distribution of new highly diverse clades of deep-sea diplonemids. *Environmental Microbiology*, *11*(1), 47-55.

Lehninger, A. L. (1993) 'Lehninger principles of biochemistry, Second Edition', *W H Freeman*. (Accessed: 13 May 2023).

Lukeš, J., Flegontova, O., & Horák, A. (2015, August 17). Diplonemids. *Current Biology*. Cell Press.

Maharana, B.R., Rao, J.R., Tewari, A.K., Singh, H., Allaie, I.M. and Varghese, A., 2014. Molecular characterisation of paraflagellar rod protein gene (PFR) of Trypanosoma evansi. *Journal of Applied Animal Research*, *42*(1), 1-5.

Marande, W. and Burger, G., 2007. Mitochondrial DNA as a genomic jigsaw puzzle. *Science*, *318*(5849), 415-415.

Marande, W., Lukeš, J. and Burger, G., 2005. Unique mitochondrial genome structure in

diplonemids, the sister group of kinetoplastids. *Eukaryotic Cell*, *4*(6), 1137-1146.
Maric, D., Epting, C.L. and Engman, D.M., 2010. Composition and sensory function of the trypanosome flagellar membrane. *Current Opinion in Microbiology*, *13*(4), 466-472.

Meléndez-Hevia, E., Waddell, T.G. and Montero, F., 1994. Optimization of metabolism: the evolution of metabolic pathways toward simplicity through the game of the pentose phosphate cycle. *Journal of Theoretical Biology*, *166*(2), 201-220.

Mensa-Wilmot, K., Hoffman, B., Wiedeman, J., Sullenberger, C. and Sharma, A., 2019. Kinetoplast division factors in a trypanosome. *Trends in Parasitology*, *35*(2), 119-128.

Morales, J., Hashimoto, M., Williams, T.A., Hirawake-Mogi, H., Makiuchi, T., Tsubouchi, A., Kaga, N., Taka, H., Fujimura, T., Koike, M. and Mita, T., 2016. Differential remodelling of peroxisome function underpins the environmental and metabolic adaptability of diplonemids and kinetoplastids. *Proceedings of the Royal Society B: Biological Sciences*, *283*(1830), 20160520.

Mukherjee, I., Salcher, M.M., Andrei, A.Ş., Kavagutti, V.S., Shabarova, T., Grujčić, V., Haber, M., Layoun, P., Hodoki, Y., Nakano, S.I. and Šimek, K., 2020. A freshwater radiation of diplonemids. *Environmental Microbiology*, *22*(11), 4658-4668.
Mul, W., Mitra, A. and Peterman, E.J., 2022. Mechanisms of regulation in intraflagellar transport. *Cells*, *11*(17), 2737.

Nakazawa, M., 2017. C2 metabolism in Euglena. *Euglena: Biochemistry, Cell and Molecular Biology*, 39-45.

Nakazawa, M., Hayashi, R., Takenaka, S., Inui, H., Ishikawa, T., Ueda, M., Sakamoto, T., Nakano, Y. and Miyatake, K., 2017. Physiological functions of pyruvate: NADP+ oxidoreductase and 2-oxoglutarate decarboxylase in Euglena gracilis under aerobic and anaerobic conditions. *Bioscience, Biotechnology, and Biochemistry*, *81*(7), 1386-1393.

Novák Vanclová, A.M., Zoltner, M., Kelly, S., Soukal, P., Záhonová, K., Füssy, Z., Ebenezer, T.E., Lacová Dobáková, E., Eliáš, M., Lukeš, J. and Field, M.C., 2020. Metabolic quirks and the colourful history of the Euglena gracilis secondary plastid. *New Phytologist*, *225*(4), 1578-1592.

Obiol, A., Giner, C.R., Sánchez, P., Duarte, C.M., Acinas, S.G. and Massana, R., 2020. A metagenomic assessment of microbial eukaryotic diversity in the global ocean. *Molecular Ecology Resources*, *20*(3), 718-731.

Okamoto, N., Gawryluk, R.M., Del Campo, J., Strassert, J.F., Lukeš, J., Richards, T.A., Worden, A.Z., Santoro, A.E. and Keeling, P.J., 2019. A Revised Taxonomy of Diplonemids Including the Eupelagonemidae n. fam. and a Type Species, Eupelagonema oceanica n. gen. & sp. *Journal of Eukaryotic Microbiology*, *66*(3), 519-524.

Ouyang, Y., Chen, S., Zhao, L., Song, Y., Lei, A., He, J. and Wang, J., 2021. Global metabolomics reveals that Vibrio natriegens enhances the growth and paramylon synthesis of Euglena gracilis. *Frontiers in Bioengineering and Biotechnology*, *9*, 652021.

Pinskey, J.M., Lagisetty, A., Gui, L., Phan, N., Reetz, E., Tavakoli, A., Fu, G. and Nicastro, D., 2022. Three-dimensional flagella structures from animals' closest unicellular relatives, the Choanoflagellates. *Elife*, *11*, 78133.

Porter, D., 1973. Isonema papillatum sp. n., a new colorless marine flagellate: a light-and electronmicroscopic study. *The Journal of Protozoology*, *20*(3), 351-356.

Portman, N. and Gull, K., 2010. The paraflagellar rod of kinetoplastid parasites: from structure to components and function. *International Journal for Parasitology*, *40*(2),135-148.

Prokopchuk, G., Tashyreva, D., Yabuki, A., Horák, A., Masařová, P. and Lukeš, J., 2019. Morphological, ultrastructural, motility and evolutionary characterization of two new Hemistasiidae species. *Protist*, *170*(3), 259-282.

Prokopchuk, G., Korytář, T., Juricová, V., Majstorović, J., Horák, A., Šimek, K. and Lukeš, J., 2022. Trophic flexibility of marine diplonemids-switching from osmotrophy to bacterivory. *The ISME Journal*, *16*(5), 1409-1419.

Reed, L.J. and Oliver, R.M., 1982. Structure-function relationships in pyruvate and α-ketoglutarate dehydrogenase complexes. *Structure and Function Relationships in Biochemical Systems*, 231-241.

Rosati, G., Verni, F., Barsanti, L., Passarelli, V. and Gualtieri, P., 1991. Ultrastructure of the apical zone of Euglena gracilis: photoreceptors and motor apparatus. *Electron Microscopy Reviews*, *4*(2), 319-342.

Rotureau, B., Ooi, C.P., Huet, D., Perrot, S. and Bastin, P., 2014. Forward motility is essential for trypanosome infection in the tsetse fly. *Cellular Microbiology*, *16*(3), 425-433.

Roy, J., Faktorova, D., BENADA, O., LUKEŠ, J. and Burger, G., 2007. Description of Rhynchopus euleeides n. sp.(Diplonemea), a free-living marine euglenozoan. *Journal of Eukaryotic Microbiology*, *54*(2),137-145.

Satir, P. and Christensen, S.T., 2007. Overview of structure and function of mammalian cilia. *Annu. Rev. Physiol.*, *69*, 377-400.

Sauvadet, A.L., Gobet, A. and Guillou, L., 2010. Comparative analysis between protist communities from the deep-sea pelagic ecosystem and specific deep hydrothermal habitats. *Environmental Microbiology*, *12*(11), 2946-2964.

Sauvadet, A.L., Gobet, A. and Guillou, L., 2010. Comparative analysis between protist communities from the deep-sea pelagic ecosystem and specific deep hydrothermal habitats. *Environmental Microbiology*, *12*(11), 2946-2964.

Schavemaker, P.E. and Lynch, M., 2022. Flagellar energy costs across the tree of life. *Elife*, *11*, 77266.

Schoenle, A., Hohlfeld, M., Hermanns, K., Mahé, F., de Vargas, C., Nitsche, F. and Arndt, H., 2021. High and specific diversity of protists in the deep-sea basins dominated by diplonemids, kinetoplastids, ciliates and foraminiferans. *Communications Biology*, *4*(1), 501.

Shigeoka, S., Onishi, T., Maeda, K., Nakano, Y. and Kitaoka, S., 1986. Occurrence of thiamin pyrophosphate-dependent 2-oxoglutarate decarboxylase in mitochondria of Euglena gracilis. *FEBS Letters*, *195*(1-2), 43-47.

Sigman, D.M. and Hain, M.P., 2012. The biological productivity of the ocean. *Nature Education Knowledge*, *3*(10), p21.

Simpson, A.G., 1997. The identity and composition of the Euglenozoa. *Archiv für Protistenkunde*, *148*(3), 318-328.

Simpson, A.G., Lukeš, J. and Roger, A.J., 2002. The evolutionary history of kinetoplastids and their kinetoplasts. *Molecular Biology and Evolution*, *19*(12), 2071-2083.

Škodová-Sveráková, I., Záhonová, K., Bučková, B., Füssy, Z., Yurchenko, V. and Lukeš, J., 2020. Catalase and ascorbate peroxidase in euglenozoan protists. *Pathogens*, *9*(4), 317.

Škodová-Sveráková, I., Prokopchuk, G., Peña-Diaz, P., Záhonová, K., Moos, M., Horváth, A., Šimek, P. and Lukeš, J., 2020. Unique dynamics of paramylon storage in the marine euglenozoan Diplonema papillatum. *Protist*, *171*(2), 125717.

Škodová-Sveráková, I., Záhonová, K., Juricová, V., Danchenko, M., Moos, M., Baráth, P., Prokopchuk, G., Butenko, A., Lukáčová, V., Kohútová, L. and Bučková, B., 2021. Highly flexible metabolism of the marine euglenozoan protist Diplonema papillatum. *BMC biology*, *19*(1), 1-21.

Skuja, H., 1948. Taxonomie des phytoplanktons einiger seen in Uppland, Schweden. *Symb. Bot. Ups.*, *9*, 1-399.

Stuart, K., Brun, R., Croft, S., Fairlamb, A., Gürtler, R.E., McKerrow, J., Reed, S. and Tarleton, R., 2008. Kinetoplastids: related protozoan pathogens, different diseases. *The Journal of Clinical Investigation*, *118*(4).1301-1310.

Sunter, J.D. and Gull, K., 2016. The flagellum attachment zone:'the cellular ruler'of trypanosome morphology. *Trends in Parasitology*, *32*(4), 309-324.

Strauch, S.M., Schuster, M., Lebert, M., Richter, P., Schmittnagel, M. and Häder, D.P., 2008, June. A closed ecological system in a space experiment. In *Proceedings of the Symposium Life in Space for Life on Earth* (22-27).

Tashyreva, D., Prokopchuk, G., Votýpka, J., Yabuki, A., Horák, A. and Lukeš, J., 2018. Life cycle, ultrastructure, and phylogeny of new diplonemids and their endosymbiotic bacteria. *MBio*, *9*(2), 02447-17.

Tashyreva, D., Prokopchuk, G., Yabuki, A., Kaur, B., Faktorová, D., Votýpka, J., Kusaka, C., Fujikura, K., Shiratori, T., Ishida, K.I. and Horák, A., 2018. Phylogeny and morphology of new diplonemids from Japan. *Protist*, *169*(2), 158-179.

Tashyreva, D., Simpson, A.G., Prokopchuk, G., Škodová-Sveráková, I., Butenko, A., Hammond, M., George, E.E., Flegontova, O., Zahonova, K., Faktorova, D. and Yabuki, A., 2022. Diplonemids–A Review on" New" Flagellates on the Oceanic Block. *Protist*, *173*(2), 125868.

Teichert, I., Nowrousian, M., Pöggeler, S. and Kück, U., 2014. The filamentous fungus Sordaria macrospora as a genetic model to study fruiting body development. *Advances in Genetics*, *87*, 199-244.

Teichert, I., Pöggeler, S. and Nowrousian, M., 2020. Sordaria macrospora: 25 years as a model organism for studying the molecular mechanisms of fruiting body development. *Applied Microbiology and Biotechnology*, *104*, 3691-3704.

Valach, M., Moreira, S., Faktorová, D., Lukeš, J. and Burger, G., 2016. Post-transcriptional mending of gene sequences: looking under the hood of mitochondrial gene expression in diplonemids. *RNA Biology*, *13*(12), 1204-1211.

Valach, M., Moreira, S., Petitjean, C., Benz, C., Butenko, A., Flegontova, O., Nenarokova, A., Prokopchuk, G., Batstone, T., Lapébie, P. and Lemogo, L., 2023. Recent expansion of metabolic versatility in Diplonema papillatum, the model species of a highly speciose group of marine eukaryotes. *BMC Biology*, *21*(1), 99.

Wan, K.Y., 2018. Coordination of eukaryotic cilia and flagella. *Essays in Biochemistry*, *62*(6), 829-838.

Whittle, E.F., Chilian, M., Karimiani, E.G., Progri, H., Buhas, D., Kose, M., Ganetzky, R.D., Toosi, M.B., Torbati, P.N., Badv, R.S. and Shelihan, I., 2023. Biallelic variants in OGDH encoding oxoglutarate dehydrogenase lead to a neurodevelopmental disorder characterized by global developmental delay, movement disorder, and metabolic

abnormalities. *Genetics in Medicine*, *25*(2), 100332.

von der Heyden, S., Chao, E.E., Vickerman., 2004. Ribosomal RNA phylogeny of bodonid and diplonemid flagellates and the evolution of Euglenozoa. *Journal of Eukaryotic Microbiology*, *51*(4), 402-416.

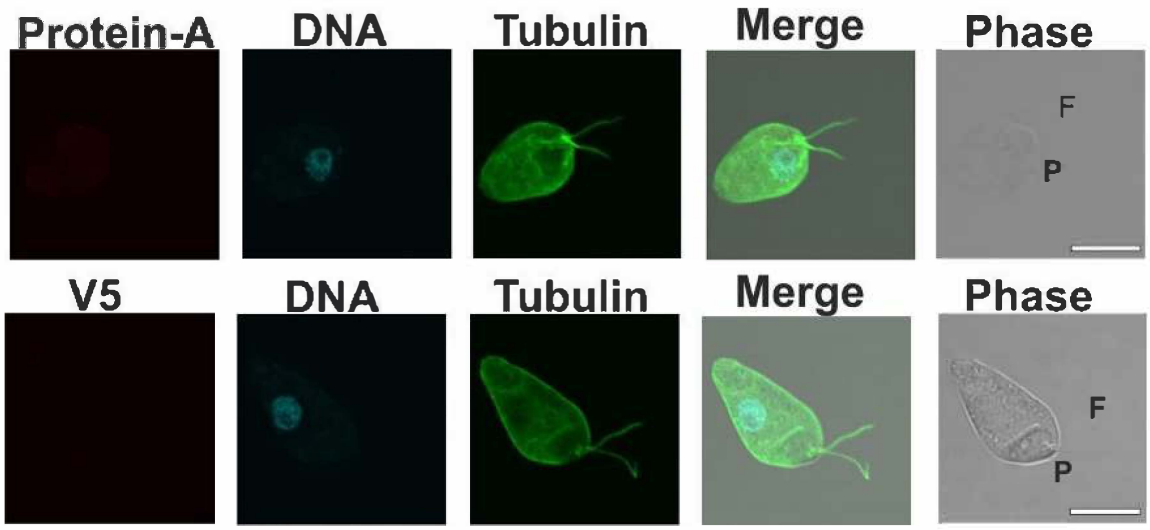
Yubuki, N., Čepička, I. and Leander, B.S., 2016. Evolution of the microtubular cytoskeleton (flagellar apparatus) in parasitic protists. *Molecular and Biochemical Parasitology*, *209*(1-2), 26-34.

Yubuki, N., Simpson, A.G. and Leander, B.S., 2013. Reconstruction of the feeding apparatus in Postgaardi mariagerensis provides evidence for character evolution within the Symbiontida (Euglenozoa). *European Journal of Protistology*, *49*(1), 32-39.

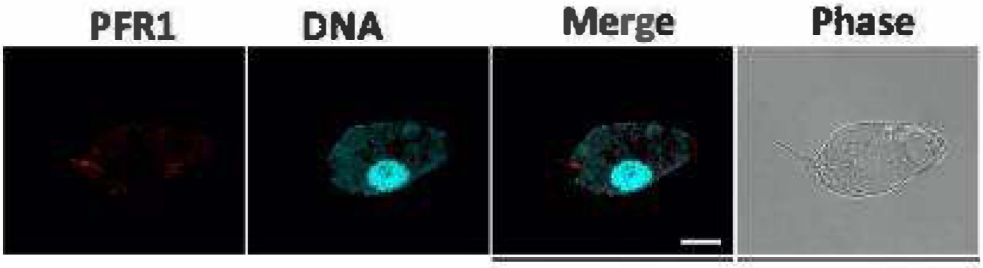
Zimorski, V., Rauch, C., van Hellemond, J.J., Tielens, A.G. and Martin, W.F., 2017. The mitochondrion of Euglena gracilis. *Euglena: Biochemistry, Cell and Molecular Biology*, 19-37.

Zoltner, M. and Field, M.C., 2022. Microbe Profile: Euglena gracilis: photogenic, flexible and hardy. *Microbiology*, *168*(9), 001241.

Supplementary Figures



Suppl. Figure 1. Wild type *P. papillatum* serving as a negative control for immunofluorescence-based experiments. Antibodies against protein-A and V5 showed barely visible signals in Wild type cells, while the anti-tubulin antibody showed expected localization of the target protein in the microtubule corset, flagella as well as in the apical cytopharynx and papilla.



Suppl. Figure 2. Localization of the tagged protein PFR1 (DIPPA_34257) by immunofluorescence via the red channel (AlexaFluor 555). DNA was tracked by DAPI via the blue channel. Merge shows the overlay of channels. Localization of the PFR1 protein with a signal concentrated near the papilla.

Suppl. Table 1. All primer sets employed for cloning in this study.

PFR1 ORF Forward	CCCCCGAAGCCTGCCTTTCAGGA
PFR1 ORF Reverse	CGATCCGCTACCCGACCCGGAGCCGCTGCCAGCCTGCGGCTGGTCTGCGATG
PFR1 3'UTR Forward	CTTGCCCGAGCGGGGTGTTACGTGCTGCAAGTTTAAGCTTGCGGAGATCCCCGCTCGCGT
PFR1 3'UTR Reverse	AGTGGCAGCGAGGGCTGTGTTGC
PFR1 Nested Forward	GGCGGGTGACGCTTCCGATCATTTTTG
PFR1 Nested Reverse	TGGCAGGGAAGGGTGGGAGAAGG
PFR2 ORF Forward	GGCGGTGTCGGAAGCCAAGAAGG
PFR2 ORF Reverse	CGATCCGCTACCCGACCCGGAGCCGCTGCCGCGGTCGGTGATGAGCTGGCCT
PFR2 3'UTR Forward	CTTGCCCGAGCGGGGTGTTACGTGCTGCAAGTTTAAGCTTAGCACTACCCACGCCCCCTGGA
PFR2 3'UTR Reverse	CGGCGGATTGCCAACACAGACTGG
PFR2 Nested Forward	GCCATTCGTGCCACCGGCATCATC
PFR2 Nested Reverse	CGGGGGAGAGTAGGGGTTGTTG
UPF1 ORF Forward	CACTGCGATGTCTCTGCGGATGTGAGAA

UPF1 ORF Reverse	CGATCCGCTACCCGACCCGGAGCCGCTGCCGCTGGTCTCGACGAGTAGGAAGGC
UPF1 3'UTR Forward	CTTGCCCGAGCGGGGTGTTACGTGCTGCAAGTTTAAGCTTTTCTCCAAGCTTTTTGTCGGAACCTTTGAACTCG
UPF1 3'UTR Reverse	GAGCGACGACGTCCTTCCTTTCCTTG
UPF1 Nested Forward	CAGGGCAGGCATTTTGTCTTGGCAATGG
UPF1 Nested Reverse	CGAATTAACGGCGGCAAAGT

PFR6 ORF Forward	AATGAAGCAGGTGTCCTGTTAAAGCGGTTTC
PFR6 ORF Reverse	CGATCCGCTACCCGACCCGGAGCCGCTGCCCGTGGGACTGTCCGGCTTGGG
PFR6 3'UTR Forward	CCGGCAGCAATGATACTTTTCGGCACTAGTACACAAACACGCGGGAGCTGCAGC
PFR6 3'UTR Reverse	TACTATCTGGCGAAACAGCAGTCAAGGGAA
PFR6 Nested Forward	TCGCTTCAACGGAAATTATAGAGGCGCATCT
PFR6 Nested Reverse	CTTCGGCGAAAGGTTGCTTTGTGGCTTT
PAR4 ORF Forward	CAGAGTACAGCACAATCGCCAGGCTGT
PAR4 ORF Reverse	CGATCCGCTACCCGACCCGGAGCCGCTGCCCGAAAGCTCGGCATTCGGCGCGAA
PAR4 3'UTR Forward	CCGGCAGCAATGATACTTTTCGGCACTAGTACGGAAGAAGAACCTGCTTCATATCCCAACT
PAR4 3'UTR Reverse	GAGCTTCCTATTGTCCTCCTCCACGAATC
PAR4 Nested Forward	GGTCCGTGAGTGTTCTAGTCTTCTTGATAGT
PAR4 Nested Reverse	TTGCAGCTTCATCGTGAGGTCGTGTATCT

Prot-A Forward	GGCAGCGGCTCCGGGTCGGGTAG
Prot-A Reverse	AAGCTTAAACTTGCAGCACGTAACACCCCGCTCGG
V5 Forward	GGCAGCGGCTCCGGGTCGGGTAG
V5 Reverse	ACTAGTGCCGAAAAGTATCATTGCTGCCGGTTCGCG

Chapter 4

Characterization of a putative kinetochore in diplonemids

The kinetochore complex of diplonemids

Introduction

A fundamental characteristic of living organisms is the ability to self-replicate, which facilitates the transmission of genetic information from one generation to the next (Chodasewicz, 2014; Tetz and Tetz, 2020). The kinetochore is a sophisticated macromolecular structure that develops on centromeric DNA, binds spindle microtubules to control chromosome movement and aids in the process of eukaryotic chromosome segregation (Cheeseman and Desai, 2008; Santaguida and Musacchio, 2009). This multi-protein complex thus directs the alignment of chromosomes, followed by their segregation during mitosis and meiosis (Matković *et al.*, 2022). Moreover, the kinetochores are responsible for facilitating the interaction between spindle microtubules and centromeric DNA (Maiato and Logarinho, 2014). The molecular mechanisms underlying all these processes have been extensively investigated in conventional model eukaryotes, such as yeasts, worms, flies, and humans (Maiato *et al.*, 2017).

The kinetochore is a mosaic and highly conserved structure (Tromer *et al.*, 2019), composed of more than 100 proteins that can be divided into three functional domains: the inner kinetochore, the outer kinetochore, and the microtubule-binding domain (Lara-Gonzalez, Westhorpe and Taylor, 2012). With very few exceptions (see below), the kinetochores contain a special histone H3 variation called CENP-A, which is unique to the centromeric region, and offers a particular chromatin environment for the kinetochore's assembly (Hori and Fukagawa, 2012; Musacchio and Desai, 2017).

However, there are prominent exceptions, such as the kinetochores of the euglenozoan protists (Akiyoshi and Gull, 2013). While they retain several conserved proteins, including TbKT1, Nuf2p and TbKIN-C, they evolved a unique tripartite structure, consisting of a central kinetochore domain, an outer microtubule-binding domain, and an inner protein scaffold (Drinnenberg and Akiyoshi, 2017). Yet the euglenozoan kinetochores still perform their conserved function, namely ensuring proper chromosome segregation during the cell division (Cheeseman, 2014).

So far, the only group of euglenozoans where the kinetochore has been studied in some detail, are the trypanosomatid flagellates, which bring together human pathogens such as *Trypanosoma brucei*, *Trypanosoma cruzi*, and *Leishmania* spp. Their sister clade of bodonids (e.g. *Bodo saltans* and *Perkinsela* spp.) have been analysed only using bioinformatics

(Butenko *et al.*, 2020) (Figure 1). All of these protists lack any CENP-A homolog, although they contain other core proteins, such as CENP-C, Ndc80 and Spc24/25 (Berriman *et al.*, 2005; Jackson *et al.*, 2012). These proteins likely form a structure able to attach to the centromeric DNA and interact with the spindle microtubules (Lowell and Cross, 2004). The inner kinetochore proteins, such as CENP-A, CENP-C, and CENP-H, play a role in centromere assembly and maintenance, while the outer kinetochore components, including Ndc80 and Spc24/25, interact with spindle microtubules (Akiyoshi and Gull, 2013).

The microtubule-binding domain is a network of proteins that connects the inner and outer domains to the microtubules. Again, some of its conserved components are absent from the euglenozoan kinetochore, while for others, such as the kinetochore-associated protein KKT4 (Kinetoplastid Kinetochore protein 4), confined to the euglenozoans, the function remains unclear (Akiyoshi and Gull, 2014; Nerusheva, Ludzia and Akiyoshi, 2019). Moreover, there are conflicting reports as to the presence of the KKT proteins in the free-living euglenozoan *Euglena gracilis*. While putative homologs of KKT10 and KKT19 were identified in its genome (Ebenezer *et al.*, no date), they were not encountered in the corresponding transcriptome (Akiyoshi, 2016). Besides Kinetoplastea, recognizable homologs of KKT10 and KKT19 have been identified only in related diplonemids and euglenids, which seem to lack other KKT proteins (Butenko *et al.*, 2020).

Another peculiar feature of the *T. brucei* kinetochore is several KKT proteins with limited if any sequence similarity with those associated with kinetochores of other organisms (Ishii and Akiyoshi, 2020). Similarly, all of its spindle checkpoint proteins seem to be absent, with the exception of Mad2 (Howell *et al.*, 2000).

Except the *in silico* analyses (Butenko *et al.*, 2020), the composition of kinetochore proteins in diplonemids remains unknown. To address this knowledge gap, we are currently investigating the composition of kinetochore proteins in *Paradiplonema papillatum* through endogenous tagging. Recently, this species became the only diplonemid for which methods of integration of extraneous DNA have been developed (Kaur *et al.*, 2018a), thus turning it into a new model marine protist (Faktorová *et al.*, 2020). Indeed, in *P. papillatum*, insertion of any DNA segment, in this case tagged genes, into a chromosome *via* homologous recombination can be easily achieved by using ~1.5 kb-long overlaps (Faktorová *et al.*, 2020). By doing so, we can endogenously tag or delete specific genes of interest, allowing us first functional studies in this group of evolutionary and ecologically highly relevant protists (Flegontova *et al.*, 2020). The tagged proteins allow the identification, by mass spectrometry analysis, of their interacting proteins or even protein complexes (see below). In this study, we

made use of the newly available methodology and explored the localization and protein components of the kinetochore complex in *P. papillatum*.

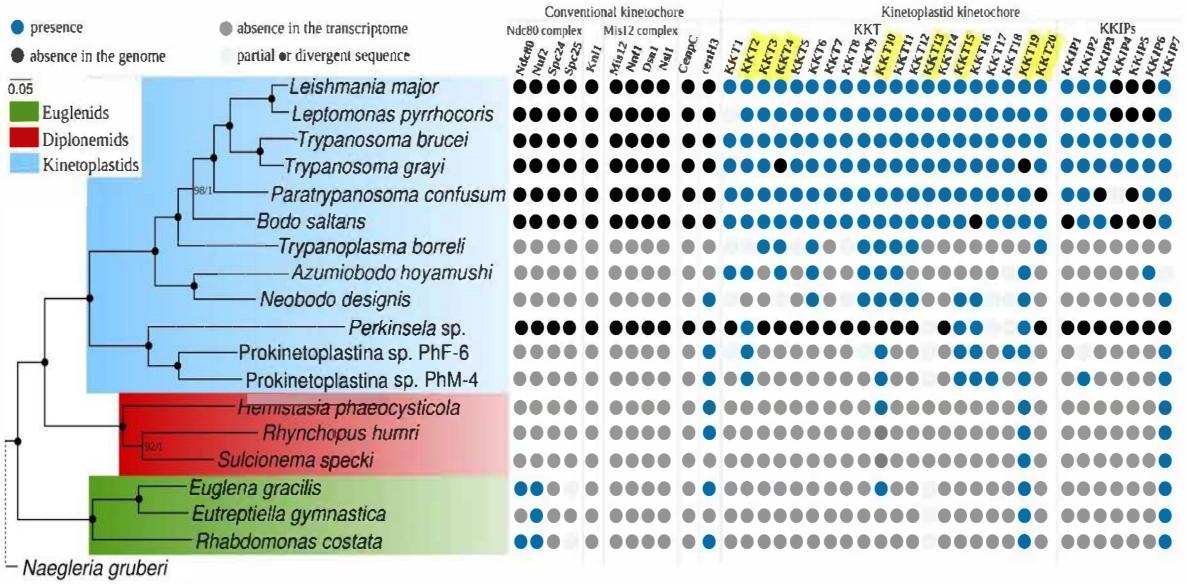


Figure 1. The distribution of kinetochore machinery components among Euglenozoa. Proteins that are present are denoted by blue circles, while those absent from a genome/transcriptome are denoted by black and gray circles, respectively. Blue-hatched circles indicate the presence of partial or divergent sequences. Kinetoplastid kinetochore proteins containing recognizable domains are marked by yellow circles. Abbreviations: cenH3, the centromeric variant of histone H3; CenpC, a centromere-associated protein; Dsn1, a dosage suppressor of NNF1; KKIPs, kinetoplastids kinetochore-interacting proteins; KKTs, kinetoplastid kinetochore proteins; Knl1, a kinetochore scaffold protein; Ndc80, a nuclear division cycle protein; Nnf1, necessary for nuclear function 1; Nsl1, a synthetic lethal protein with Nnf1; Nuf2, a nuclear filament-containing protein; Spc24, a homolog of spindle pole body component 24; and Spc25, a homolog of spindle pole body component 25 (Butenko *et al.*, 2020).

Material and Methods

Tagging of KKT-10, KKT-17, Mad2 and CENP-A

Parts of the open reading frame (ORF) and 3' untranslated region (UTR) of *P. papillatum* genomic DNA were amplified using specific primers containing sequences overlapping with the protein A-neomycin cassette of the plasmid pDP002 to endogenously tag KKT-10, KKT17, and Mad2 with a C-terminal protein A-tag. Using primers Fw protA-Neo cassette

and Rv protA-Neo cassette, another PCR was performed to amplify the protein A-neomycin cassette from pDP002 (Faktorová *et al.*, 2020). To ligate all three fragments together, Phusion polymerase (NEB®) was used in a nested PCR using these fragments as a template, and the PCR product was A-tailed and cloned into pcrTM 2.1-TOPO™ (ThermoFisher). Restriction enzyme digestion and sequencing were used to confirm the correct structure of the generated plasmid. To release the tagging segment, 10 µg of the final plasmid was cut with EcoRI (NEB), ethanol precipitated, resuspended in 10 µl of water, and used to transfect *P. papillatum* as described elsewhere (Faktorová *et al.*, 2020).

Electroporation of *p. papillatum*

Amaxa Nucleofector® II was used to transform a total of 5 x 10^7 cells, as previously described by (Kaur *et al.*, 2018b; Faktorová *et al.*, 2020). Clones were selected in 24-well plates at 27 ºC using varied doses of G418 (45 to 83 g/ml). Successful transfectants could be seen after 2 weeks of selection. Before being tested by western blot, each clone was expanded to a volume of 20 ml and cultured for up to 2 to 3 weeks.

Immunofluorescence

A 5 to 10 mL log phase culture was centrifuged at 1000 x g for 5 min. Cells were fixed in 4% paraformaldehyde in seawater at room temperature for 30 min. The fixative was washed out of the cells with seawater and 1 x PBS (Phosphate buffered saline) (1:1). The cells were rinsed once more in 1 x PBS before being spotted on a gelatine-coated slide, and permeabilized for 20 min for antibody labelling in 100% ice-cold methanol. Throughout the procedure, the slides were stored in a humidified cabinet, washed in PBS after 20 min. and blocked in 5% milk-PBS-T (0.05% Tween in PBS) for 45 min (0.05% Tween). After removing the blocking solution, the cells were again washed in 1 x PBS. The primary antibody (anti-protein A; Sigma®, 1:2,000) was applied on slides covered with parafilm and incubated overnight at 4°C. After removing the primary antibody, the slides were washed three times with 1 x PBS. The secondary antibody (AlexaFluor555- goat-anti-rabbit; Invitrogen, 1:1,000) was added and the slides were incubated for 1 hour in the dark at room temperature, covered with parafilm. Finally, the slides were washed in 1 x PBS and covered with 4'6-diamidino-2-phenylindole (DAPI) containing the antifade reagent ProlongGold (Life Technologies) after the secondary antibody. Images were obtained using camera Olympus DP73 (Axioplan 2 imaging).

Western blot

Protein samples from 5 x 10^5 cells were prepared by pelleting and resuspending them in 25 µl of 2 x SDS sample buffer, separated on 4-20 % Mini-protein TGX stain-free gels (Bio-Rad), and subsequently transferred to a PVDF membrane. After blocking with 5% milk in PBS-T for at least 30 min at room temperature, the membrane was incubated with an anti-protein A antibody (Sigma®; 1:10,000) in 5% milk in PBS-T overnight at 4^0C. After 3 washes in PBS-T, the membrane was incubated with anti-rabbit-HRP (Sigma®; 1:1,000) and incubated at room temperature for 1 hour. The membrane was then washed 3x in PBS-T and the signal was developed using Clarity Western ECL Substrate (Bio-Rad). The anti-tubulin mouse antibody (Sigma-Aldrich; 1:10,000) was used as a loading control.

Immunoprecipitation

Approximately 5 x 10^8 cells expressing V-tagged Mad2 or Cenp-A proteins, as well as the wild-type control cells were grown in Diplonema growth media with the appropriate selection antibiotic (neomycin for Protein-A tag and hygromycin for V5 tag). Cells were harvested at 1,000 g for 10 min, resuspended in 5 ml ice-cold 1 x PBS, centrifuged again at 1,000 g for 10 min, and the supernatant was discarded. The cells were then lysed using lysis buffer (10 mM Tris [pH 6.8], 150 mM NaCl, 0.1% Igepal, 1% (v/v) glycerol) and passing through a 30 gauge needle several times. The cell lysate was cleared at 10,000 g for 20 min. 50 µl V5-tagged magnetic bead (Sigma) was added to the cleared cell lysate and rotated at 4°C for 2 to 3 hours. The beads were washed 3x with the washing buffer (10 mM Tris [pH 6.8], 250 mM NaCl, 1% (v/v) glycerol) supplemented with 0.1% (v/v) igepal, and then twice with the washing buffer without the detergent. Bound proteins with beads were processed for immunoblotting and mass spectrometry.

Mass-spectrometry analysis

The eluted protein V-tagged Mad2, Cenp-A, and wild-type control underwent trypsin-digestion before liquid chromatography-tandem mass spectroscopy (LC-MS/MS) analysis following previously described methods (Pyrih *et al.*, 2020). MaxQuant v1.6.14 (Cox, J., Mann, M.,2008) was employed for data processing, which employed the Andromeda search engine (Cox *et al.*, 2011) for protein identification. The protein sequence database of *P. papillatum* (comprising 43,871 sequences) was custom-made and supplemented with frequently observed contaminants.

The search parameters used specified an MS tolerance of 6 ppm, an MS/MS tolerance of 0.5 Da, and full trypsin specificity with allowance for up to two missed cleavages. Fixed

modification of cysteine carbamidomethylation and variable modifications such as methionine oxidation and N-terminal protein acetylation were set. The experimental design included matching between runs for biological replicates. Peptides were required to be at least seven amino acids long, and false discovery rates (FDRs) of 0.01 were calculated at the peptide, protein, and modification site levels based on the number of hits against the reversed sequence database. iBAQ indices were employed for protein quantification, which involved raw intensities divided by the number of theoretical peptides, enabling the comparison of protein abundance within and between samples. Finally, Perseus v1.6.14 was used for further data processing, as previously described (Zoltner *et al.*, 2020).

Results

C-terminal endogenous tagging of KKT-10, KKT-17, Mad2 and CENP-A

By bioinformatic analysis, we identified sequences of the kinetochore proteins in *P. papillatum* homologous to *T. brucei* and other eukaryotes. We narrowed it down to 2 kinetochore CLK-like kinase proteins KKT-10, and KKT-17, one mitotic arrest deficient (Mad2) protein, and one centromere protein-A (CENP-A). For experimental analysis, we have endogenously tagged KKTs, Mad2 and CENP-A following our previously described strategies (Faktorová *et al.*, 2020) as follows: KKT-10, KKT-17, and Mad2 were tagged with the protein-A tag at their C terminal end, while CENP-A was tagged with V5.

Plasmid pDP002 was used as a template with homologous sequences to the 5' and 3' flanking regions of the corresponding gene with a neomycin resistance cassette, and plasmid pD011 was used for CENP-A protein. We assayed the expression of the tagged proteins by western blotting using an antibodies against the common protein A or V5 tag. We selected 3 random cell lines to check for the protein expression. Western blot of a representative clone for each tagged protein is shown in (Figures 1- A and B). Tubulin was used as a loading control for all cell lines. Although we did not quantify the protein expression level, all three proteins seemed to be expressed in western blot images.

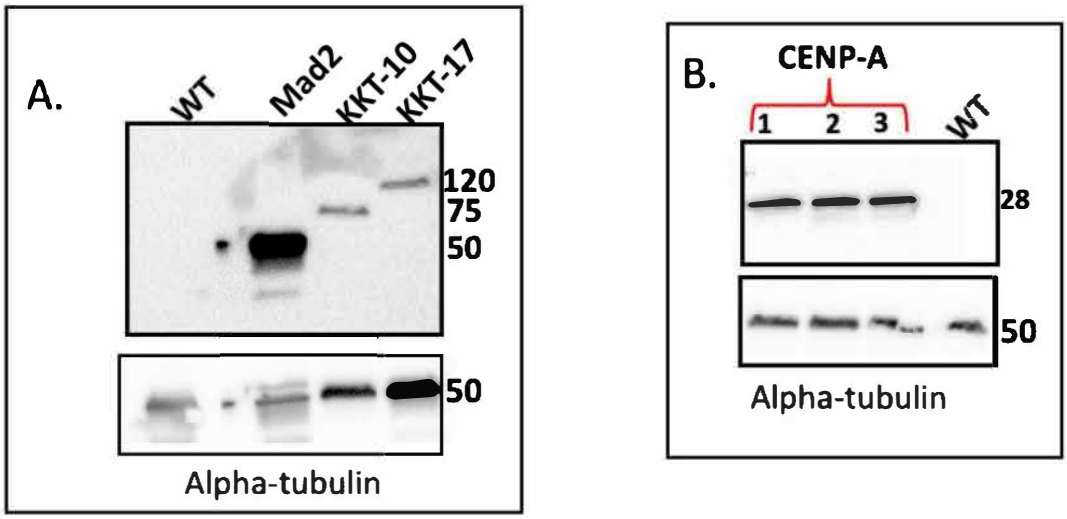


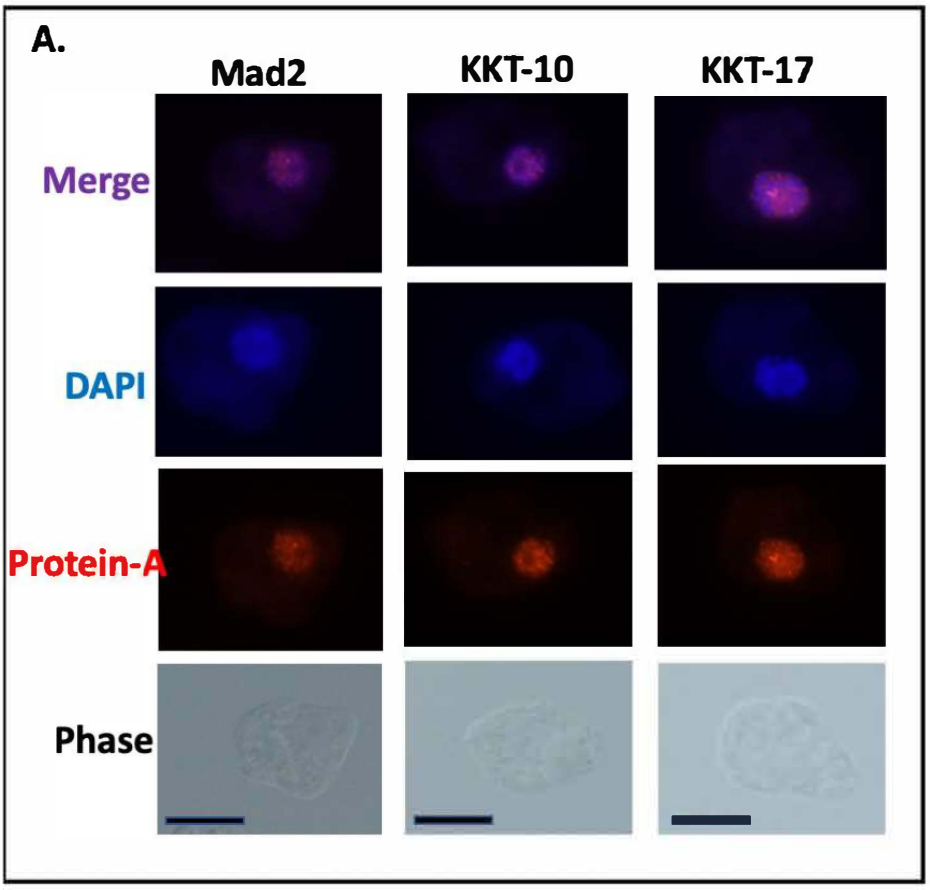
Figure 1-A. Western blot analysis of *P. papillatum* wild-type and transformant cell lines Mad2, KKT-10, and KKT-17 (primary anti-rabbit protein-A antibody; 1:10,000). Anti-alpha tubulin (1:10,000) was used as a loading control. **B.** Western blot analysis of *P. papillatum* wild-type and transformant cell lines CENP-A (primary anti-mouse V5; 1:1,000).

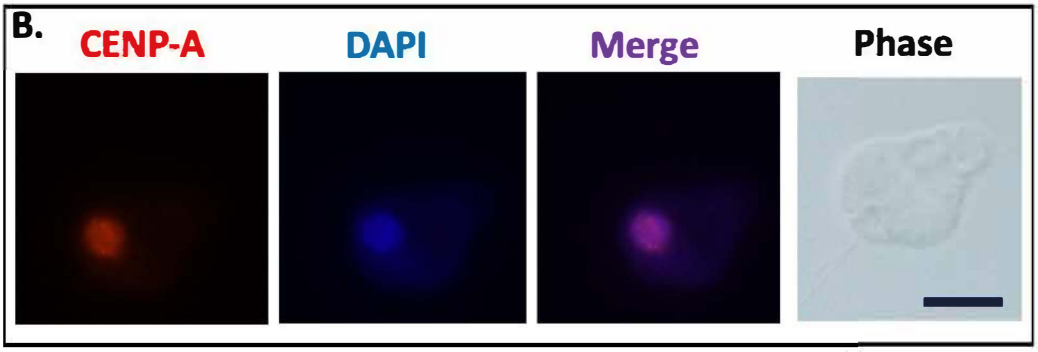
Localization of KKT-10, KKT-17, Mad2 and CENP-A

Previous studies showed that in *T. brucei* KKT10 and KKT-17 are in one complex with other proteins (Akiyoshi and Gull, 2014). We were curious if this close association is functionally also present in *P. papilatum.* Therefore, we performed immunoprecipitation of the cell lines expressing tagged KKT-10, KKT-17, Mad2, and CENP-A, and analysed the pull downs by mass spectrometry. Unfortunately, the mass-spectrometric analysis did not show any significant interacting partners with the KKT proteins (data not shown).

In *T. brucei* KKT-10, and KKT-17 is localized in the nucleus, while Mad2 is associated with the basal body area, and CENP-A is localized in the centromere region of the chromosome (Zhang *et al.*, 2018; Mahlke and Nechemia-arbely, 2020; Mitra, Srinivasan and Jansen, 2020). Therefore, we performed an immunofluorescence assay to see the localization of KKT-10, KTT-17, Mad2, and CENP-A. Consistent with their localisation in *T. brucei,* we also found KKT-10 and KKT-17 localized in the nucleus, but Mad2 showed an unexpected nuclear localization in *P. papillatum*, as confirmed by co-localization with DAPI (Figure 2A). Moreover, fluorescent and confocal microscopy analysis revealed a distinct nuclear

localization pattern for CENP-A. The protein was observed to be enriched in discrete foci throughout the nucleus (Figures 2B and C).





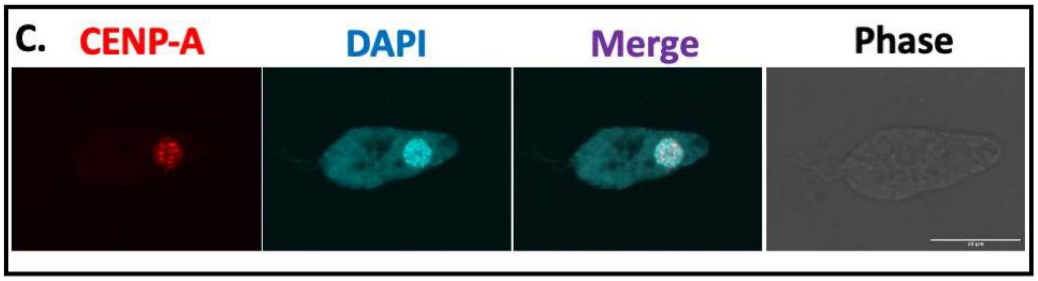


Figure 2. A-B. Localization of Mad2, KKT-10, KKT-17, and CENP-A by immunofluorescence. All transformed cell lines expressing the protein-A tag, except CENP-A which was tagged with V5. Target proteins localize to the nucleus (red signal). Scale bar: 10 μm. **C.** Localization of CENP-A was visualized using a confocal microscope to investigate its distinct distribution pattern within the nucleus. Scale bar: 10 μm.

Pull-down of tagged Mad2

In this study, we sought to investigate how kinetochore and Mad2 proteins interact with one another. In order to achieve this, we used a pull-down assay with the C-terminally-tagged Mad2 protein from *P. papillatum* and LC-MS/MS analysis of the protein complexes that were successfully captured. Our rationale was that if the Mad2 protein was a kinetochore or spindle checkpoint machinery member, it would capture all the complexes connected with it. Conversely, if Mad2 was solely connected with certain complexes, such as Mad1 or CdC20, which are spindle checkpoint proteins known from other eukaryotes, it would only capture that particular complex. The pull-down fraction was subjected to immunoblotting, revealing an enrichment of the Mad2 protein in the eluate as well as in the mass-spectrometry data (Figure 3).

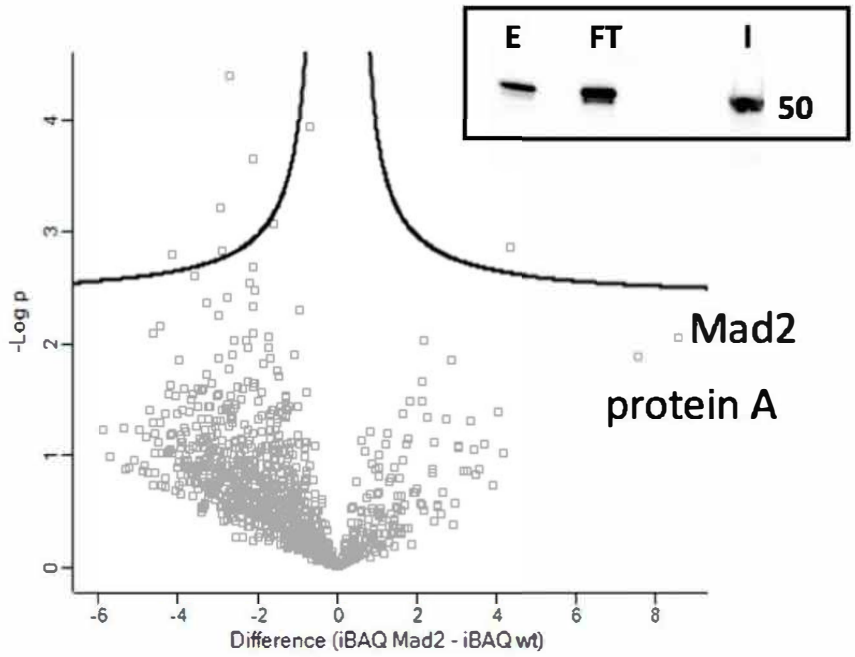


Figure 3. Volcano plots showing the differential expression of proteins and control. Western blot analysis of Pull-down Mad2 protein in *p. papillatum*. Abbreviations are I- Input, FT-flow through, and E- elution.

Pull down of V-5 tagged CENP-A

The aim of pulling down CENP-A tagged with V5 protein is to investigate its protein-protein interactions and identify potential binding partners that are involved in centromere function and maintenance. CENP-A is a critical component of the centromere, and its proper localization and interactions are essential for accurate chromosome segregation during cell division. By pulling down CENP-A and analyzing its associated proteins, we found an enrichment of DIPPA_32769.mRNA.1 Histone H3.2 ("CENP-A" bait, tagged with V5) itself, DIPPA_34374.mRNA.1 Histone H2B protein, probably DIPPA_34783.mRNA.1 histone H2A.2, DIPPA_14460.mRNA.1 Sideroflexin, DIPPA_27824.mRNA.1 Histone deacetylase HDA1, DIPPA_27824.mRNA.3 Histone deacetylase HDA1, DIPPA_03119.mRNA.1 DNA topoisomerase 1 and some hypothetical protein.

The enrichment of these proteins suggests that they may be part of the centromere complex and may contribute to the formation of a stable centromere structure in *P. papillatum*. However, further investigations are needed to confirm these findings and to better understand the role of these proteins in centromere function.

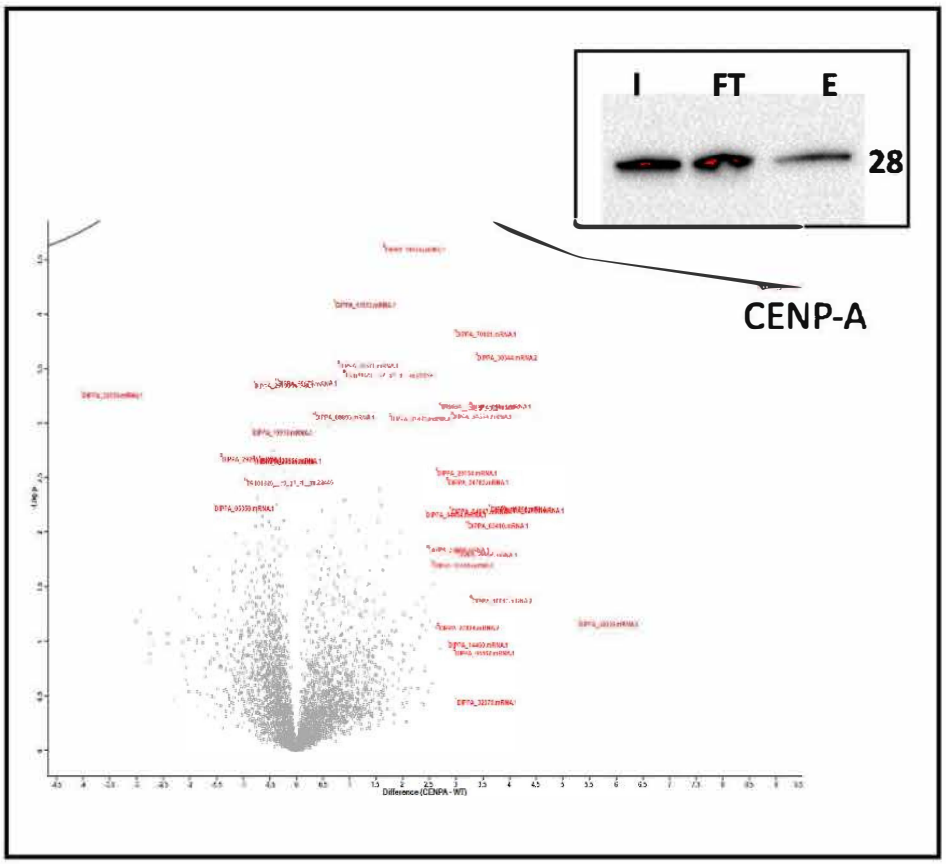


Figure 4. Volcano plots showing the differential expression of proteins and control. Western blot analysis of Pull-down CENP-A protein in *p. papillatum*. Abbreviations: I - Input; FT -flow through; and E - elution.

Investigation of the co-localization of Mad2 and KKT-10

To examine the potential interaction between Mad2 and KKT-10, we created a V5-tagged Mad2 cell line and subsequently generated a protein-A-tagged KKT-10 cell line using the Mad2 cell line as a parental line (Figure 5A). Co-immunoprecipitation was performed to assess the interaction between both proteins. Our Input and Flow-through data confirmed the expression of both Mad2 and KKT-10 proteins. However, no band of the expected size for KKT-10 at 75 kDa was observed in the elution. Conversely, a strong band of Mad2 was detected in the elution, indicating that Mad2 does not directly interact with KKT-10 in *P. papillatum* (Figure 5B).

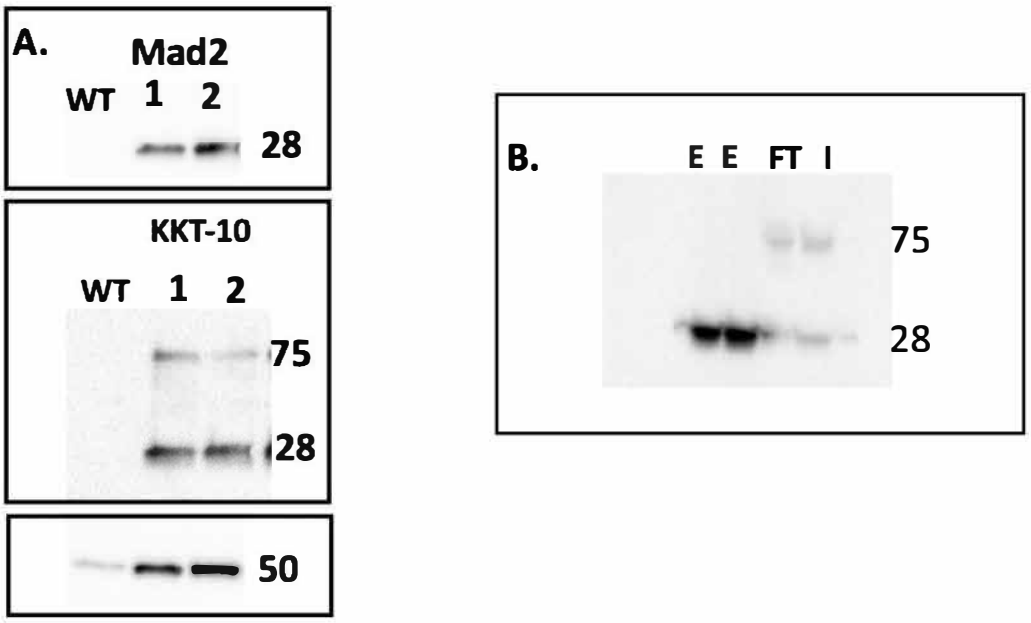


Figure 5. A. Western blot analysis of *P. papillatum* wild-type and transformant cell lines carrying Mad2 tagged with V5 (primary anti-mouse antibody; 1:1,000), and co-transformant cell lines KKT-10 tagged with protein-A ((primary anti-rabbit protein-A antibody; 1:10,000). **B.** Western blot analysis of co-localization of Mad2 tagged with V5 and KKT-10 tagged with protein-A. Abbreviation: I – Input; FT – flow-through; E - elution.

Discussion

T. brucei contains a unique assortment of kinetochore proteins KKT1 through KKT25, among which there are 6 protein kinases, namely KKT10, KKT19, KKT2, KKT22, KKT025, and KKT3 (Musacchio and Desai, 2017). The proteins KKT2 and KKT3, which localize to the centromere, are crucial for the development of its kinetochore (Marcianò *et al.*, 2021). It was also shown that the knock-down of KKT10 or KKT19 has a strong effect on growth of both the procyclic (insect) stage and the bloodstream (mammalian) stage of the parasite (Akiyoshi and Gull, 2014). Overall, kinase proteins play a crucial role in cell growth in analysed eukaryotes as well as in *T. brucei.*

CENP-A is a highly conserved histone variant that replaces the canonical histone H3 at centromeric chromatin. It is essential for proper kinetochore assembly, spindle attachment, and faithful chromosome segregation during the mitosis (Hori and Fukagawa, 2012; Shrestha

et al., 2021; Ishii and Akiyoshi, 2022). CENP-A-containing nucleosomes have distinct biophysical properties and epigenetic marks compared to canonical nucleosomes, contributing to their specific localization and function at centromeres. Recent studies have shed light on the complex higher-order structure of centromeric chromatin and the role of CENP-A in its organization and maintenance (Maiato *et al.*, 2017). Moreover, emerging evidence suggests that CENP-A may have non-centromeric functions, such as regulating gene expression, DNA repair, and chromosome organization (Renaud-Pageot *et al.*, 2022). Overall, CENP-A is a crucial protein for faithful chromosomal segregation and genome stability, with important implications for human health and disease (Mahlke and Nechemia-arbely, 2020). Its complex regulation and function continue to be an active area of research in the field of chromatin biology.

We wanted to see how kinetochores look like in diplonemids, with our approach based on kinetochore proteins for which homologs could be identified *in silico*. While we were able to demonstrate the localization of tagged KKT10, KKT19, Mad2, and CENP-A in the nucleus of *P. papillatum*, our attempts to identify genuine interacting proteins by mass spectrometry analysis so far failed. However, the pull-down assays might have missed intriguing possibilities due to ineffective incorporation into the protein complex, possibly caused by several reasons, such as the size of the tag (e.g. protein-A; 25 kDa), the used buffers and immunoprecipitation protocols. Another hypothesis is that a cluster of proteins may undergo self-assembly and form a complex, analogous to the spontaneous formation of the Dam complex in bacteria, which is involved in DNA replication and repair (Palmer and Marinus, 1994). We also tried the cryo-grinding technique which is used to disrupt and lyse cells for subsequent biochemical analysis (Phillips *et al.*, 2021). In this method, cells are first frozen in liquid nitrogen to minimize the degradation of cellular components. Next, the frozen cells are ground to a fine powder using a mortar and pestle that has been pre-chilled in liquid nitrogen. The resulting cell powder is subsequently used for further analysis for immunoprecipitation. To accomplish a pull-down, we tried a variety of buffer compositions, but sadly, all failed (data not shown).

One possible alternative approach is to see if the cells under investigation have a checkpoint response, although it can be difficult to spot such cells. This is due to the likely possibility that only a rather small portion of the culture actually undergoes mitosis. This issue could be resolved by using synchronization using various drugs, such as nocodazole, roscovitine and others, which would halt a greater percentage of cells in the target cell cycle

phase. Further experimental approaches will include gene knock-outs, once this technique is available for diplonemids.

References

Akiyoshi, B., 2016. The unconventional kinetoplastid kinetochore: from discovery toward functional understanding. *Biochemical Society Transactions*, *44*(5), 1201-1217.

Akiyoshi, B. and Gull, K., 2013. Evolutionary cell biology of chromosome segregation: insights from trypanosomes. *Open Biology*, *3*(5), 130023.

Akiyoshi, B. and Gull, K., 2014. Discovery of unconventional kinetochores in kinetoplastids. *Cell*, *156*(6), 1247-1258.

Berriman, M., Ghedin, E., Hertz-Fowler, C., Blandin, G., Renauld, H., Bartholomeu, D.C., Lennard, N.J., Caler, E., Hamlin, N.E., Haas, B. and Böhme, U., 2005. The genome of the African trypanosome Trypanosoma brucei. *Science*, *309*(5733), 416-422.

Butenko, A., Opperdoes, F.R., Flegontova, O., Horák, A., Hampl, V., Keeling, P., Gawryluk, R.M., Tikhonenkov, D., Flegontov, P. and Lukeš, J., 2020. Evolution of metabolic capabilities and molecular features of diplonemids, kinetoplastids, and euglenids. *BMC Biology*, *18*(1), 1-28.

Cheeseman, I.M., 2014. The kinetochore. *Cold Spring Harbor perspectives in biology*, *6*(7), a015826.
Cheeseman, I.M. and Desai, A., 2008. Molecular architecture of the kinetochore–microtubule interface. *Nature Reviews Molecular Cell Biology*, *9*(1), 33-46.

Krzysztof, C., 2014. Evolution, reproduction and definition of life. *Theory in Biosciences*, *133*(1), 39-45.

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R.A., Olsen, J.V. and Mann, M., 2011. Andromeda: a peptide search engine integrated into the MaxQuant environment. *Journal of Proteome Research*, *10*(4), 1794-1805.

Drinnenberg, I.A. and Akiyoshi, B., 2017. Evolutionary lessons from species with unique kinetochores. *Centromeres and Kinetochores: Discovering the Molecular Mechanisms Underlying Chromosome Inheritance*, 111-138.

Ebenezer, T.E., Zoltner, M., Burrell, A., Nenarokova, A., Novák Vanclová, A.M., Prasad, B., Soukal, P., Santana-Molina, C., O'Neill, E., Nankissoor, N.N. and Vadakedath, N., 2019. Transcriptome, proteome and draft genome of Euglena gracilis. *BMC Biology*, *17*, 1-23.

Faktorová, D., Kaur, B., Valach, M., Graf, L., Benz, C., Burger, G. and Lukeš, J., 2020. Targeted integration by homologous recombination enables in situ tagging and replacement of genes in the marine microeukaryote Diplonema papillatum. *Environmental Microbiology*, *22*(9), 3660-3670.

Flegontova, O., Flegontov, P., Londoño, P.A.C., Walczowski, W., Šantić, D., Edgcomb, V.P., Lukeš, J. and Horák, A., 2020. Environmental determinants of the distribution of planktonic diplonemids and kinetoplastids in the oceans. *Environmental Microbiology*, *22*(9), 4014-4031.

Hori, T. and Fukagawa, T., 2012. Establishment of the vertebrate kinetochores. *Chromosome Research*, *20*, 547-561.

Howell, B.J., Hoffman, D.B., Fang, G., Murray, A.W. and Salmon, E.D., 2000. Visualization of Mad2 dynamics at kinetochores, along spindle fibers, and at spindle poles in living cells. *The Journal of Cell Biology*, *150*(6), 1233-1250.

Ishii, M. and Akiyoshi, B., 2020. Characterization of unconventional kinetochore kinases KKT10 and KKT19 in Trypanosoma brucei. *Journal of Cell Science*, *133*(8), jcs240978.

Jackson, A.P., Berry, A., Aslett, M., Allison, H.C., Burton, P., Vavrova-Anderson, J., Brown, R., Browne, H., Corton, N., Hauser, H. and Gamble, J., 2012. Antigenic diversity is generated by distinct evolutionary mechanisms in African trypanosome species. *Proceedings of the National Academy of Sciences*, *109*(9), 3416-3421.

Kaur, B., Valach, M., Peña-Diaz, P., Moreira, S., Keeling, P.J., Burger, G., Lukeš, J. and Faktorová, D., 2018. Transformation of Diplonema papillatum, the type species of the highly diverse and abundant marine microeukaryotes Diplonemida (Euglenozoa). *Environmental Microbiology*, *20*(3), 1030-1040.

Lara-Gonzalez, P., Westhorpe, F.G. and Taylor, S.S., 2012. The spindle assembly checkpoint. *Current Biology*, *22*(22), R966-R980.

Lowell, J.E. and Cross, G.A., 2004. A variant histone H3 is enriched at telomeres in Trypanosoma brucei. *Journal of Cell Science*, *117*(24), 5937-5947.

Mahlke, M.A. and Nechemia-Arbely, Y., 2020. Guarding the genome: CENP-A-chromatin in health and cancer. *Genes*, *11*(7), 810.

Maiato, H., Gomes, A.M., Sousa, F. and Barisic, M., 2017. Mechanisms of chromosome congression during mitosis. *Biology*, *6*(1), 13.

Maiato, H. and Logarinho, E., 2014. Mitotic spindle multipolarity without centrosome amplification. *Nature Cell Biology*, *16*(5), 386-394.

Marciano, G., Ishii, M., Nerusheva, O.O. and Akiyoshi, B., 2021. Kinetoplastid kinetochore proteins KKT2 and KKT3 have unique centromere localization domains. *Journal of Cell Biology*, *220*(8), e202101022.

Matković, J., Ghosh, S., Ćosić, M., Eibes, S., Barišić, M., Pavin, N. and Tolić, I.M., 2022. Kinetochore-and chromosome-driven transition of microtubules into bundles promotes spindle assembly. *Nature Communications*, *13*(1), 7307.

Mitra, S., Srinivasan, B. and Jansen, L.E., 2020. Stable inheritance of CENP-A chromatin: Inner strength versus dynamic control. *Journal of Cell Biology*, *219*(10).

Musacchio, A. and Desai, A., 2017. A molecular view of kinetochore assembly and function. *Biology*, *6*(1), 5.

Nerusheva, O.O., Ludzia, P. and Akiyoshi, B., 2019. Identification of four unconventional kinetoplastid kinetochore proteins KKT22–25 in Trypanosoma brucei. *Open Biology*, *9*(12), 190236.

Phillips, E.O., Giovinazzi, S., Menz, S.L., Son, Y. and Gunjan, A., 2021. Preparation of cell extracts by cryogrinding in an automated freezer mill. *J. Vis. Exp*, *167*, e61164.

Pyrih, J., Rašková, V., Škodová-Sveráková, I., Pánek, T. and Lukeš, J., 2020. ZapE/Afg1 interacts with Oxa1 and its depletion causes a multifaceted phenotype. *Plos One*, *15*(6), e0234918.

Renaud-Pageot, C., Quivy, J.P., Lochhead, M. and Almouzni, G., 2022. CENP-A regulation and cancer. *Frontiers in Cell and Developmental Biology*, 1165.

Santaguida, S. and Musacchio, A., 2009. The life and miracles of kinetochores. *The EMBO Journal*, *28*(17), 2511-2531.

Shrestha, R.L., Rossi, A., Wangsa, D., Hogan, A.K., Zaldana, K.S., Suva, E., Chung, Y.J., Sanders, C.L., Difilippantonio, S., Karpova, T.S. and Karim, B., 2021. CENP-A overexpression promotes aneuploidy with karyotypic heterogeneity. *Journal of Cell Biology*, *220*(4).

Tetz, V.V. and Tetz, G.V., 2020. A new biological definition of life. *Biomolecular Concepts*, *11*(1), 1-6.

Tromer, E.C., van Hooff, J.J., Kops, G.J. and Snel, B., 2019. Mosaic origin of the eukaryotic kinetochore. *Proceedings of the National Academy of Sciences*, *116*(26), 12873-12882.

Zhang, Y., Huang, Y., Srivathsan, A., Lim, T.K., Lin, Q. and He, C.Y., 2018. The unusual flagellar-targeting mechanism and functions of the trypanosome ortholog of the ciliary GTPase Arl13b. *Journal of Cell Science*, *131*(17), jcs219071.

Zoltner, M., Campagnaro, G.D., Taleva, G., Burrell, A., Cerone, M., Leung, K.F., Achcar, F., Horn, D., Vaughan, S., Gadelha, C. and Zíková, A., 2020. Suramin exposure alters cellular metabolism and mitochondrial energy production in African trypanosomes. *Journal of Biological Chemistry*, *295*(24), 8331-8347.

.

Conclusions and future directions

Diplonemids have gained recognition for their remarkable diversity and wide distribution, establishing themselves as a prominent group within the euglenozoans (Kostygov *et al.*, 2021). Due to a lack of comprehensive genomic data, diplonemid research has been somewhat restricted. However, with the recent release of the whole genome, we now have an opportunity to discover more about their biology (Valach, Moreira, Petitjean, Benz and Butenko, 2023). The comprehensive analysis of the gene complement in *P. papillatum* has shed light on the so far unknown significance of carbohydrate degradation in this organism, enabling inferences regarding its ecological function. However, it is crucial to recognize that these findings may not be universally applicable to other diplonemids, as their carbohydrate-active enzyme (CAZyme) complements differ from that of *P. papillatum*. Furthermore, the investigated diplonemid species are limited to the Diplonemidae and Hemistasiidae families, which are experimentally tractable. In contrast, our understanding of the metabolic capabilities and ecological roles of the DSPDII group, especially the highly diverse Eupelagomenidae clade, remains extremely limited. Recent advancements in single-cell (SC) technologies, encompassing SC-genomics and SC-metabolomics approaches (Lähnemann, *et al*, 2020), hold the potential to bridge this knowledge gap in the near future. Considering the availability of genetic manipulation techniques, the acquisition of nuclear genome information for *P. papillatum* holds significant promise for facilitating systematic investigations into its functional and cell biology. Furthermore, this information will serve as a valuable reference for exploring the evolutionary dynamics of gene complements within the Kinetoplastida, the sister group of diplonemids encompassing diverse species, including those with implications for human health.

Additionally, significant progress has been achieved in developing diplonemid tagging procedures in recent years. The successful utilization of targeted integration by homologous recombination in *P. papillatum* enables gene tagging and replacement. This approach provides us with powerful tools to investigate gene function, protein localization, and cellular processes within this marine microeukaryote (Kaur *et al.*, 2018a; Faktorová *et al.*, 2020) It has been determined that the pyruvate dehydrogenase complex is only partially present in *P. papillatum* at the proteomic level. Following the approach of gene tagging, to elucidate the localization and presence of pyruvate dehydrogenase in *P. papillatum*, we conducted investigations aimed at obtaining relevant information. It is highly probable that the E3 protein serves as a common element among the three DH (Dihydrolipoamide dehydrogenase) complexes in diplonemids. Although the PDH (Pyruvate Dehydrogenase), OGDH (α-Ketoglutarate Dehydrogenase), and BCKDH (Branched-chain α-keto acid Dehydrogenase) complexes exhibit mechanistic and structural similarities, they perform distinct enzymatic reactions within the cellular context (Nemeria, *et al.*, 2018; Kinugawa *et al.*, 2020).
Based on the information available for alveolates, specifically dinoflagellates and apicomplexans, we aimed to investigate the PDH complex in *P. papillatum*. Dinoflagellates exhibit a wide range of metabolic capabilities and harbour noncanonical variants of the pyruvate dehydrogenase complex (PDH complex). It is hypothesized that these noncanonical forms have emerged through gene duplication events followed by neofunctionalization processes, thereby enabling dinoflagellates to adapt to diverse environmental conditions. In contrast, apicomplexans, which encompass parasitic organisms like Plasmodium and Toxoplasma, were traditionally presumed to have lost the PDH complex during their evolutionary transition toward a parasitic lifestyle (Danne *et al.*, 2012).

Typically, the activity of pyruvate dehydrogenase (PDH) is supported by pyruvate generated through the glycolysis pathway (Park S, *et al.*, 2018). However, in the case of *P. papillatum*, it differs from trypanosomatids as glycolysis is not the predominant route in the presence of glucose. The observed lower physiological abundance of the complete pyruvate dehydrogenase (PDH) complex may be attributed to this phenomenon, thereby introducing complexities in its analysis. In summary, while the OGDH and BCKDH complexes in diplonemids exhibit a conventional composition, the PDH complex has undergone divergent evolutionary changes (Figure 1). Rather than comprising E1p-α and β proteins as the E1 subunit, our findings demonstrate that diplonemids possess a prokaryotic-type aceE. This prokaryotic type aceE is highly likely to have been acquired through horizontal gene transfer from an archaeon, occurring in the last common ancestor of diplonemids (Figure 1). By conducting surveys on the distribution of dihydrolipoamide dehydrogenase (DH) components within understudied Euglenozoa, valuable information can be obtained regarding additional functional substitutions, losses of subunits, and the relationship between the acquisition of aceE and the loss of E2p, followed by subsequent replacement with E2o or E2b, as observed in diplonemids and dinoflagellates (Danne *et al.*, 2012). Future studies utilizing functional genetics hold the potential to offer a broader understanding of the mechanisms employed by organisms to incorporate foreign components into a vital enzyme complex.

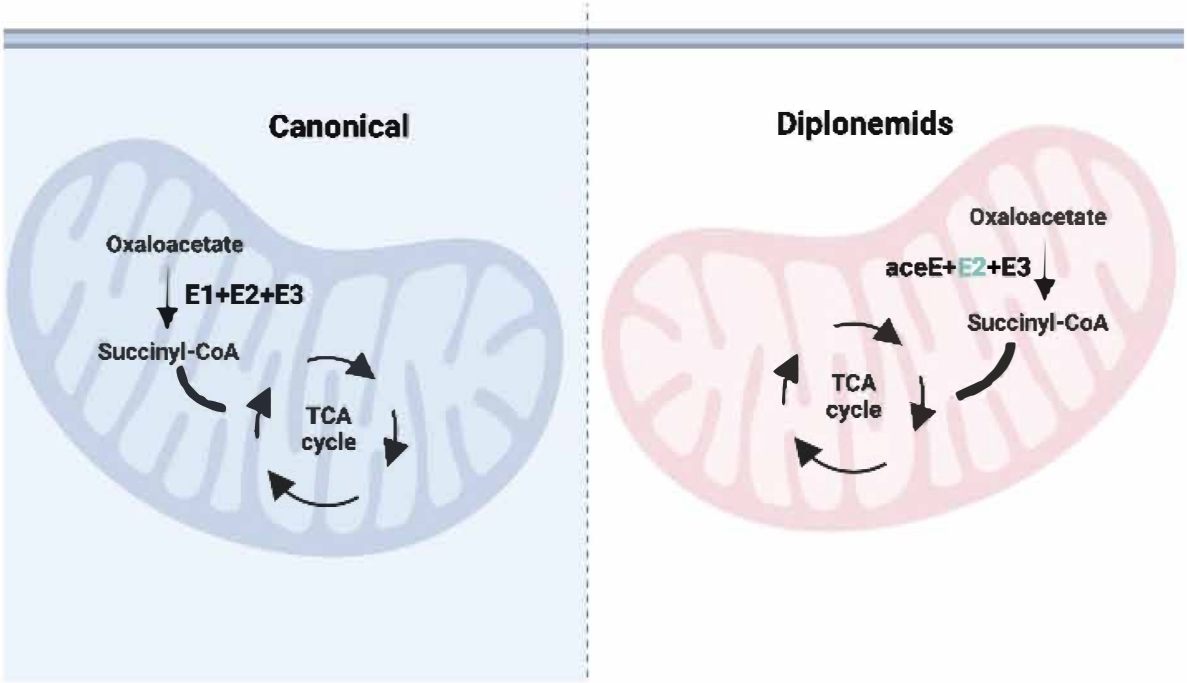


Figure 1. Comparative analysis of canonical and inferred diplonemid pyruvate dehydrogenase complexes. It states that a specific E2 subunit, which is typically found in the pyruvate dehydrogenase complex (PDH), could not be identified in diplonemids. However, it is hypothesized that the function of this missing E2 subunit is fulfilled by E2o and/or E2b from other dehydrogenase complexes in diplonemids. Abbreviations E1- pyruvate dehydrogenase, E2- dihydrolipoamide transacetylase, E3- dihydrolipoamide dehydrogenase, and aceE- dihydrolipoyl acetyltransferase. The scheme is purely illustrative.

In our preliminary genome analysis of the model organism *P. papillatum*, we identified highly conserved homologues of core paraflagellar components found in sister clades Euglenida and Kinetoplastea. The paraflagellar rod (PFR) eventually emerged in several diplonemid species that initially lacked one during the "swimming" stage brought on by starvation, according to a recent survey of those species (Tashyreva, Prokopchuk, Votýpka, *et al.*, 2018b). As a result, a PFR was discovered to be present in every species of Euglenozoa examined, with the

exception of *P. papillatum*, the model organism for diplonemids. Regarding the putative function of the paraflagellar rod (PFR), both Euglenids and Diplonemids demonstrate a distinctive "spinning lasso" pattern of their dorsal flagella, which enables swift cell locomotion (Leander BS, *et al.* 2017; Yubuki and Leander, 2013). A recent review focusing on diplonemids emphasized that the "spinning lasso" flagella pattern is exclusively observed in diplonemids that possess a PFR (Tashyreva *et al.*, 2022b). In contrast, the absence of a PFR in *P. papillatum*, as conclusively demonstrated by this study, results in more constrained flexing of the flagella, leading to slower propulsion. We speculate that the loss of selective pressure for rapid movement, facilitated by the spinning flagellum, in our model organism, has subsequently led to a diminished targeting pressure for the PFR along the flagellum. Our research endeavors have provided evidence of the absence of core paraflagellar rod (PFR) proteins in the flagellar structure of *P. papillatum*. These proteins have instead been retargeted to other cellular regions, establishing *P. papillatum* as the first truly PFR-deficient species within the Euglenozoan group.

In contrast to the absence of the PFR, Diplonemids possess a distinctive cytoskeletal feature called the papilla, characterized by a protrusion of the cell membrane supported by microtubules and microfilaments (Tashyreva *et al.*, 2022b). The papilla's complex structure comprises a basal body, a central filament, and several microtubules that extend into the papilla. Although the precise function of the papilla remains enigmatic, it is hypothesized to play a vital role in cell motility and feeding processes (Leander BS, *et al.*, 2014). Our investigation has conclusively determined that *P. papillatum* is truly deficient in the paraflagellar rod (PFR), distinguishing it from other surveyed Euglenozoans. While PFR proteins were detected in the cytoplasm and associated with microtubule structures near the apical region of the cell, they were notably absent

from the flagellar structure. Our findings suggest that the loss of PFR in *P. papillatum* may have been driven by a lack of selective pressure for rapid movement.

This study highlights the need for further research to elucidate the functions of PFR proteins within the cytoplasm and their potential interactions with other proteins in the papilla of *P. papillatum*. Investigating these aspects will contribute to a deeper understanding of the molecular mechanisms underlying the evolutionary divergence of PFR in this species.

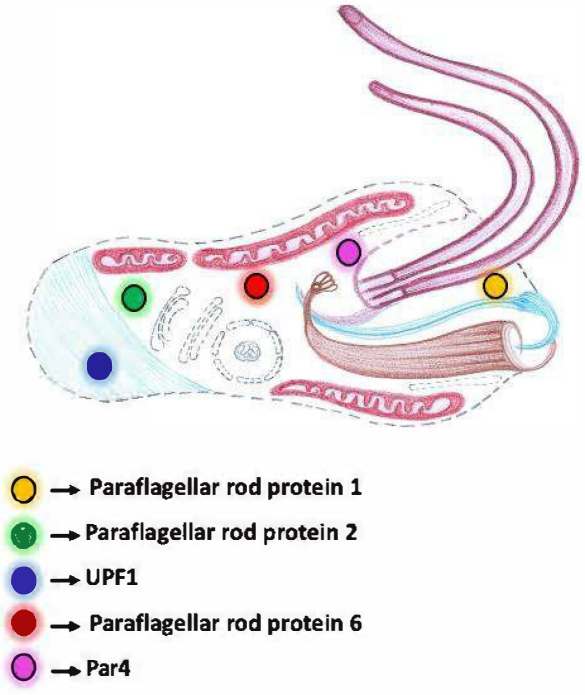


Figure 2. A schematic representation of *P. papillatum* cells depicting the cytoplasmic localization of paraflagellar rod proteins. PFR1 is specifically localized at the papilla, while other paraflagellar rod proteins are distributed throughout the entirety of the cells, as indicated by highlighted circles representing different proteins.

Further, our objective was to investigate the morphology of kinetochores in diplonemids by utilizing kinetochore proteins with identifiable homologs through in silico analysis. Through our experimental approach, we successfully demonstrated the localization of tagged KKT10, KKT19, Mad2, and CENP-A within the nucleus of *P. papillatum* (Figure 3). However, our efforts to identify authentic interacting proteins through mass spectrometry analysis have thus far been unsuccessful. Further investigations are warranted to uncover the protein interactions associated with diplonemid kinetochores. Nonetheless, it is essential to acknowledge the potential limitations of the pull-down assays employed in this study, which may have hindered the detection of certain interactions within the protein complex. Factors such as the size of the tag utilized, the specific buffers employed, and the immunoprecipitation protocols could have influenced the effective incorporation of the tagged protein.

Building upon these observations, further investigations can be conducted to generate an RNAi cell line in *P. papillatum*. This approach will enable exploration of the consequences of knocking out specific kinetochore proteins in *P. papillatum*, providing insights into the cellular response and potential functional implications. Also in the future, using synchronization techniques with drugs like nocodazole and roscovitine, which can arrest a higher proportion of cells in the target cell cycle phase, is an alternative way to examine whether a checkpoint response exists in these cells. Additionally, it was shown that the T. brucei parasite's procyclic (insect) and bloodstream (mammalian) stages both grow more slowly when KKT10 or KKT19 is knocked down (Akiyoshi and Gull, 2013).

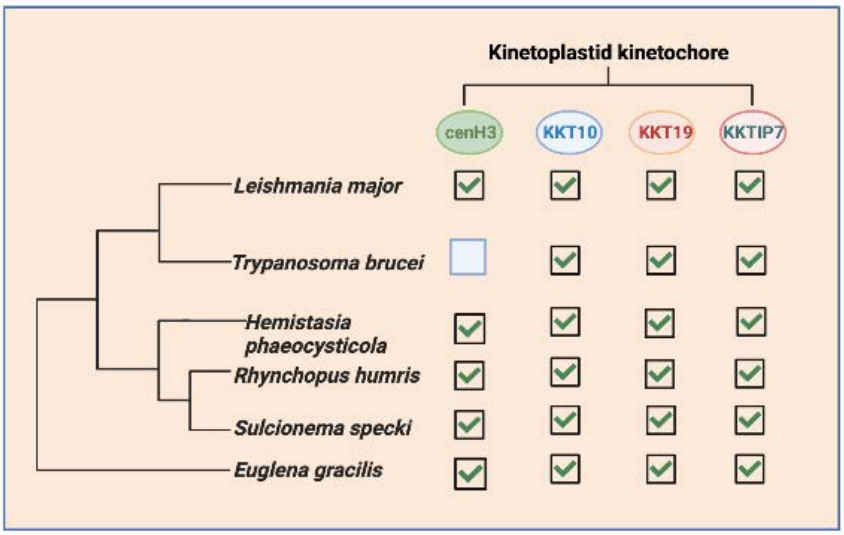


Figure 3. The distribution of the parts of the kinetochore machinery in Euglenozoa. Abbreviations: cenH3, the centromeric variant of histone H3; KKIP, kinetoplastids kinetochore-interacting protein; KKTs, kinetoplastid kinetochore proteins. Information obtained from (from Butenko *et al*., 2020).

Curriculum Vitae

Personal information

Name	Pragya Tripathi
Present Address	Laboratory of Molecular Biology of Protists Institute of Parasitology, Biology Centre CAS Ceske Budejovice, 37005 Czech Republic
Telephone	+420 – 739854315
E – mail	tripathi.pragya540@gmail.com
Nationality	Indian
Gender	Female
Date of Birth	08/07/1989

Academic highlights

- Master of Science in 'Zoology' from VBS Purvanchal University, Jaunpur, India 2011.
- Master of Science in 'RNA Biology, University of Lorraine (**QS world ranking #332**), **Nancy,France**: 2017-2018
- Ph.D. from Institute of Parasitology, Biology Centre, Ceske Budejovice, **Czech Republic**: 2018-Ongoing.
- Qualified all India-level '**Graduate Aptitude Test in Engineering** (GATE), 2011.

Work experience

December 11, 2017- Ongoing	
Designation	PhD student
Main activities	Working on "Investigation of the Subunit Composition of Mitochondrial Dehydrogenase Complexes and Putative Kinetochore and Localization of Paraflagellar Rod Proteins in Marine Diplonemids.
Laboratory	**Prof. RNDr. Julius Lukeš** Laboratory of Molecular Biology of Protists Institute of Parasitology, Biology Centre CAS Ceske Budejovice, 37005 Czech Republic

University Education

Master of Science

Subject	Zoology
University	VBS Purvanchal University, Jaunpur, India

Bachelor of Science

Subject	Zoology, Botany, Chemistry
University	VBS Purvanchal University, Jaunpur, India

Conference Talk

- Participated in International conference on 21st Symposium of the International Society of Endocytobiology, České Budějovice, July 19-22, 2022. An unusual composition of mitochondrial pyruvate dehydrogenase complex of diplonemids, widespread marine protists

- Participated in Ph.D. seminar retreat 2022 14th - 16th October. An unusual composition of mitochondrial pyruvate dehydrogenase complex of diplonemids, widespread marine protists.

- Participated in 15 International Workshop on Opportunistic on 15-17 June 2021 in České Budějovice, Czech Republic. "Tagging of pyruvate dehydrogenase candidates in *Diplonema Papillatum*.

- Participated in 50th Jirovec's Protozoological Days of Czech Societyfor Parasitology, The Protozoological section of the Czech Society for Parasitology, Nove Hrady, Czech Republic 2021. Tagging of pyruvate dehydrogenase E3 subunit in Diplonema papillatum.

Poster

- Best Poster award (rank #3) in international conference ADNAT-2023, held at Banaras, Hindu University from March 10-12, 2023. The genome of *Paradiplonema papillatum* reveals a recent expansion of metabolic versatility.

Publications

1. Kristína Záhonová[1,2,3,4,#], Matus Valach[5,#], **Pragya Tripathi[1,6,#]**, Corinna Benz[1], Fred R. Opperdoes[7], Peter Barath[8,9], Veronika Lukáčová[9], Maksym Danchenko[8], Drahomíra Faktorová[1,6], Anton Horváth[10], Gertraud Burger[5], Julius Lukeš[1,6,*] and Ingrid Škodová-Sveráková[1,2,10,*] **,** Subunit composition of mitochondrial dehydrogenase complexes in widespread marine diplonemids.

1. Matus Valach[1]*, Sandrine Moreira[1], Celine Petitjean[2], Corinna Benz[3], Anzhelika Butenko[3,4,5], Olga Flegontova[3,5], Anna Nenarokova[2,3], Galina Prokopchuk[3,4], Tom Batstone[2,6], Pascal Lapébie[7], Lionnel Lemogo[1,8], Matt Sarrasin[1], Paul Stretenowich[1,9], **Pragya Tripathi[3,4]**, Euki Yazaki[10], Takeshi Nara[11], Bernard Henrissat[7,12], B. Franz Lang[1], Michael W. Gray[13], Tom A. Williams[2], Julius Lukeš[3,4] and Gertraud Burger[1]*, Recent expansion of metabolic versatility in *Diplonema papillatum*, the model species of a highly speciose group of marine eukaryotes.