Evolution of the eukaryote (epi)genome : Insights from orphan protists

Jazmin Blaz Sánchez

▶ **To cite this version:**

Jazmin Blaz Sánchez. Evolution of the eukaryote (epi)genome : Insights from orphan protists. Populations and Evolution [q-bio.PE]. Université Paris-Saclay, 2023. English. NNT : 2023UPASL129 . tel-04403505

HAL Id: tel-04403505
https://theses.hal.science/tel-04403505

Submitted on 18 Jan 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

université
PARIS-SACLAY

Evolution of the eukaryotic (epi)genome: Insights from orphan protists

Évolution de l' (épi)génome des eucaryotes: perspectives offertes par les protistes orphelins.

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° n°577 : Structure et Dynamique des Systèmes Vivants(SDSV)
Spécialité de doctorat: Évolution
Graduate School : Sciences de la vie et santé
Référent : Faculté des Sciences d'Orsay

Thèse préparée dans l' unité de recherche **Ecologie Systématique et Évolution** (Université Paris-Saclay, CNRS, AgroParisTech), sous la direction de **Laura EME**, chargée de recherche, et la co-direction de **David MOREIRA,** directeur de recherche.

Thèse soutenue à Paris-Saclay, le 15 décembre 2023, par

Jazmín Itzel BLAZ-SÁNCHEZ

Composition du Jury

Membres du jury avec voix délibérative

Mireille BÉTERMIER Directeur de recherche, Université Paris-Saclay	Présidente
Anna KARNKOWSKA Maîtresse de conférences, University of Warsaw	Rapporteure et Examinatrice
Arnau SEBÉ-PEDRÓS Directeur de recherche, Institute of Comparative Regulatory Genomics	Rapporteur et Examinateur
Jon JERLSTRÖM-HULTQVIST Maître de conférences, Uppsala University	Examinateur

THÈSE DE DOCTORAT
NNT : 2023UPASL129


cnrs
CENTRE NATIONAL
DE LA RECHERCHE
SCIENTIFIQUE

erc
Macro
EpiK

ESE
Écologie Systématique Évolution

université PARIS-SACLAY

ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Titre : Évolution de l'(épi)génome des eucaryotes: perspectives offertes par les protistes orphelins.

Mots clés : phylogénétique, évolution, génomique comparative, modifications épigénétiques, protistes

Résumé : Depuis leur origine, les eucaryotes ont évolué dans une myriade de lignées. A partir de la combinaison de données ultrastructurales et moléculaires, il a été montré que la diversité des eucaryotes peut être divisée en un petit nombre de super-groupes phylogénétiques. Cependant, la monophylie de certains de ces super-groupes, leurs interrelations et le positionnement de la racine de l'arbre des eucaryotes restent controversés. Cette incertitude peut s'expliquer par le manque de données sur des lignées clés. En outre, la représentation déséquilibrée des cultures et des données génomiques disponibles pour de nombreux clades de protistes a entravé les recherches sur la véritable diversité des caractéristiques clés des eucaryotes ainsi que sur les schémas d'évolution de leur génome et des questions fondamentales sur la diversité des génomes et le rôle de la régulation épigénétique dans l'évolution des eucaryotes microbiens restent encore ouvertes.

Les ancyromonades et les mantamonades sont deux clades de flagellés hétérotrophes avec un mode de vie libre, considérés comme orphelins en raison de leur profonde divergence par rapport à tous les autres super-groupes eucaryotes. Ce projet de thèse avait trois objectifs principaux: i) générer les premières séquences génomiques pour les ancyromonades et les mantamonades; ii) comprendre les principaux processus qui ont conduit à l'évolution des répertoires génétiques de ces organismes; et iii) explorer l'évolution des mécanismes épigénétiques et leur relation avec les modèles d'expression génique en réponse aux changements environnementaux chez les ancyromonades.

Nous avons obtenu un assemblage très complet du génome de *Mantamonas sphyraenae* et rapporté ses principales caractéristiques. Parallèlement, nous avons séquencé et analysé les génomes de plusieurs espèces d'ancyromonades. Malgré les similitudes morphologiques entre les ancyromonades, nous avons constaté que les génomes de ces organismes sont différents en termes de taille, de nombre de gènes et de contenu en séquences répétées.

Nos analyses phylogénomiques ont confirmé la divergence précoce des ancyromonades dans l'arbre des eucaryotes et la monophylie du genre *Mantamonas* dans le supergroupe CRuMs. En outre, la reconstruction du contenu génétique ancestral à partir de 371,634 familles de gènes réparties sur l'ensemble des eucaryotes a révélé un gain important de gènes chez l'ancêtre des ancyromonades, et beaucoup de remplacements des familles de gènes sur l'ensemble de l'arbre des ancyromonades. La plupart des familles de gènes avec une évolution très dynamique ont des fonctions inconnues mais englobent également de nombreuses protéines liées aux mécanismes de transduction des signaux et le cytosquelette.

En utilisant le séquençage du génome entier au bisulfite (Whole Genome Bisulfite Sequencing, WGBS), nous avons exploré les motifs de méthylation de l'ADN des ancyromonades. Nos analyses ont montré de faibles niveaux globaux de méthylation. Notament, les ancyromonades présentent une méthylation du corps des gènes similaire à celle d'espèces eucaryotes éloignées, ce qui suggère qu'il s'agit d'une caractéristique ancestrale de la méthylation de l'ADN chez les eucaryotes.

Enfin, nous avons caractérisé les profils d'expression génique d'*Ancyromonas sigmoides* dans des conditions environnementales changeantes, mettant en lumière leurs réponses moléculaires aux changements de salinité, de température et d'oxygène, ainsi que les rôles possibles de la fraction non caractérisée du contenu génique de ces organismes.

En approfondissant la diversité génomique et épigénomique de ces lignées peu étudiées, ce travail a élargi notre compréhension de l'évolution des eucaryotes et mis en lumière des mécanismes de régulation essentiels régissant l'expression du génome chez les eucaryotes microbiens. Cela ouvre des perspectives prometteuses pour les recherches futures dans ces domaines.

université PARIS-SACLAY

ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Title : Evolution of the eukaryotic (epi)genome: Insights from orphan protists

Keywords : phylogenetics, evolution, comparative genomics, epigenetic modifications, protists

Abstract : Since their origin, eukaryotes have evolved in a myriad of lineages. Based on the combination of ultrastructural and molecular data, it has been proposed that the diversity of eukaryotes can be divided into a small number of phylogenetic supergroups. However, the monophyly of some of the super-groups, their interrelationships and the placement of the root of the eukaryotic tree remain contentious, even in the most recent analyses. This uncertainty can be explained by the lack of data from key lineages. In addition, the unbalanced representation of cultures and available genomic data for numerous protist clades has hampered our investigation of the true diversity of key eukaryotic features as well as the patterns of genome evolution across this domain of life. In particular, fundamental questions about the genome diversity and the role of the epigenetic regulation on the evolution of microbial eukaryotes still remain open.

Ancyromonads and mantamonads are two clades of free-living heterotrophic flagellates considered orphans due to their deep divergence from any eukaryotic supergroup. This project had three main objectives: i) generating the first genomic sequences for ancyromonads and mantamonads; ii) understanding the main processes that have driven the evolution of the genetic repertoires of these organisms; iii) exploring the diversity of epigenetic mechanisms and their relationship to gene expression patterns in response to environmental shifts in ancyromonads. We obtained a highly complete assembly of the genome of *Mantamonas sphyraenae* and reported its main characteristics. In parallel, we sequenced and analyzed the genomes of several ancyromonas species. Despite the morphological similarities among ancyromonads, we found that the genomes of these organisms are diverse in size, number of genes and repeat content.

Our phylogenomic analyses confirmed the early divergence of ancyromonads within the eukaryotic tree of life and the monophyly of the genus *Mantamonas* within the CRuMs supergroup. Furthermore the ancestral gene content reconstruction over 371,634 gene families distributed across eukaryotes revealed a significant gene origination rate in the ancestor of ancyromonads and a high turnover of gene families across the ancyromonad tree. Most of the gene families with dynamic evolution have unknown functions but also encompass many proteins related to signal transduction mechanisms and the cytoskeleton.

Using Whole Genome Bisulfite Sequencing (WGBS) we explored the DNA methylation landscapes of ancyromonads. Our analyses showed low global levels of methylation. Interestingly, ancyromonads display gene body methylation similarly to distant species of eukaryotes, suggesting this is an ancestral feature of DNA methylation in eukaryotes.

Finally, we characterized the gene expression patterns of *Ancyromonas sigmoides* under shifting environmental conditions, shedding light into their molecular responses to salinity, temperature and oxygen changes andon the possible roles of the uncharacterized fraction of the gene content and improve our knowledge of the biology of these enigmatic organisms.

By deepening into the genomic and epigenomic diversity of these understudied lineages, this work has expanded our understanding of eukaryotic evolution and shed light on essential regulatory mechanisms governing genome expression in microbial eukaryotes, opening exciting possibilities for future research in these fields.

Acknowledgments

I would like to thank the members of the jury for accepting to evaluate this work and our enriching discussion during my defense. Thanks also to the members of my thesis committees for their valuable advice. During these years, I have been surrounded by wonderful people without whom this work would not have been possible and I would like to deeply thank the persons who supervised me, collaborated with me and accompanied me during these years. I feel extremely grateful for them.

I want to thank my supervisors Laura Eme and David Moreira for their support, advice and guidance during my PhD as well as the opportunity to work in such a fascinating subject. Thank you for the time you dedicated to share your ideas, teach me and figure out the various challenges we encountered during my project. I can only imagine that mentoring students with such diverse background it's anything but easy, however doing a PhD surrounded by the brilliant and creative scientists forming the DEEM team has been one of the most intense and rich experiences of my life.

Laura, thank you for trusting me to pursue this project. Thank you for your warm welcoming and for all your kind help and patience during the roughest times of the Covid pandemics, my small and big personal crises, and my inhability to survive french administration. I remember my first days in the valley, discovering the Yvette on the bike you lent me and being marveled by all the colors that autumn doesn't have in my hometown. I had a similar feeling when I started working with our tiny cute protists. I feel deeply grateful for the adventure that these years have been.

David, gracias por siempre encontrar tiempo para responder mis preguntas. Gracias por tu ayuda en el laboratorio cuidando (¿y torturando?) los cultivos de ancyros. Gracias por todas las veces que me motivaste, me corregiste pacientemente y por propiciar el trabajo en equipo. **Puri**, aunque no fuiste mi supervisora formalmente, también he aprendido muchísimo de tí y de las discusiones durante los comités de tesis, seminarios y otros momentos. Gracias por hacerme sentir bienvenida en el equipo que has formado con David, siempre admiraré vuestra curiosidad y creatividad.

I would like to specially thank **Eunsoo Kim** for generously sharing the precious genomic data generated at her lab with us and being a crucial collaborator during my project. I hope we will be able to meet you in person someday.

To my talented co-authors **Naoji** Yubuki and **Maria** Ciubanu whose work was absolutely essential for this project. Sharing the days with you at the lab and learning

about protist culturing and molecular biology from your wisdom was a great pleasure Thank you for all the energy and passion you guys put on everything!. Quiero agradecer también a los maravillosos **Luis** Galindo y **Guifré** Torruella**,** por mostrarme en vivo al microscopio los pequeños sujetos de mi futura fascinación y frustración en los primeros días del doctorado y también por enseñarme la diferencia entre frijolitos y elefantes.

I would also thank the people who helped me a lot along the journey. To **Philippe** Deschamps without whom our computing cluster could not work, thank you for all the patience and help with several aspects related to bioinformatics in my project. I would also like to thank **Paola** Bertolino whose work is very important for keeping the wet lab running. To **Jolien** van Hooff**, Julien** Massoni**, Sergio** Muñoz and **Guillaume** Louvel thank you guys for our discussions about science and life, for sharing with me your ideas and for helping me so many times to solve analytic problems.

To the members of the lab who all the time encouraged me **Feriel, Pauline, Marina, Romain, Bledina, Miguel, Ana** and **Jacqui**. Thank you guys for making the lab a happier place even after the shocking move to the Plateau de Saclay. Thanks for being so kind and supportive during the moments in which I was so stressed. Thank you **Kristina** for always being so honest and at the same time so sweet. Thank you for all the amazing monsters you show me in your fascinating videos, for you now I always travel with Falcon tubes when I go on vacations. **Brittany**, thank you for sharing the good and hard parts of this adventure with me, you are an important piece of the beauty of this journey and at the heart of the best memories I will keep from these times.

A **Carolina** y **Rodrigo**, gracias por todo su cariño, siempre que estoy con ustedes me siento como si estuviera en casa. Caro, el trabajo que haces es admirable y me inspira a seguir cada día y seguir soñando.

I also dedicate a very special thought to **Thomas** and **Fabian,** to whom I am very grateful for their warm welcoming into the team when I arrived in France. Thank you for all the help science related and also for the non science related experiences we have shared. I miss your special sense of humor guys!

Rémy, merci de m'avoir appris à croire en moi même et à faire confiance à mes pieds pendant les séances d'escalade. Tu imagines pas à quel point ça a été important pour moi. Merci aussi à **Arthur, Thibault, Baptiste, Fabien** et **Max**, pour votre générosité et votre amitié lorsque vous m'avez reçu chez vous en tant que coloc non officiel à l'époque des confinements de Covid.

A mis amigos en México, **Anahí, Diana, Shama, Cristóbal, Viv, Itzel, Dulce, Lupita**, **Isabel** y **Rosalba** gracias por hacer de este mundo un lugar más lindo. Gracias por su cariño, ustedes siempre están en mi corazón.

A mi querida **Valerie** de Anda, gracias por tu generosidad y entusiasmo contagiosos. No puedo agradecerte lo suficiente por leer las primeras versiones de mi manuscrito, por tu ayuda en python, por todas las ideas que me diste para darle sentido a mis resultados y el apoyo emocional que me has dado estos años. Eres una persona brillante y te admiro muchísimo. Confío en que nuestra amistad dure toda la vida y sé que cuando seamos viejitas nos acordaremos de todo esto con humor.

Enrique Ibarra-Laclette, a pesar de la distancia siempre has sido un maravilloso mentor y amigo. Gracias por haberme recibido en tu lab en uno de los momentos más difíciles de mi vida, y por seguir apoyándome en la distancia. Eres una de las personas mas brillantes y generosas que conozco y me siento increíblemente afortunada de haber trabajado y aprendido tantas cosas contigo.

A **Galel** y **Maud**, gracias por compartir conmigo la belleza de Saorge y tantos momentos importantes. Gracias por llenar mi corazón de Son Jarocho, musica barroca y su buen humor.

A mis padres **Violeta** y **Emmanuel**, gracias por siempre haber creído en mí, por haberme animado a estudiar ciencia desde pequeña, a seguir mis instintos y mi curiosidad. Gracias por abrigarme con su amor inclusive en la distancia trasatlántica que ahora nos separa. Se que estos fueron años difíciles y espero poder compensarles todo lo que me han dado. A mi hermananita **Rubí,** te extraño muchísimo, gracias por compartir tus historias. La idea de seguir leyéndote me hace una persona feliz! Espero que pronto podamos vivir nuevas aventuras para contar. A mi hermana **Nayeli**, yo no sería quien soy sin tí. Te fuiste demasiado pronto, pero estoy convencida de que una parte de tí vive en mí, en las cosas que hago y en los lugares a los que voy.

A mi querida ***Erianacea***, desde que naciste has ocupado el lugar más importante de mi corazón, y has hecho siempre las preguntas más interesantes. Persigue siempre tus sueños, yo te apoyaré sin importar ninguna distancia. Gracias por tu amor.

Et finalement, merci **Florian**. Merci de m'avoir soutenue pendant toutes ces années. Je n'aurais jamais terminé ce travail sans ton amour, ton esprit incroyable et ton bon humour qui font de chaque moment, même pendant l'écriture d'une thèse, une expérience *magique et musicale*. Franchement ce travail est aussi le tien, je pense que tu en sais déjà plus sur les ancyromonads que n'importe quel citoyen moyen haha. Cela me rend joyeuse de rêver à tout ce qu'il nous reste à partager. Je t'aime !

Contents

“Every scrap of biological diversity is priceless, to be learned and cherished,

and never to be surrendered without a struggle.”

E. O. Wilson

1. Introduction

1.1 Studying the deep evolution of life

One of the major evolutionary transitions of the history of life has been the emergence and further diversification of eukaryotes, which occurred more than two billion years ago (Eme et al. 2014). Eukaryotic cells are characterized by the presence of a nucleus, energy generated by mitochondria, a complex cytoskeleton, and an intricated endomembrane system. These features, along with a fundamentally different genome organization and regulation compared to their prokaryotic ancestors, have played a crucial role in enabling the compartmentalization of biochemical processes in eukaryotic cells. As well, this has also catalyzed the evolution of complex cell morphologies, life cycles, and multicellularity in various lineages (Figure 1).

Understanding the processes that led to the origin and further diversification of this complex form of life remains one of the most fascinating and significant challenges in biology. During my PhD, I approached this deep evolutionary history through the examination of protist species belonging to ancient and divergent lineages as a study case of the evolution of the eukaryotic (epi)genome. In this introductory chapter, I will describe some of the milestones that have revolutionized our understanding of the evolution of eukaryotes and the ideas that have sparked the questions that I addressed in this project.

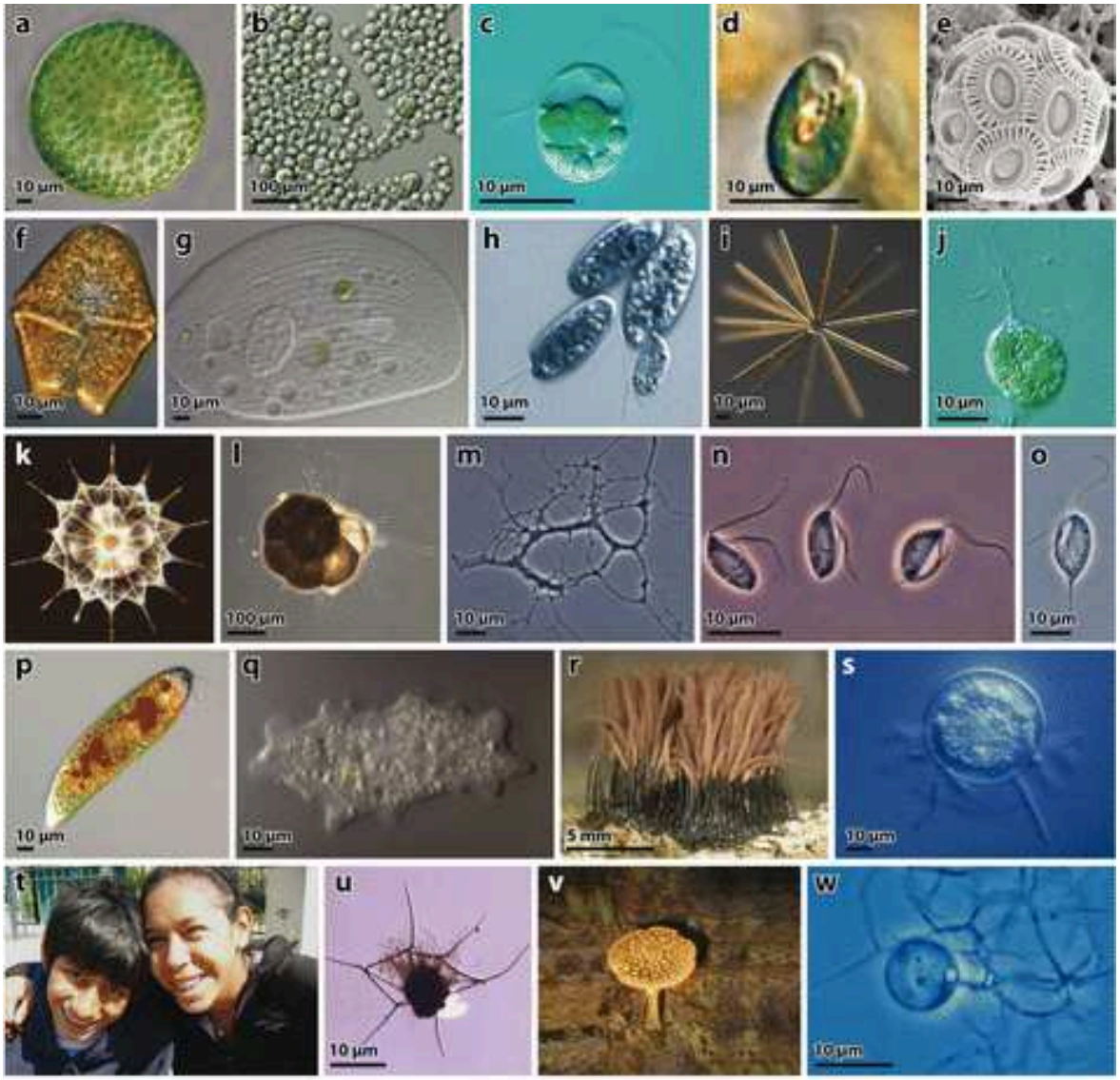

Figure 1. Eukaryotes can take a miriard of forms. (a) *Eremosphaera viridis*, a green alga. (b) *Cyanidium sp.*, a red alga. (c) *Cyanophora sp.*, a glaucophyte. (d) *Chroomonas sp.*, a cryptomonad. (e) *Emiliania huxleyi*, a haptophyte. (f) *Akashiwo sanguinea*, a dinoflagellate. (g) *Trithigmostoma cucullulus*, a ciliate. (h) *Colpodella perforans*, an apicomplexan. (i) *Thalassionema sp.*, a colonial diatom. (j) *Chlorarachnion reptans*, a core cercozoan. (k) *Acantharea sp*., formerly known as a radiolarian. (l) *Ammonia beccarii*, a calcareous foraminiferan. (m) *Corallomyxa tenera*, a reticulate rhizarian amoeba. (n) *Jakoba sp.*, a jakobid with two flagella. (o) *Chilomastix cuspidata*, a flagellate in Fornicata. (p) *Euglena sanguinea*, an autotrophic Euglenozoa. (q) *Trichosphaerium sp.*, a naked stage (lacking surface spicules) of an unusual amoeba with alternation of generations, one naked and one with spicules. (r) *Stemonitis axifera*, a dictyostelid. (s) *Arcella hemisphaerica*, a testate amoeba in Tubulinea. (t) *Homo sapiens*, animal. (u) *Campyloacantha sp.*, a choanoflagellate. (v) *Amanita flavoconia*, a basidiomycete fungus. (w) *Chytriomyces sp.*, a chytrid. Figure and captions adapted from Katz (2012).

1.1.1 Origin of the eukaryotic domain

All eukaryotes share a common ancestor. It is now recognized that this ancestor descends from the endosymbiosis of formerly free living organisms through one of the most important evolutionary transitions of life, the eukaryogenesis.

By the beginning of the twentieth century, Konstantin Mereschkowski proposed that chloroplasts in photosynthetic eukaryotes could be symbiotic cyanobacteria (Mereschkowsky 1905, 1910). Similar ideas were later elaborated by Lynn Margulis in her famous serial endosymbiosis hypothesis to explain the origin of the eukaryotic organelles putting forward the idea that symbiosis is a powerful creative force in the evolution of life (Margulis 1971; Margulis and Bermudes 1985). In parallel, Emile Zuckerkandl and Linus Pauling pioneered the use of amino acid sequences to reconstruct species relationships (Pauling et al. 1963; Zuckerkandl and Pauling 1965). This inaugural era of molecular biology and evolution allowed the establishment of the first universal molecular phylogeny by Carl Woese that led to the discovery of Archaea as another domain of life (Woese and Fox 1977). This has drawn the *renaissance* of microbial phylogenetics (Woese 1994) and opened a gateway for new hypotheses in which an archaeon was the symbiotic partner of the bacterial mitochondrial ancestor (López-García, Eme, and Moreira 2017; Spang 2023).

The further development of microbial ecology based on the isolation and study of environmental DNA has illuminated the vast extent of microbial diversity on our planet and allowed the examination of the previously hidden diversity of organisms without established cultures (Hug et al. 2016). Asgard archaea are one of such lineages and were cultured only recently. The discovery of Asgard archaea as a closer phylogenetic sister branch of eukaryotes compared to other archaea represented another breakthrough into the understanding of eukaryogenesis. Asgards encode eukaryotic signature proteins (ESPs) in their genomes, which additionally provides compelling evidence that the eukaryotic branch is placed within archaea, as a sister to the Heimdall−Hodarchaeota clade (Eme et al. 2018, 2023). Experimental studies of asgard archaeal profilins have confirmed the functional conservation on the activity of these proteins in assembling actin filaments (a cytoskeleton component) (Akıl and Robinson 2018). Additionally, the cultivation and microscopic observation of

Lokiarchaeota has revealed cellular protrusions putatively supported by filaments comprising lokiactins (Imachi et al. 2020). Altogether, these evidences suggest that the asgard lineage already possessed some of the complexity exhibited by modern eukaryotes, such as a cytoskeleton-like system (Spang 2023).

In contrast, the bacterial ancestor of mitochondria remains elusive, and the most recent phylogenomic analyses show that modern mitochondria clade is related to but outside known alphaproteobacteria (Muñoz-Gómez et al. 2022).

Under the light of these findings, eukaryotes constitute a third but secondary and merger domain of life (Figure 2, left panel). Other questions about the tempo and mode of eukaryogenesis remain open. When and how mitochondria were acquired is one of the most important ones (Roger, Susko, and Leger 2021), as well, the origin of the eukaryotic membrane lipid chemistry remains enigmatic (López-García and Moreira 2015; Eme et al. 2018). Moreover, several alternative scenarios of how eukaryogenesis happened contrasting in cellular mechanistic aspects, the number and nature of possible ancestral symbionts and the order of acquisition of key eukaryotic features have been proposed over the last years (Recently reviewed and discussed in Donoghue et al. 2023).

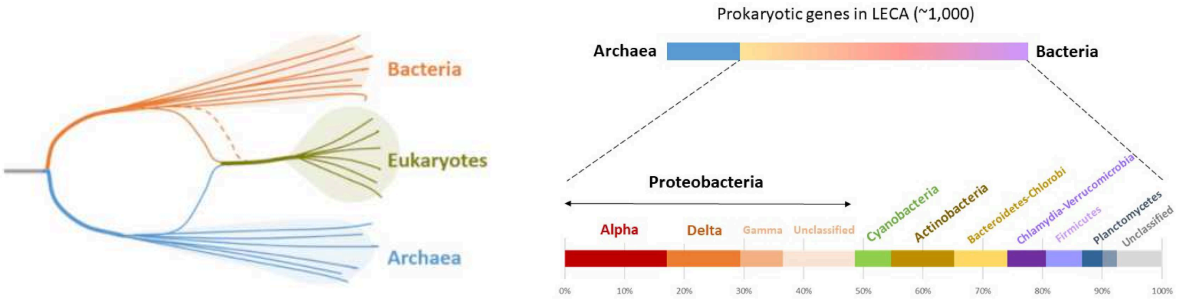

Figure 2. Left: Current model of the domains of life. Right: Ancestral eukaryotic genes with different prokaryotic origins based on previous reconstructions of the LECA gene content. Figure and captions adapted from López-García and Moreira (2023).

As we can learn from modern examples of endosymbiosis, lateral gene transfers between the symbionts are frequent, and because of the merger nature of eukaryotes, it is expected that the ancestral eukaryotic nuclear genome contained genes with

multiple origins (Doolittle 1998) (Figure 2, right panel). Different eukaryogenesis hypotheses make distinct predictions about the phylogenetic origins of the proteins present at the last eukaryotic common ancestor (LECA). This origin has left an imprint in the genomes of modern species and could be traced back to the bacterial and archaeal ancestors. For example, several components of the eukaryotic informational machinery (e.g. transcription and translation) are closer to the archaeal rather than the bacterial one, while the genes involved in metabolism are usually bacterial-like (Spang 2023). Additional prokaryotic genes could have been acquired at different evolutionary distant time points by the eukaryotic ancestors, between the first eukaryotic common ancestor (FECA) and LECA (Eme et al. 2018). As well, lateral gene transfer during the further diversification of eukaryotes also explains the chimerism of the modern eukaryotic genomes, however, phylogenetic methods should be able to detect if these are recent or ancient transfers.

Indeed, to gain insight into the tempo of gene acquisition during eukaryogenesis, some authors have used the branch-length of the phylogeny of anciently originated gene families (Pittis and Gabaldón 2016), and gene duplications (Vosseberg et al. 2021) to tackle the age of key eukaryotic features. These works have suggested that the duplication of several genes-families of the cytoskeleton and membrane-trafficking system predate the acquisition of mitochondria and that gene families involved in the transduction of signals and the regulation of transcription expanded after this endosymbiosis was fixed.

Through the comparison of modern eukaryotes from diverse and early divergent lineages, several authors have reconstructed the gene content of LECA and used phylogenomics to disentangle the origin and evolution of essential machinery such as the membrane trafficking system (More et al. 2020), the kinetochore (van Hooff et al. 2017), the epigenetic toolkit (Weiner et al. 2020a), and the chromatin (Grau-Bové et al. 2022).

The inclusion of diverse and disparate species in these studies has dramatically improved our understanding of the ancient evolution of eukaryotes. As we have seen with the Asgards, the growing body of evidence from key lineages challenge or strengthen our hypotheses and refine our view of this ancient evolutionary history as we improve our sampling of eukaryotes, archaea, and bacteria.

1.1.2 The eukaryotic tree of life

> "Evolution is the essence of systematics. That point has dual meaning: not only are relationships consequent on evolutionary background the major part of what systematics seeks to represent in classification; classifications themselves are subject to a kind of cultural evolution as understanding of biological evolution changes. Changing classifications in turn suggest still further inquiry into, and change of views of biological evolution."
>
> Whittaker & Margulis (1978)

The study of eukaryotic microorganisms, or protists, has a long history that goes back to the earliest microscopes in the seventeenth century. The capacity to identify these organisms due to their high morphological diversity improved with the development of microscopy, allowing the characterization and cataloging of a myriad of lineages (Adl et al. 2007; Caron et al. 2009). These descriptive studies were common up to the end of the twentieth century and have been crucial for the understanding of the roles of protists concerning ecosystem functioning and human health (Caron et al. 2009).

The term *protista* appeared with the first classifications of life that included microorganisms (Figure 3). Today, the term protist is still being used to refer to all the forms of eukaryotic life that are not plants, fungi, or animals (Rothschild 1989). The naming, grouping, and splitting of protist taxons have been tightly tied to the development of our conceptions of evolution and therefore has dramatically changed since the first species were formally described (Whittaker and Margulis 1978), and especially in the recent decades (Adl et al. 2012; Burki et al. 2020). Now, it is recognized that most of the eukaryotic diversity is indeed represented by protists and that they provide the foundation for understanding the origins and diversification of all eukaryotes (Sibbald and Archibald 2017; Blaxter et al. 2022).

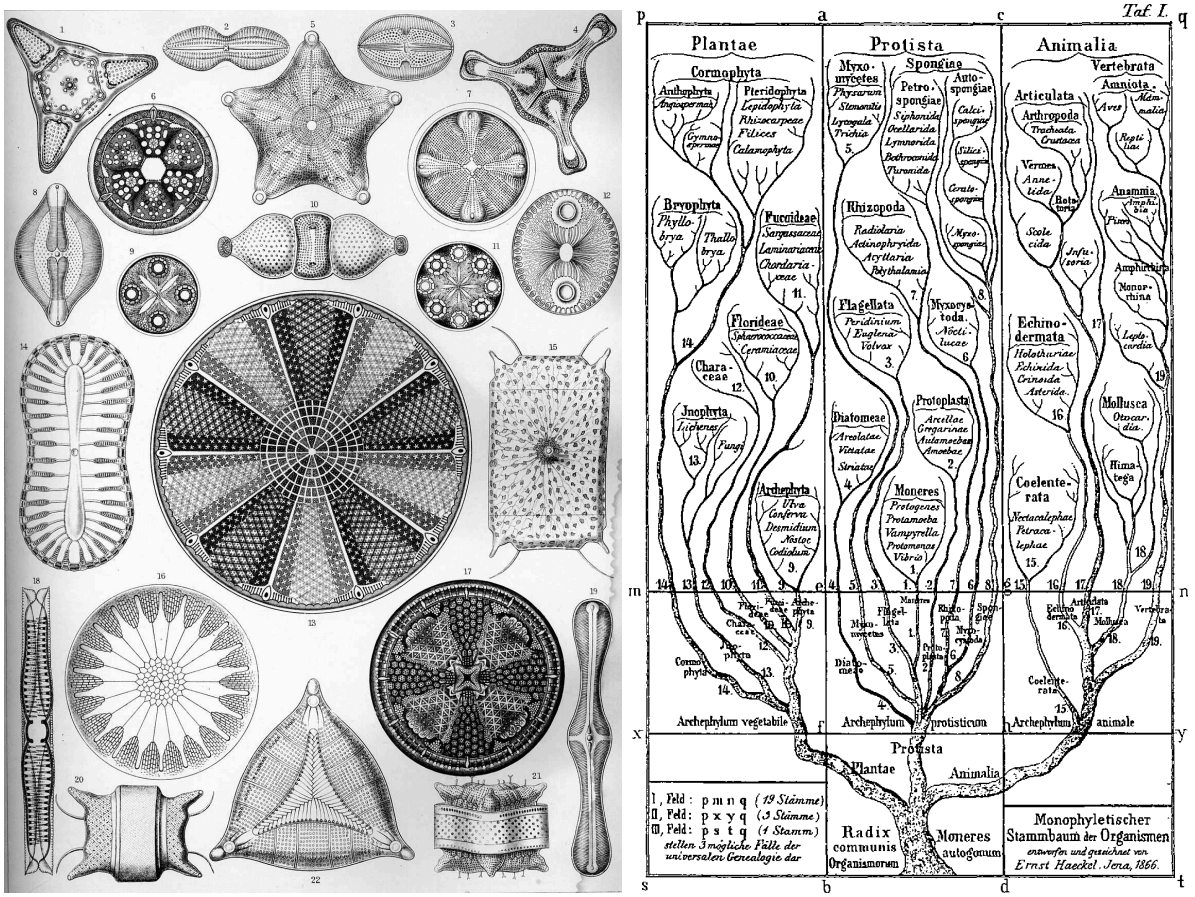

Figure 3. Left: Ernst Haeckel's diatom plate (Haeckel 2004). Right: Three kingdom tree of life depicted in by Haeckel (1866).

With the improvement of phylogenetic methods and models, as well as the generation of transcriptomic and genomic data from diverse eukaryotic species over the last few decades, it has become possible to resolve the relationships of ancient eukaryotic lineages based on the concatenation of several genes (Delsuc, Brinkmann, and Philippe 2005). The integration of these studies has led to the current model of the eukaryotic tree of life (eToL) (Burki 2014; Burki et al. 2020), in which most of the species diversity is comprised within major eukaryotic clades coined as supergroups (Figure 4).

Amorphea comprises Obazoa (including all animals, fungi, their relatives, as well as apusomonads and breviates) and Amoebozoa (diverse amoebae, slime molds, etc.). Altogether Amorphea is a well-supported clade and is represented by many well-known model species. In contrast, CruMs, very recently ranked as a supergroup, comprises only few representative species with molecular data, displaying very diverse morphologies, that were previously considered orphan lineages (Brown et al. 2018).

Archaeplastida includes green algae and land plants (further termed Chloroplastida) as well as red algae (Rhodophyta) and glaucophytes. All these are photosynthetic organisms and harbor plastids originating from the primary endosymbiosis of cyanobacteria (Ponce-Toledo et al. 2017).

Cryptista include organisms relevant for the study of the acquisition of photosynthesis through the secondary endosymbiosis of eukaryotes bearing primary plastids (secondary red plastids). Haptista includes algae such as *Emiliania huxleyi* and centrohelids.

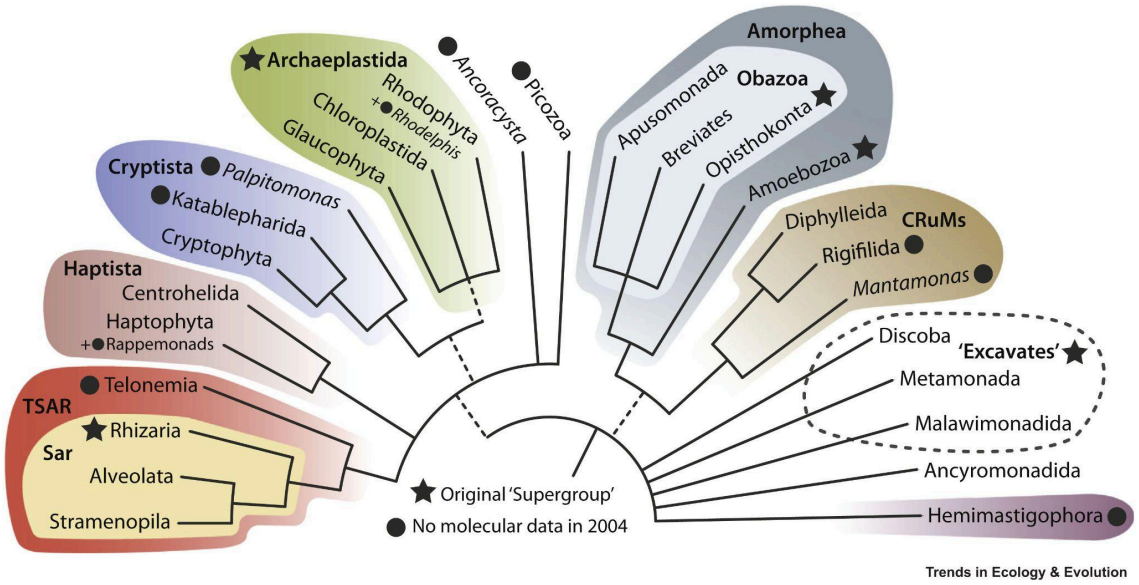

Figure 4. Schematic representation of the current consensus eukaryotic tree of life (eToL). The colored groupings correspond to the current 'supergroups'. Unresolved branching orders among lineages are shown as multifurcations. Broken lines reflect lesser uncertainties about the monophyly of groups. Star symbols denote taxa that were considered as supergroups in early versions of the eToL and circles show major lineages that had no molecular data when the supergroup model emerged. Rappemonads are placed on the basis of plastid rRNA data only. Figure and caption adapted from Burki et al. (2020).

The SAR clade (acronym Stramenopila, Alveolata, and Rhizaria) includes organisms with extremely diverse morphologies and life history traits such as ciliates, diatoms, radiolarians, foraminifera, dinoflagellates and the multicellular brown algae

(Grattepanche et al. 2018). As a sister group of SAR, is Telonemia, which only has two molecularly characterized representative species (Shalchian-Tabrizi et al. 2006).

Some more inclusive taxa (ranking higher than supergroups) have been also proposed over the years. For example, alongside TSAR, Cryptista and Archaeplastida form a group referred to as Diaphoretikes (Burki et al. 2020). In contrast, previous groups characterized on the basis of morphological and life story characters have been split due to their lack of support in phylogenomic analyses, such as Excavata or Chromalveolata (Burki et al. 2020). This radical remodeling has a profound effect on the way we understand the evolution of key features across eukaryotes as a whole. For example, Excavata was a previously defined supergroup based on the conservation of a conspicuous morphological characteristic: the feeding groove (Simpson 2003). However, the monophyly of Excavata has been questioned, therefore, depending on the position of the root of the eToL this implies that the origin of such morphology is ancient and could have been exhibited by the last eukaryotic common ancestor (LECA) (Burki et al. 2020; Derelle et al. 2015).

1.1.3 The elusive root of the eToL and the enigmatic orphan branches.

The root of the eukaryotic tree of life is the hypothetical common ancestor of all eukaryotes and, therefore, a fundamental question in biology. Based on the use of bacterial and archaeal outgroups, the root of the eToL has been previously proposed to be between major groups comprising Diphoda and Opimoda (Derelle et al. 2015), while more recent analyses have suggested a placement of the root between Opisthokonta and everything else (Cerón-Romero et al. 2022) as well as within the paraphyletic group of excavates (Al Jewari and Baldauf 2023).

In addition to this lack of consensus on the root of the eToL, fossil records and molecular clocks suggest that the early diversification of eukaryotes since LECA, was characterized by an important radiation during the proterozoic (Knoll 2014; Betts et al. 2018). This fast process of cladogenesis, that some authors have compared with the *Big*

Bang (Koonin 2007), probably contributes to the difficulty of disentangling the deep structure of the tree.

The current model of the eToL highlights the presence of novel and previously unrecognized lineages of eukaryotes that branch deeply within the tree and lack affinity to any of the supergroups coined as orphans (Burki et al. 2020; Simpson and Roger 2004). In addition, Discoba and Metamonada, represented by many characterized species, have been historically hard to place within global eukaryotic phylogenies, probably because these organisms generally have reduced and fast-evolving genomes. Additionally, the culturing and phylogenetic characterization of some of the other lineages has further clarified their affiliation to other species which yet are still represented by only a few species.

For example, *Collodyction, Rigifila* and the gliding flagellate *Mantamonas* are free-living protists with very different basic morphologies (swimming flagellates, filose amoeboid cells, and tiny gliding cells, respectively) and longly considered each an orphan lineage (Yabuki, Ishida, and Cavalier-Smith 2013; Glücksman et al. 2011; Zhao et al. 2012; Cavalier-Smith et al. 2014). However, four of these species were recently shown to robustly cluster in a new supergroup named CRuMs (Brown et al. 2018).

Moreover, *Ancoracysta twista*, bearing one of the biggest mitochondrial genomes across eukaryotes (Janouškovec et al. 2017) is now placed in a group alongside other predatory flagellates in the recently inaugurated Provora supergroup (Tikhonenkov et al. 2022). Similarly, Hemimastigophora (Figure 5) position has also been recently reevaluated and revealed that they are probably related to *Meteora* (Eglit et al. 2023) a charismatic lineage for which molecular data was available only until recently (Galindo, López-García, and Moreira 2022).

In contrast, the position of ancyromonads and malawimonads remains elusive. Malawimonads exhibit the characteristic feeding groove of excavates and only count with two formally described species (Heiss et al. 2018). Ancyromonads, on their end, are bean shaped flagellates that glide in aquatic sediments or soil and that feed from prokaryotes that graze from diverse environments that range from soils to aquatic sediments (Saville-Kent 1882; Heiss, Walker, and Simpson 2011).

Despite their probably relevant ecological role and evolutionary importance, little is known about the biology of most of all these orphan protists. If fast evolution and long branch attraction is discarded (Philippe et al. 2000) these lineages represent early divergent eukaryotes and therefore hold profound implications into the inference of the root and the early radiation of the eToL.

As described in the previous section, the understanding of eukaryotic diversity and relationships continues evolving with additional data and taxon sampling. One of the greatest obstacles to resolving the position of most of these orphan lineages is the limited sampling of genes and taxa associated with them. Although ultrastructural studies have identified some of these species since several decades ago, only a few of them have been subject of genomic analyses and they are generally represented by a handful of species. As the sparse sampling increases the risk of phylogenetic artifacts that can result in false relationships, statements of monophyly may be premature when taxonomic sampling is insufficient.

The recent clarification of the position of *Ancoracysta* within Provora demonstrates that the isolation, and detailed examination of enigmatic heterotrophic flagellates species plays a crucial role in addressing significant evolutionary questions. Therefore a further genomic exploration of lineages such as ancyromonads and the poorly sampled CRuMs is critical to clarify the placement of the root of the eukaryotic tree of life and will likely impact the understanding of the deep evolution of key eukaryotic features.

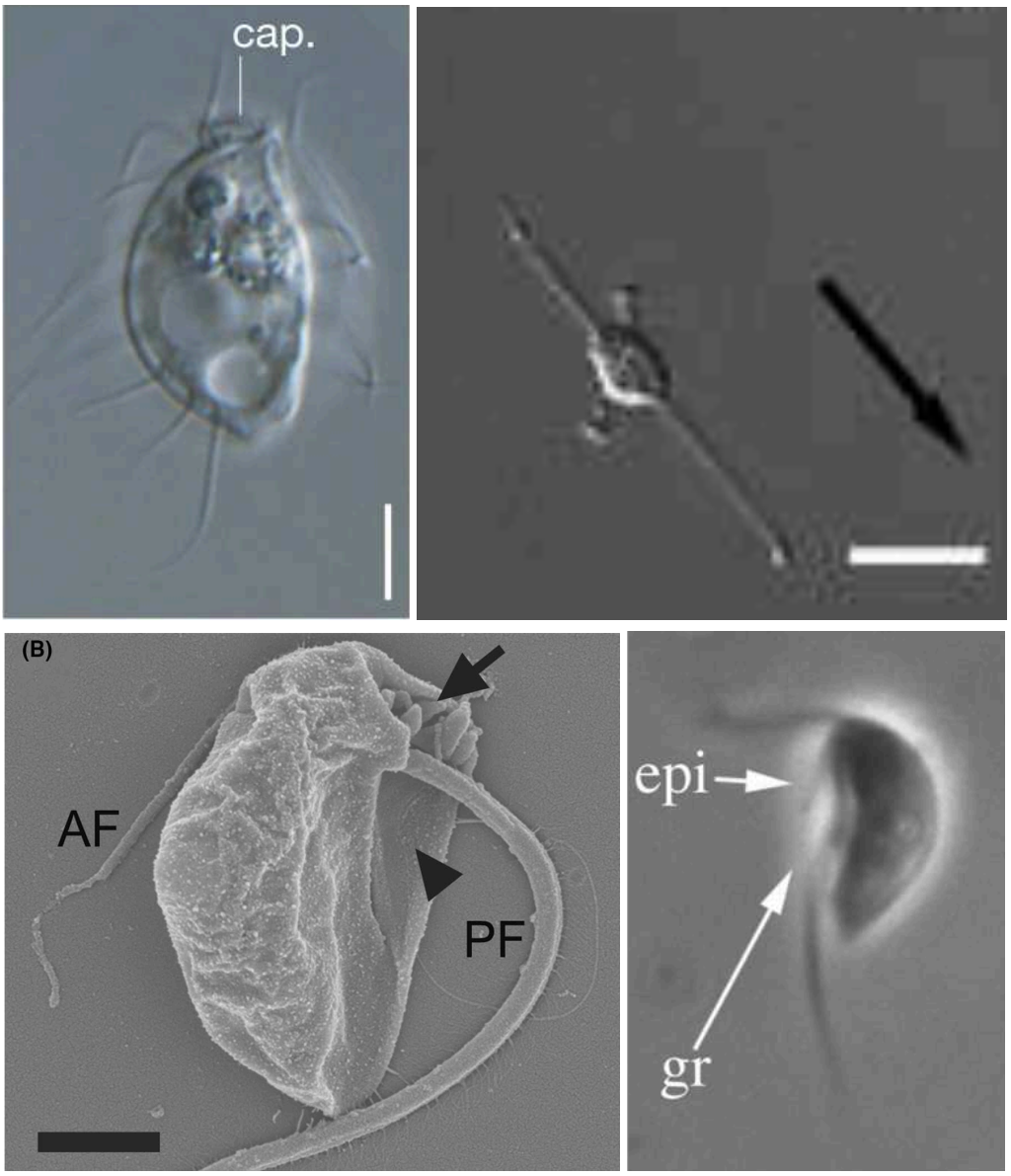

Figure 5. Micrographs of representative deep branching protists. Up and left: *Hemimastix kukwesjijk*, cell with capitulum (cap) (Lax et al. 2018). Up and right: Differential interference contrast (DIC) and phase contrast microscopy observations of *Meteora sporadica* cells (Galindo, López-García, and Moreira 2022). Down left: Scanning electron micrographs (SEM) of *Nyramonas silfraensis gen. et sp. nov.* an ancyromonads species (Yubuki et al. 2023). Down right: Phase contrast microscopy observation of *Gefionella okellyi* (Heiss et al. 2018).

1.2 Paradigms of eukaryote genome evolution

1.2.1. Organization and diversity of eukaryotic genomes

In eukaryotic nuclear genomes DNA is wrapped around octamers of histone proteins, forming nucleosomes, the fundamental unit of the chromatin (Talbert and Henikoff 2010). Chromatin exhibits a further hierarchical three-dimensional organization and is ultimately packed in linear chromosomes. While fascinating exceptions exist, most eukaryotes conserve the fundamental characteristics of this genome organization, in contrast, the genome size, architecture and content varies widely across different lineages (Misteli 2020). Indeed, the realization that the genome size and number of genes does not correlate with the number of cell types (a proxy of organismal complexity) has perplexed several researchers and was known for some time as the C-value enigma (Figure 6). Free living microeukaryotes with comparable cellular complexity but disparate genome architecture can be exemplified by the streamlined genome of green alga *Ostreococcus tauri* of 12.5 Mbp (Derelle et al. 2006) and the repetitive genome of the foraminifera *Reticulomixa filosa,* spanning around 320 Mbp (Glöckner et al. 2014).

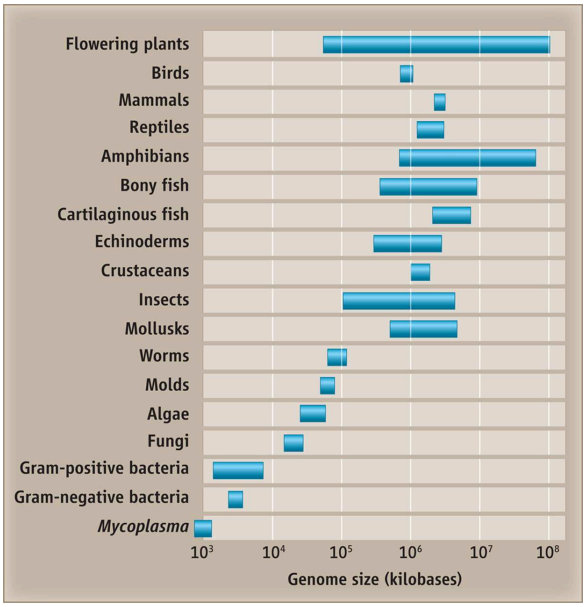

Figure 6. The C-value enigma. The range of haploid genome sizes is shown in kilobases for the groups of organisms listed on the left. Figure and caption from Fedoroff (2012).

One of the main hallmarks of the eukaryotic genome architecture is the high abundance of non coding sequences of diverse nature, explaining in part this apparent paradox. Early in 1948, even before the model of double helix structure of the DNA was published, Barbara McClintock discovered mobile loci in *Zea maize* capable of changing their localization in the genome and being responsible for the generation of different pigmentation patterns in this species (McClintock 1950; Feschotte 2023). Although this finding was initially received with skepticism, the further discovery of “jumping genes”, also known a transposable elements in diverse microbial and multicellular species substantiated the fact that they are pervasive across the tree of life (Fedoroff 2012; Feschotte 2023).

Transposable elements (TEs) have been classified based on the different molecular mechanisms that they use to move in the genome and the nature of their intermediary molecules into two main types: DNA transposons, which move through a "cut-and-paste" mechanism involving excision and reinsertion at new locations, and retrotransposons, which utilize an RNA intermediate for their mobility, involving transcription, reverse transcription, and insertion (Feschotte and Pritham 2007). TEs also can inactivate when losing the proteins necessary to self-replicate and transpose, and many of the repetitive sequences found in nuclear genomes constitute fossils of ancient events of burst and decay of TEs (Wallau et al. 2014; Bourque et al. 2018). Some gene families of transposable elements can be defined based on their common origin when this is tractable, however these sequences evolve quickly, and also have the capacity to move adjacent genetic material and even to hijack other mobile elements, making their study at large evolutionary scales a fascinating but challenging quest (Bourque et al. 2018). Mechanistic similarities as well as conserved sets of genes such as integrases and capsids have helped to trace a tight relationship between some TE families and both RNA and DNA Viruses (Fischer and Suttle 2011; Skala 2014).

Because of their prevalence and dynamic nature, TEs and EVEs are now recognized as powerful source of genomic variation, raising questions on their short and long term impact on the genome stability (Finnegan 1989; Oggenfuss and Croll 2023) and the diversification of regulatory networks (Miller et al. 2000; Feschotte 2008). Recent evidence from diverse organisms suggest that transposable elements can play

adaptive roles in the host genomes by co-opting machineries involved in environmental response (Maumus et al. 2009; Hunter et al. 2015).

Moreover, endogenous viral elements (EVEs) have been recently shown to be abundant and highly diverse in eukaryotic genomes (Bellas et al. 2023). Furthermore, viral interactions are recognized to be a driver of gene acquisition and gene exchange across eukaryotes (Koonin 2010; Frank and Feschotte 2017; Irwin et al. 2022; Barreat and Katzourakis 2022). Finally, some studies have shown that EVEs can play an immunity role in multipartite viral coinfections in *Cafeteria* flagellates (Fischer and Hackl 2016; Roitman et al. 2023), whether this is a common mechanisms in other eukaryotic lineages is an interesting question.

Another kind of non-coding element contributing to the larger genome sizes in eukaryotes compared with prokaryotes are the spliceosomal introns. Introns are non-coding regions interspersed within eukaryotic genes, they can contain repeats, therefore significantly increasing the size of a gene (Gilbert 1978; Rogozin et al. 2012). Introns are further removed from transcripts by the splicing machinery (Nilsen 2003), which in turn can generate different isoforms from the same stairting protein coding gene being an additional source of protein diversity (Mayr 2016). The evolutionary analysis of the ancestral intron position of ancient gene families have traced the origin of spliceosomal introns before the emergence of LECA and a pervasive loss during eukaryogenesis (Vosseberg et al. 2022).

Finally, the exploration of eukaryotic genomes beyond traditional models of animal plants and fungi has further revealed a rich diversity of features as well as genomic oddities in some lineages. A notable example of extreme genome organization is represented by the dinoflagellate *dinokaryon* which unlike the nuclei of most eukaryotic cells, harbors permanently condensed chromosomes bound to Dinoflagellate/Viral NucleoProteins (DVNPs) instead of canonical histones (Fukuda and Suzaki 2015; Gornik et al. 2012).

Striking variation of protein coding gene architecture can be also found ciliates which have evolved gene-sized chromosomes, several alternative genetic codes and a separation of somatic and germline in different nuclei (Smith and Keeling 2016; Boscaro and Keeling 2023).

Why do eukaryotes exhibit a greater variation in the genome architecture than prokaryotes is an important question that is probably related to the complexity of their regulatory mechanisms (discussed in the following introductory sections) as well as the effect of random genetic drift because of the typically smaller populations that some eukaryotes have (Szitenberg et al. 2016).

1.2.2 Processes driving the evolution of gene content across eukaryotes

Over macroevolutionary scales, genomes evolve losing and acquiring new functions through mechanisms of protein coding gene gain and and loss (O'Malley, Wideman, and Ruiz-Trillo 2016). New genes can evolve from non coding sequences through random mutations (Tautz and Domazet-Lošo 2011). Other important mechanisms of genetic innovation include events of domain fusion by which some proteins can acquire new activities (Marsh and Teichmann 2010).

Moreover, when a DNA sequence is duplicated, the copies of a gene can either conserve the function, or evolve under more relaxed selective pressures into a new activity by the fixation of substitutions over time (Innan and Kondrashov 2010). Additionally, whole genome duplications, which result in the duplication of an organism's entire set of chromosomes, can also happen, providing even more genetic material susceptible to evolve and be neo-functionalized (Clark and Donoghue 2018). Duplication not only contributes to changes in genome size but also has been demonstrated to be one of the most important forces into the innovation of the genetic repertories across eukaryotes (Sémon and Wolfe 2007; Fernández and Gabaldón 2020). Expanded gene families in fungi, animals and plants include for example diversen gene families of transcription factors, underscoring the diversity of the regulatory networks in these organisms (de Mendoza and Sebé-Pedrós 2019).

In contrast, the importance of lateral gene transfer (LGT) as a source of genomic innovation across eukaryotes has been less studied and is highly debated. One of the main arguments for this is that eukaryotes have strong barriers against LGT (Sibbald et al. 2020) although several mechanisms of LGT in eukaryotes have been proposed (Figure 7). In endosymbiosis these barriers can be broken, for example *Bigelowiella*

natans a cryptomonad that acquired photosynthesis through the secondary endosymbiosis of a green alga still conserves a very reduced genome of this endosymbiont in the form of a nucleomorph and has transfer a high number of genes to the nuclear genome of *B. natans* (Archibald et al. 2003; Raymond and Blankenship 2003). Other well studied examples of the acquisition of important functions are represented in several anaerobic lineages that acquired key genes from prokaryotes (Stairs, Leger, and Roger 2015). Interestingly, the genome of the foraminifer *Reticulomyxa filosa* also harbors an important proportion of prokaryotic genes (Glöckner et al. 2014) and more recently, LGT between Rhizaria and other eukaryotes have been also characterized (van Hooff and Eme 2023). Finally, a recent study suggests that fungi have acquired several metabolic capabilities through LGT, being a more important mechanism of gene gain than in metazoa (Ocaña-Pallarès et al. 2022). These studies have highlighted the role of LGT as an important source of innovation in diverse microbial eukaryotes.

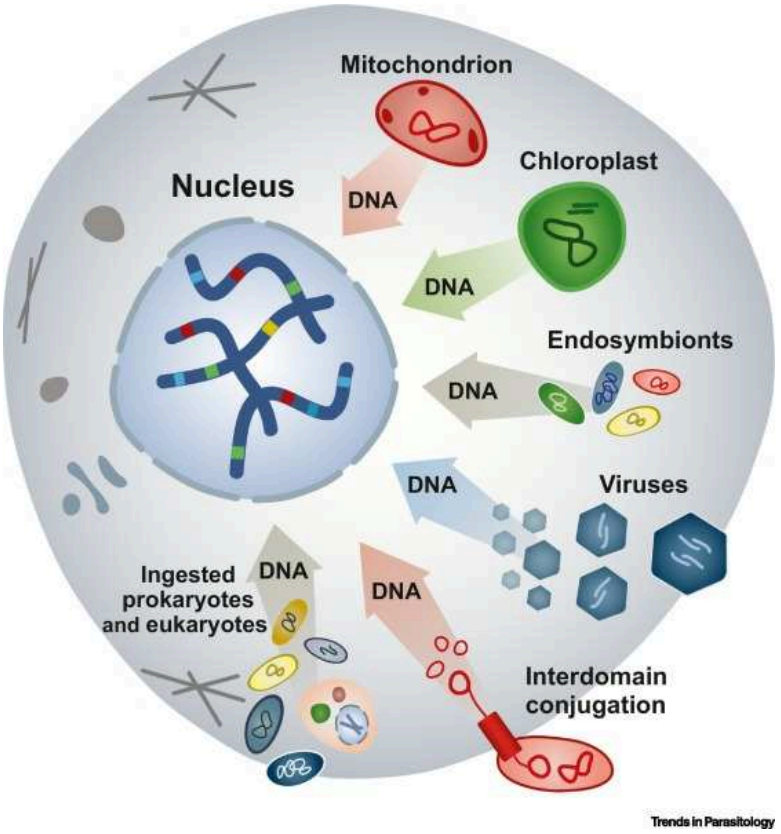

Figure 7. Scheme of the possible Sources of Foreign DNA Contributing to Lateral Gene Transfer (LGT) in Eukaryotes. Figure and caption from Sibbald et al. (2020).

1.3 Eukaryotic epigenome and epigenetic toolkit

The realization that identical cells can express different genes, led to the first speculations that there are changes in gene expression that are not caused by changes in DNA sequences (Holliday 2006). This is possible, in part, because genomes contain heritable information superimposed to the DNA sequence that can be inherited through generations. This information, also known as epigenome, constitutes a regulatory interface to the eukaryotic genome (Bestor, Chandler, and Feinberg 1994).

The cornerstones of the eukaryotic epigenome are the DNA methylation marks and the histone post-translational modifications. These marks interact with other mechanisms (such as the polycomb-trithorax system or non-coding RNA systems) in complex regulatory networks that determine the local activity of the genome by diverse mechanisms such as regulating the access of DNA-interacting proteins (such as transcription factors) to the DNA (Lamka et al. 2022; Gibney and Nolan 2010). The epigenome is influenced by environmental factors, such as diet, stress, and exposure to toxins and altogether, these mechanisms underlie a great extent of the complexity and physiological plasticity displayed by microbial and multicellular species (Weiner and Katz 2021; Lamka et al. 2022).

How the epigenetic mechanisms aroused and diversified in eukaryotes has been a long standing question. Recent studies suggest that the last eukaryotic common ancestor already encoded a complex toolkit of epigenetic related proteins (Weiner et al. 2020b). Some components of the chromatin even predate the origin of eukaryotes. Histones that can be found in archaea (Ammar et al. 2012; Stevens et al. 2020).

Unlike eukaryotic histones, archaeal histones generally lack tails and do not exhibit post-translational modifications, which are most likely eukaryotic innovations (Grau-Bové et al. 2022). Some works (Irwin and Richards 2023) have proposed that unique histone fusions observed in extant viruses could constitute relics of ancient histones found in stem eukaryotes and raised new questions on the evolutionary history of these proteins in deep evolutionary scales.

Some authors have proposed that, among other epigenetic mechanisms, DNA methylation has driven the diversification of the genome architecture in eukaryotes

through a process of genomic conflict and arms race between the regulatory mechanisms and mobile elements (Fedoroff 2012; Maurer-Alcalá and Katz 2015; Yi and Goodisman 2021).

1.3.1 DNA methylation is an ancient and widespread epigenetic mechanism

DNA methylation marks are ancient and phylogenetically widespread components of the epigenome. Indeed, these covalent modifications can be found in bacterial, archaeal and eukaryotic genomes in which play diverse roles. In prokaryotes, DNA methylation primarily functions alongside restriction modification (RM) systems acting to discriminate and destroy invading foreign DNA (Hampton, Watson, and Fineran 2020) as a defense mechanism. Nonetheless, several "orphan" DNA methyltransferases have been found to perform regulatory functions such as chromosome replication and regulation of transcription (Løbner-Olesen, Skovgaard, and Marinus 2005; Sánchez-Romero and Casadesús 2020), extending the notion of DNA methylation as an epigenetic mark towards prokaryotes.

In Eukaryotes DNA methylation is predominantly found as C5-methylcytosine (5mC). Often called "the fifth base", 5mC plays an important role in genome defense against mobile genetic elements in fungi (Bewick et al. 2019) and plants (Ritter and Niederhuth 2021) and is often associated with transcriptional silencing, establishment of the closed chromatin configuration, and repressive histone modifications (de Mendoza, Lister, and Bogdanovic 2019). In vertebrates, DNA methylation has been shown to be dynamic during development (Li and Zhang 2014) and to underlie the regulation of gene expression involved in the emergence of specialized traits such as cognitive function (de Mendoza et al. 2021). A recent study has also shown that 5mC patterns are tied with the lifespan across mammals (Haghani et al. 2023), although the causes of this association are still unclear.

The 5mC mark is introduced in the genome by C5-MTases. The origin of these enzymes can be traced to prokaryotic ancestors which have transferred DNMT proteins several times into eukaryotic species (Arkhipova et al. 2023). DNMT1 maintains DNA methylation in CpG sites of mammal genomes, while DNMT3 methylates new sites de

novo (de Mendoza, Lister, and Bogdanovic 2019). Other DNMTs such as DNMT2, 5 and 6 have been found in diverse eukaryotes (de Mendoza, Lister, and Bogdanovic 2019).

Since the first large-scale comparison of the DNA methylation landscapes were done (Zemach et al. 2010) a high diversity of DNA methylation landscapes has been revealed (Figure 8), highlighting the diverse biological roles that DNA methylation is playing in diverse species such as expression repression and DNA repair (de Mendoza, Lister, and Bogdanovic 2019; Zemach et al. 2010; Schmitz, Lewis, and Goll 2019).

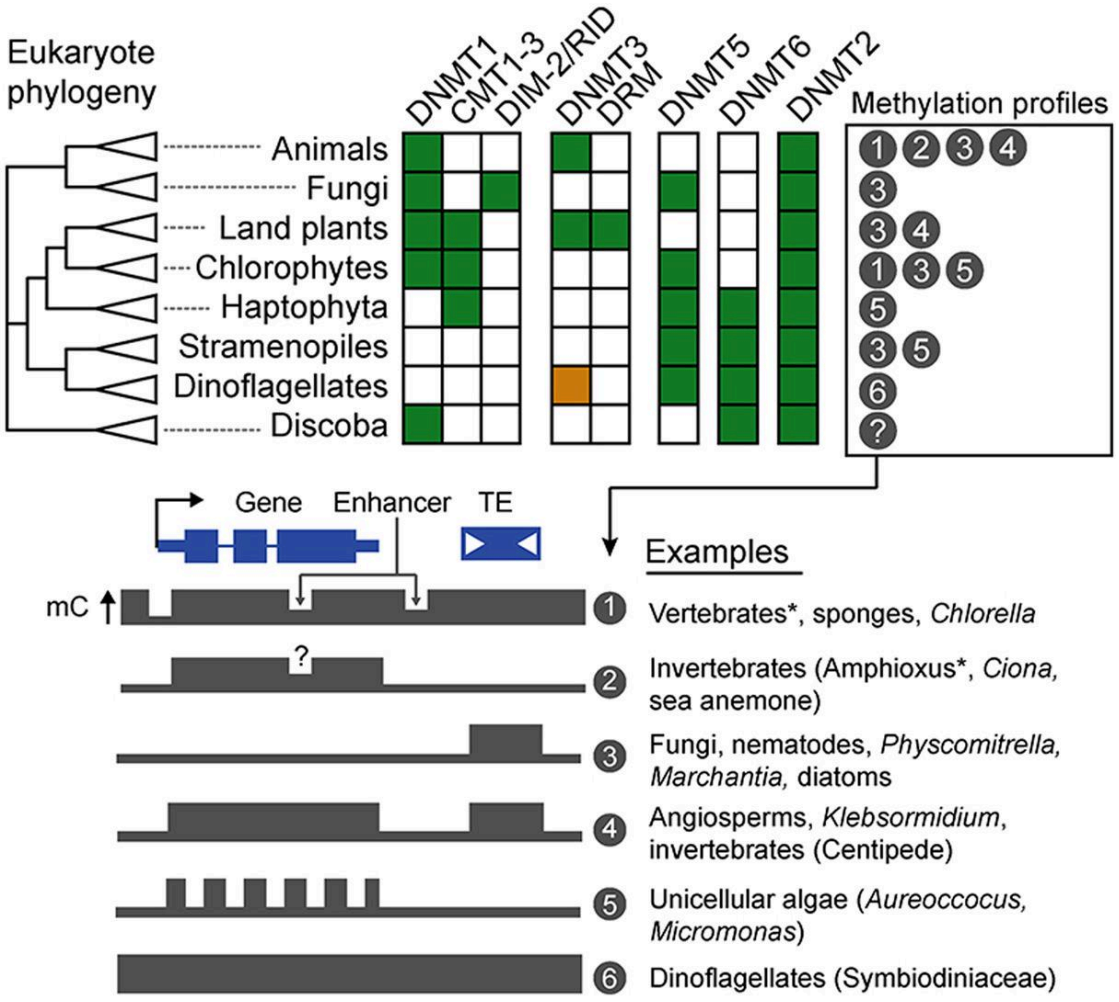

Figure 8. DNA methylation systems in eukaryotes. Up: Presence (green) or absence (white) of enzymes within major eukaryotic groups. Orange cells indicate the presence of enzymes that are encoded as part of a retrotransposon. Down: Major DNA methylation profiles described to date, with representative species. Asterisks indicate lineages where enhancer DNA demethylation has been described. Figure and caption adapted from (de Mendoza, Lister, and Bogdanovic 2019).

Moreover, DNA adenine methylases appear to have been acquired across eukaryotes such as ciliates, heterolobosea amoeboflagellates, and certain chlorophyte algae (Iyer, Abhiman, and Aravind 2011). N6-methyladenine (6mA), gained attention as a possible form of epigenetic modification in diverse eukaryotes, for example it has been shown to play a role in the chromatin 3D organization of the parasite trichomonas vaginalis (Lizarraga et al. 2020) and has been also shown to be prevalent across diverse viruses (Jeudy et al. 2020), although its conserved role in animals and plants remains unclear (Iyer, Zhang, and Aravind 2016; Boulias and Greer 2022; Bochtler and Fernandes 2021).

Finally, N4-methylcytosine (4mC), the third type of DNA methylation naturally occurring in bacteria, has recently been demonstrated to be present in rotifera (Rodriguez et al. 2022). These authors have also shown that 4mC can be recruited as an epigenetic mark in the genomes of these organisms and identified its underlying enzymatic machinery, supporting the idea that a horizontally transferred gene can become part of a complex regulatory system maintained by selection over a large evolutionary scale.

1.3.2 The intertwined evolution of the genome and the epigenome

Some of these ideas have been explored focusing on the diversity of the genome wide methylation landscapes of some clades such as metazoa (Zhou et al. 2020), land plants (Ritter and Niederhuth 2021) and fungi (Bewick et al. 2019). These studies have pointed out that DNA methylation has contributed to the integration and stabilization of mobile elements.

Moreover, under certain conditions, 5mC can undergo spontaneous deamination, leading to C to T mutation in the DNA sequence. If this mutation is not corrected during DNA replication or repair processes, it becomes a permanent change in the DNA sequence. Zhou and collaborators (2020) argue that this process has been related with the emergence and expansion of regulatory elements in animals.

Some authors have proposed that an epigenetic mark can be stably inherited across generations, leading to "epigenetic assimilation" and potentially providing a selectable advantage to increase population fitness and promote adaptation,

diversification, and speciation (Weiner and Katz 2021) (Figure 9). In addition, if an epigenetic modification is followed by a genetic mutation, it could become permanently integrated into the genome through the process of genetic assimilation.

So far, the main limitation to further test these hypotheses and investigate the macroevolution of the methylome and other components of the epigenome has been the lack of epigenomic studies in the bulk diversity of the tree of eukaryotes.

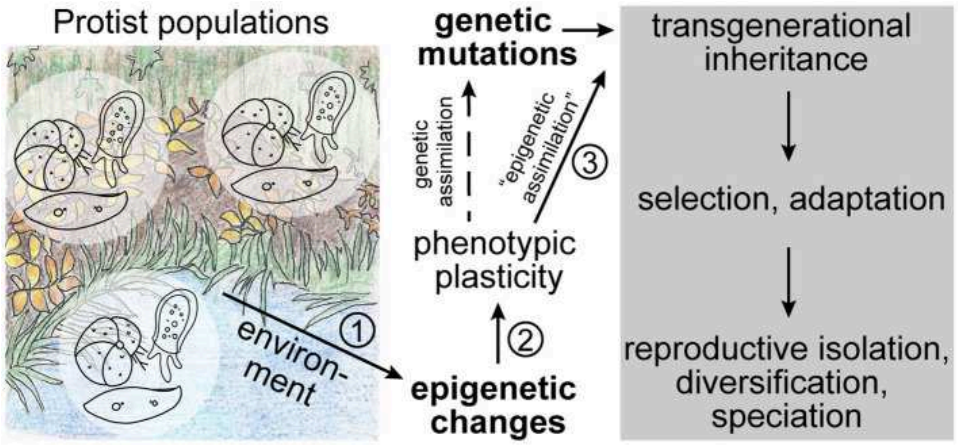

Figure 9. Theoretical sequence of events in ecological speciation driven by epigenetics. Figure and caption adapted from Weiner and Katz (2021).

2. Objectives

The understanding of the eukaryotic (epi)genome diversity and evolution has been historically limited by our access to model species across the tree of life. Indeed, the genomic sequences required to conduct functional studies of protists, which represent the vast majority of the eukaryotic diversity, are scarce and strongly biased towards parasitic model species.

Moreover, despite their evolutionary relevance, protists belonging to deeply branching lineages within the eukaryotic tree of life have been poorly explored, hampering the investigation of the deep evolution of eukaryotes. Therefore, the general objective of this thesis was to include new and diverse model species of protists in the reconstruction of ancient evolutionary events shaping the eukaryotic (epi)genome by the generation and comparative analysis of genomic data for these species. The particular objectives were:

I. Expanding the genome resources for orphan eukaryotes.

The orphan branches of the eukaryotic tree of life (eToL) potentially constituting anciently diverged eukaryotic lineages are often represented by very few species and no genomic data. Therefore, the first objective of my PhD was generating, assembling and curating high quality genomic data for of mantamonads and ancyromonads previously gathered in our culture collection.

We combined state of the art long and short read sequencing methods, genome assembly, and custom approaches of data curation with this purpose. The assessment of genome completeness and the first characterization of the architecture and functions encoded by the nuclear genomes of *Mantamonas spyhraneae* and *Ancyromonas sigmoides* are described in the manuscript 1 and 2 of this thesis, respectively.

II. Studying the evolutionary processes shaping the genome content of ancyromonads in a deep evolutionary scale using a phylogenetically informed framework.

The second objective of this thesis was to reconstruct the evolutionary history of the genomic content of ancyromonads since their divergence from other major eukaryotic supergroups. To improve the resolution of ancient evolutionary events across eukaryotes we sequenced six additional genomes of several new species of ancyromonads. Then, we reconstructed the evolutionary history of diverse eukaryotes and the gene families distributed across them with the purpose of gaining insight into the importance of gene duplication, transfer, loss and origination into the genome evolution of these organisms in comparison to other eukaryotes. As well we aimed to shed light into some understudied aspects of the biology of ancyromonads. These results are presented in the draft manuscript 3 of this thesis.

III. Exploring the diversity of the epigenetic mechanisms in ancyromonads.

We aimed to gain insights into the nature and diversity of genome regulation systems of ancyromonads using the previously generated genomic resources and the cultures of these organisms. We explored the 5mC DNA methylation patterns in ancyromonads by using Whole Genome Bisulfite Sequencing. We also characterized the gene expression profiles of *Ancyromonas sigmoides* under different environmental conditions. Through the integration of these datasets and their further comparison with available information from other organisms we aim to shed light into the conservation of these systems across eukaryotes as well as the potential particularities and role of these systems in this ancient and divergent eukaryotic lineage. The preliminary results tied to this objective are described in the draft of the manuscript 4 of this thesis

3. Material and methods

The DEEM team, in collaboration with the Kim's lab (formerly at the National Museum of Natural History in New York, USA), has previously isolated and gathered a collection of cultures of diverse protists. During my project, we generated diverse genomic datasets from several mantamonads and ancyromonads species of this collection (Table 1).

Table 1. Isolates and sampling sites for all cultured ancyromonad and mantamonad strains characterized in this study. Previously described species are marked with an asterisk (*). G: genome, T: transcriptome, M: methylome data.

Taxonomic name (isolate)	Culture collection	Sampling site	Sampling environment	Generated datasets
Mantamonas sphyraenae (STR306)	DEEM	Iriomote Island, Japan	Barracuda skin	G,T
Mantamonas vickermanii (CROMAN19)	DEEM	lagoon Malo jezero, Croatia	Sediment	T
*Ancyromonas sigmoides** (B70)	CCAP1958/3	Near Srednii Island, Russia	Littoral waters	G, T, M
Ancyromonas mediterranea (C362)	DEEM	Villefranche Bay, France	Mesopelagic water column (250 m depth)	G, T, M
*Nutomonas limna** (ORSAYFEB19ANCY)	DEEM	University campus, Orsay, France	Ditch water	G, T, M
*Striomonas longa** (ncfw)	CCAP1958/5	Boiling Springs North Carolina, USA	Sediment	G, T, M
Nyramonas silfraensis (ORSLAND19S2)	DEEM	Silfra rift, Iceland	Sediment from the walls of the rift (~ 2 m depth)	G, T, M
Planomonas micra (PMROSKO2018)	DEEM	Roc'h ar Bleïz, Roscoff, France	Sediment in a tide pool	G, T, M
Fabomonas mesopelagica (A153)	DEEM	Villefranche Bay, France	Mesopelagic water column (250 m depth)	G, T, M

3.1 Workflows for genome sequencing and assembly

DNA and RNA was extracted from the bulk cultures of six ancyromonad species (Figure 10a-b) and sent for sequencing using Illumina HiSeq to Eurofins Genomics, Germany. To improve the coverage of species hard to grow in high quantities, additional genomic data was generated from sorted samples of ancyromonad cells that were further lysed and amplified using Whole Genome Amplification and sequenced in a minIT Oxford Nanopore Technologies (ONT) platform (Figure 10c). These experiments were primarily conducted by Naoji Yubuki and Maria Ciobanu. The genomic and transcriptomic data generated in Kim's Lab for the species Mantamonas *sphyraenae* and *Ancyromonas sigmoides* was generated through a different approach, in which nucleic acids were extracted from bulk cultures (Figure 10d).

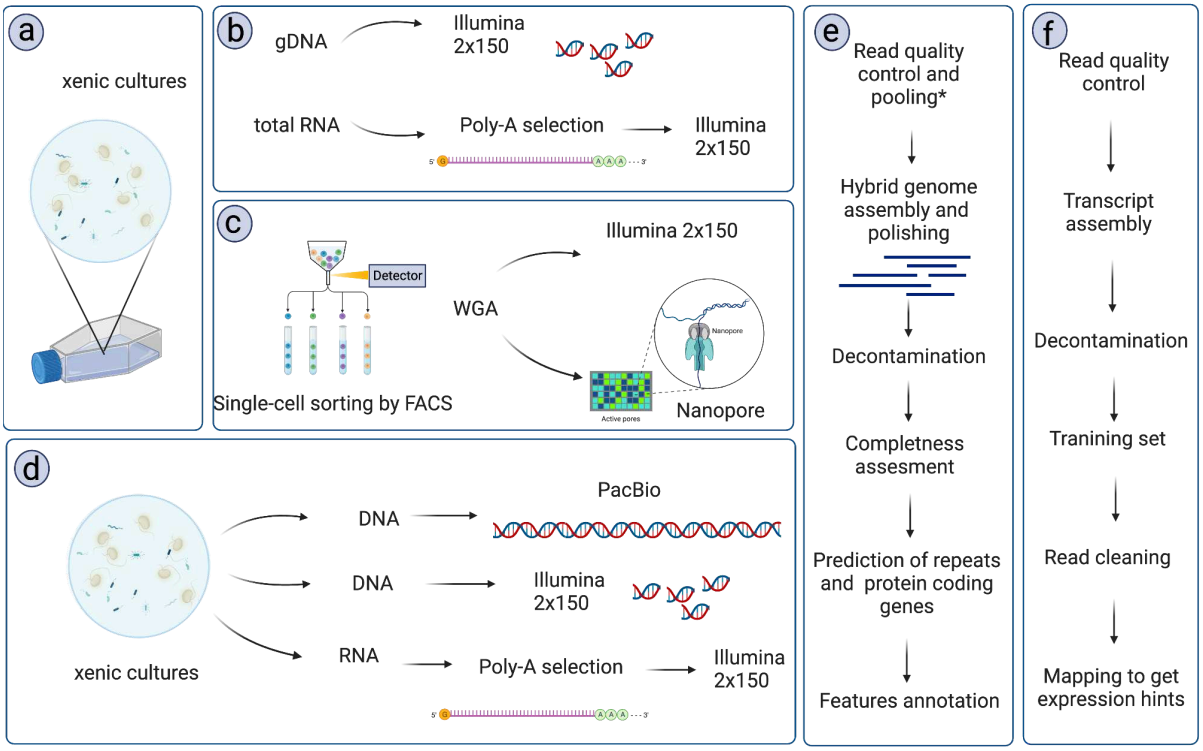

Figure 10. Overview of methods employed for the genome and transcriptome generation of protists. a. Established cultures of protists b. Nucleic acid extraction and sequencing from bulk cultures c. Combination of Fluorescence activated Cell Sorting (FACS) and WGA for the generation of additional genomic data d. Genomic data was obtained from bulk cultures using a different approach in Kim's lab at the NMNH. e. Genome assembly and analysis (*pooling of WGA amplified data was done after read coverage normalization). f. Transcriptome datasets processing to generate a training set and expression hints employed in the prediction of protein coding genes on the genomic sequences.

A summary of the downstream assembly and analysis of these genomic datasets is presented in Figure 10(e-f). The genomic libraries were assembled using several strategies for comparison purposes and the best assemblies were chosen based on the contiguity and completeness of the resulting genomic sequence. For further details on the methods employed see the method sections of the following chapters.

3.2 Comparative genomic analyses and ancestral reconstructions

To reconstruct the evolutionary history of genomes and gene families distributed in ancyromonads and other eukaryotes we clustered homologous gene families distributed across our genomes and diverse eukaryotes employing Orthofinder (Emms and Kelly 2019) and the comprehensive eukaryotic proteome database Eukprot v3 (Richter et al. 2022). The Genome Taxonomy Database (GTDB) (Parks et al. 2022) and the non redundant (nr) database of the National Center of Biotechnology Information were also employed to look for homologous sequences that originated outside eukaryotes. We reconstructed maximum likelihood (ML) phylogenies for the gene families using IQ-Tree v2 (Minh et al. 2020). Additionally we reconstructed a ML species phylogeny combining the set of single gene copy orthologues retrieved from the Orthofinder pipeline and a phylogenetic constraint based on the currently unsolved model of the eukaryotic tree of life.

Furthermore, Amalgamated Likelihood Estimation (ALE) reconciliation analysis (Szöllõsi et al. 2013) was employed. This analysis implements an algorithm that maximizes the phylogenetic likelihood representing the relationship between the gene tree and species tree (Williams et al. 2023). The results of ALE reconciliations include branch-wise estimates of gene duplication, transfer, vertical inheritance and loss events, as well as the estimated parameters for each gene family. To contrast these inferences we also conducted a Dollo parsimony analysis using COUNT (Csurös 2010), that infers the ancestral gene content and gene family gain and loss based only on the phylogenetic distribution of the gene families across species.

3.3 DNA methylation profiling of ancyromonads

To perform an initial exploration of the epigenome in ancyromonads we used Whole Genome Bisulfite (WGB) sequencing to study the landscapes of 5-methylcytosine (5mC). During WGB sequencing, the genomic DNA is treated with sodium bisulfite, which converts unmethylated cytosines to uracils while leaving methylated cytosines unchanged. This conversion is specific to unmethylated cytosines and provides a way to distinguish methylated and unmethylated sites. The bisulfite-converted DNA is then used to generate a sequencing library. The library is sequenced using high-throughput Illumina sequencing. In our case, DNA was obtained from the bulk cultures of seven species of ancyromonads. This material was sent to Eurofins Genomics (Germany), where bisulfite libraries were generated with EZ-96 DNA Methylation-Lightning MagPrep kit (zimo Research) and sequenced in a Illumina HiSeq platform.

The sequenced reads are then aligned to a reference genome of each species using Bismarck (Krueger and Andrews 2011) and Batmeth2 (Zhou et al. 2019) was employed to obtain methylation calls for the mapped cytosine positions of the genome with a minimum coverage of 10x.

The presence of a cytosine in the sequenced read at a cytosine site indicates that the original cytosine was methylated, while a thymine indicates an unmethylated cytosine. The aligned reads are processed to determine the methylation status at each cytosine site during the methylation calling. The output consists in a quantitative measure of methylation, represented as a percentage of methylated cytosines among the total reads mapped to a site. This information can be used to generate genome wide methylation profiles with single-base resolution.

3.4 Environmental shifts experiment on *Ancyromonas sigmoides*

To gain further insight into the genome regulation mechanisms of ancyromonads we design an experiment in which we could study the effect of several conditions into the variations of the DNA methylation patterns across the genome and the gene expression

(Figure 11). We chose *Ancyromonas sigmoides* species with this purpose because it has the most contiguous genome and also because it grows faster than other species. Its culture doubling time lasts approximately four days after which they are usually transferred to new media to scale the culture and to avoid starvation. After scaling, cells were retrieved and pooled into an homogenous inoculum that was then redistributed in culture flasks that were put in five different conditions as well as a control with three replicates of each.

We tested several conditions in which ancyromonad could survive but were potentially stressed and therefore we could observe a response in their transcriptomes and methylomes. For example, we could observe that ancyromonads cannot survive in temperatures over 35 degrees. The experiment was then performed over seven days, in which the first phase consisted in culture scaling. In contrast ancyromonads seem to grow well under low oxygen conditions in which cells were observed to be abundant and with movement after 4 days. Moreover, under low temperature, ancyromonad cells were observed to be static, although cells were abundant. No differences were observed between the control and the changes in salinity. In the day seven cells were harvested and DNA and RNA were extracted from each the replicates to be sequenced in Eurofins Genomics, Germany as previously described.

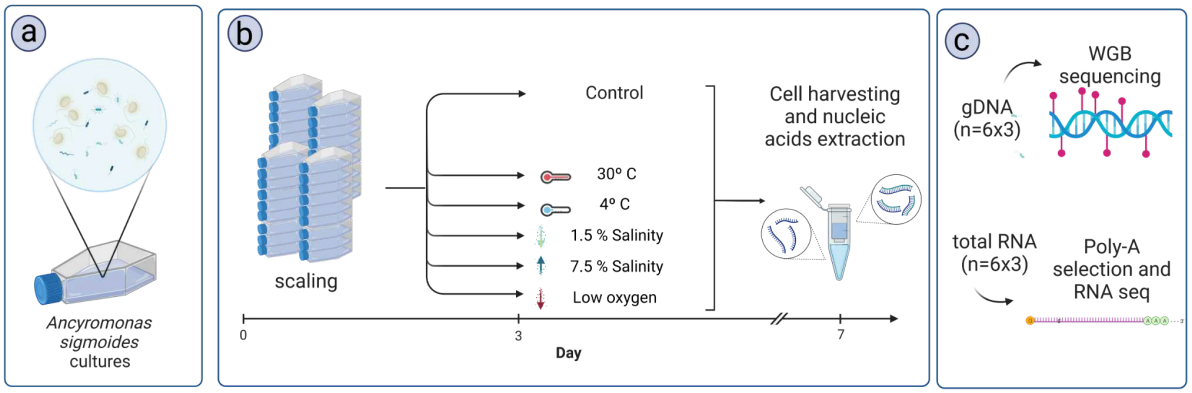

Figure 11. Summary of the methodological approach to explore the responses to the shifting environment of *Ancyromonad sigmoides*.

4. Description of two new species of *Mantamonas* and their genomic datasets

Context and results summary

Mantamonas is a genus of marine gliding flagellates which was initially thought to be related to the lineages Apusomonadida and Ancyromonadida (Glücksman et al. 2011), but recent transcriptome-based phylogenomic analyses placed it as sister to the new CRuMs supergroup-ranking clade comprising also Collodictyonidae and Rigifilidae (Brown et al. 2018). Currently CRuMs is only represented by four species with partial transcriptomic data. In this work we isolated and described two new species *Mantamonas sphyraenae sp. nov.* and *Mantamonas vickermani sp. nov.*.

Through the combination of PacBio and Illumina reads, we assembled a contiguous and highly complete genome sequence for *M. sphyraenae*. The genome of this species was inferred to be diploid and bears telomeric repeats in both ends of the majority of the contigs. Similarly the sequenced transcriptome of *M. vickermani* conserves a high proportion of BUSCO markers suggesting it nearly represents the gene complement of this species.

Our phylogenetic analysis using 182 conserved protein markers confirmed the monophyly of *Mantamonas* genus within the CRuMs supergroup and places *M. sphyraenae* as sister to a clade containing *M. vickermani* and *M. plastica*.

Moreover, when comparing the gene set for the CRuMs species, only 1.7K gene families were found to be conserved across all CRuMs. In comparison, *Mantamonas* unique core comprised around 4K gene families, mostly comprising genes with unknown functions.

Finally, the presence of rare paralogues of the membrane-trafficking system proteins such as the AP5 complex and syntaxin 17 in *Mantamonas* species suggests retention of anciently originated protein machineries during evolution.

This work constitutes a significant improvement into the genomic representation of CRuMs and also an important foundation for further comparative functional studies. This manuscript was prepared in collaboration with the teams lead by Enunsoo Kim and Joel Dacks and published in *ScientificData* on September 9th, 2023.

scientific data

Check for updates

OPEN

DATA DESCRIPTOR

One high quality genome and two transcriptome datasets for new species of *Mantamonas*, a deep-branching eukaryote clade

Jazmin Blaz[1], Luis Javier Galindo[1,2], Aaron A. Heiss[3,4,5], Harpreet Kaur[6], Guifré Torruella[1], Ashley Yang[4], L. Alexa Thompson[6], Alexander Filbert[6], Sally Warring[4,7], Apurva Narechania[4], Takashi Shiratori[3], Ken-ichiro Ishida[3], Joel B. Dacks[5,8], Purificación López-García[1], David Moreira[1], Eunsoo Kim[4,9] ✉ & Laura Eme[1] ✉

Mantamonads were long considered to represent an "orphan" lineage in the tree of eukaryotes, likely branching near the most frequently assumed position for the root of eukaryotes. Recent phylogenomic analyses have placed them as part of the "CRuMs" supergroup, along with collodictyonids and rigifilids. This supergroup appears to branch at the base of Amorphea, making it of special importance for understanding the deep evolutionary history of eukaryotes. However, the lack of representative species and complete genomic data associated with them has hampered the investigation of their biology and evolution. Here, we isolated and described two new species of mantamonads, *Mantamonas vickermani* sp. nov. and *Mantamonas sphyraenae* sp. nov., for each of which we generated transcriptomic sequence data, as well as a high-quality genome for the latter. The estimated size of the *M. sphyraenae* genome is 25 Mb; our de novo assembly appears to be highly contiguous and complete with 9,416 predicted protein-coding genes. This near-chromosome-scale genome assembly is the first described for the CRuMs supergroup.

Background & Summary

Free-living heterotrophic flagellates play important roles in the nutrient cycling of marine and freshwater ecosystems. However, the extent of their genomic diversity is still dramatically uncharacterized. Amongst the lesser-known of these is *Mantamonas*, a genus of marine gliding flagellates initially described as very divergent from all other known eukaryotes[1]. Although *Mantamonas* was originally thought to be related to the poorly-known lineages Apusomonadida and Ancyromonadida, based on ribosomal RNA gene phylogenies and some of their morphological characteristics[1], recent transcriptome-based phylogenomic analyses instead robustly placed *Mantamonas plastica* as sister to a clade comprising Collodictyonidae (also known as diphylleids) and Rigifilidae, altogether forming the "CRuMs" supergroup[2,3]. This clade presents diverse cell morphologies and branches at the base of Amorphea[2,4] (Amoebozoa plus Obazoa, the latter including animals and fungi, among others). The genomic exploration of members of this supergroup therefore represents an important resource for uncovering the characteristics of this deep-branching clade, and may help us better understand evolutionary transitions within the eukaryotic tree of life, such as the acquisition of complex multicellularity

[1]Unité d'Ecologie Systématique et Evolution, CNRS, Université Paris-Saclay, AgroParisTech, Gif-sur-Yvette, France. [2]Department of Biology, University of Oxford, Oxford, United Kingdom. [3]Institute of Life and Environmental Sciences, University of Tsukuba, Tsukuba, Japan. [4]Division of Invertebrate Zoology, American Museum of Natural History, New York, NY, USA. [5]Department of Oceanography, Kyungpook National University, Daegu, South Korea. [6]Division of Infectious Disease, Department of Medicine, University of Alberta and Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada. [7]Earlham Institute, Norwich Research Park, Norwich, United Kingdom. [8]Centre for Life's Origin and Evolution, Department of Genetics, Evolution & Environment, University College London, London, United Kingdom. [9]Division of EcoScience, Ewha Womans University, Seoul, South Korea. ✉e-mail: eunsookim@ewha.ac.kr; laura.eme@universite-paris-saclay.fr

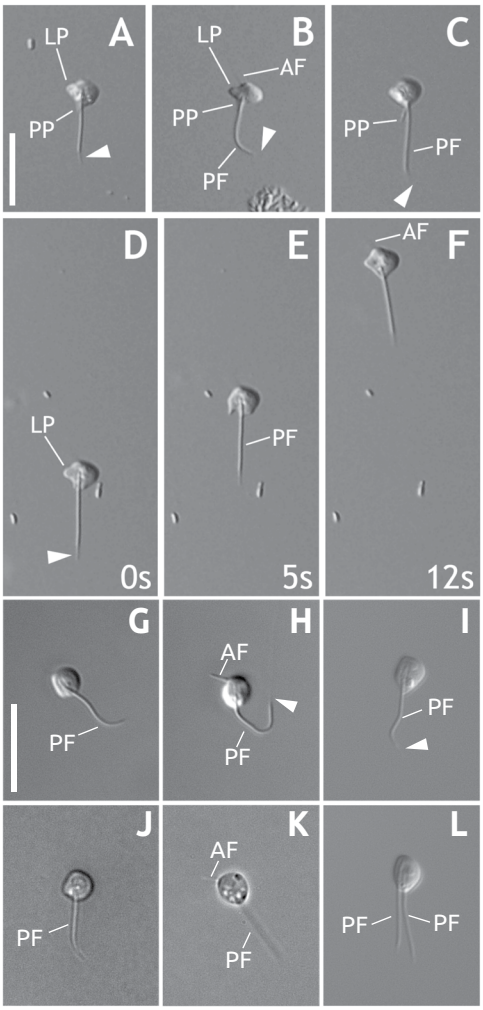

Fig. 1 General morphology of *Mantamonas sphyraenae* sp. nov. and *Mantamonas vickermani* sp. nov. (**a**–**c**) Differential interference contrast light micrographs of living M. sphyraenae cells. Note acroneme (white arrowheads), most visible in panel (**a**) but present in all micrographs. The extremely thin anterior flagellum is visible in panel (**b**). The left projection, present in all cells, is most distinct in (**b**). A posterior protrusion is often visible, usually parallel and immediately adjacent to the posterior flagellum (**a**,**b**), but sometimes at an angle to it (**c**). (**d**–**f**) Individual M. sphyraenae cell imaged over a 12-second period; numbers in lower right indicate elapsed time in seconds. Note the plastic nature of the cell and lack of movement of the posterior flagellum except to trail behind the cell body. (**g**–**l**) Phase and differential interference contrast light micrographs of living interphase M. vickermani cells. Note contrast between thick and long posterior flagellum and thin and short anterior flagellum in (**g**,**k**). (**l**) Laterally dividing cell of M. vickermani with two posterior flagella. Scale bars: 10 µm. AF = anterior flagellum; LP = left projection; PF = posterior flagellum; PP = posterior protrusion; arrowhead = acroneme.

in several lineages of the Obazoa. However, to date, only partial transcriptomic data is available for a handful of CRuMs taxa, including *M. plastica*[2,3]. Here, we isolated and described two new species of mantamonads, *Mantamonas sphyraenae* sp. nov. and *Mantamonas vickermani* sp. nov., and generated a high-quality nuclear genomic assembly for the former and transcriptomic assemblies for both species.

Overall, the cell morphology and behavior under light microscopy of these two new species (Fig. 1, Movie 1 and Movie 2) are comparable to what was reported in the original description of the genus *Mantamonas*[1] and to our own observations of the type strain of *M. plastica*. Nonetheless, our strains appear to be slightly smaller than the 5 × 5 µm dimensions of *M. plastica*. Cells of this genus have one anterior and one posterior flagellum. They are flattened and somewhat plastic, with shapes ranging from wide, with more or less pointed lateral "wings" resembling the fins of a manta ray, to kite-shaped, to oval, to spherical. The left side of the cell body often displays a characteristic blunt projection, which we sometimes observed in our new strains, although less conspicuously (Fig. 1; see the detailed morphological description of each of the new species in *Methods* and formal species description in *Data Usage Notes*).

All previously known mantamonad strains were isolated from marine sediments[1], which was also the case for our strain *M. vickermani* sp. nov., isolated from marine lagoon sediment. However, we isolated the other strain (*M. sphyraenae* sp. nov.) from the skin surface of a barracuda, which could suggest that either this species

	Mantamonas sphyraenae	*Mantamonas sphyraenae*	*Mantamonas vickermani*
Assembly type	genome	transcriptome	transcriptome
Assembly length (Mb)	25.06 (31.49)	20.52	19.78
Number of contigs	78 (199)	9,255	9,796
Contig mean length (Kb)	321.30	2.218	2.019
Longest contig (Kb)	751.365	32.28	21.503
Shortest contig (Kb)	17.266	0.0202	0.0201
N50 (Kb)	375.07	3.05	2.79
L50	26	1,917	2,083
GC content	59.19	59.02	46.4
Total repeat content	12.12%	—	—

Table 1. Genomic and transcriptomic assemblies statistics for *Mantamonas sphyraenae* sp. nov. and *Mantamonas vickermani* sp. nov. Values within parentheses correspond to primary plus associate contigs produced by FALCON.

Assembly approach	Canu	Falcon	MaSuRCa
Total length (Mb)	27.35	25.06	26.11
Number of contigs	172	78	136
Mean length (bp)	159,014.56	321,299.41	191,995.24
Longest contig (bp)	732,584	751,365	1,133,621
Shortest contig (bp)	20,756	17,266	1,222
N_count	0	0	4,688
Gaps	0	0	9
N50 (bp)	303,774	375,077	386,663
N50n	32	26	24
N70 (bp)	224,361	300,753	269,297
N70n	52	41	40
N90 (bp)	50,673	226,430	146,039
N90n	95	60	65
BUSCO eukaryota odb10	C:89.1%[S:82.0%,D:7.1%], F:2.0%,M:8.9%	C:91.4%[S:90.6%,D:0.8%],F:2.0%,M:6.6%	C:89.8%[S:86.7%,D:3.1%], F:2.0%,M:8.2%

Table 2. Mantamonas sphyraenae sp. nov. genome assembly statistics produced by the tested assembly strategies.

is epizootic (normally inhabiting the skin of the fish) or that the cells that we isolated were dislodged from their normal habitat and adhered to the fish skin by chance. Additional sampling and culturing efforts should help resolve this matter.

The assembled nuclear genome sequence of *M. sphyraenae* is highly contiguous (Table 1). This genome sequence was generated using long (PacBio) and short (Illumina) reads (see Methods). The average sequencing coverage was 112x for PacBio and 115x for Illumina. Three different genome assembly strategies, using Canu[5], FALCON[6], and MaSuRCA[7], yielded comparable results (see Methods, Table 2), with >90% representation of the 255 Benchmarking Universal Single Copy Orthologs (BUSCO[8]) of the eukaryota_odb10 dataset (Fig. 2), indicating high completeness. For downstream analyses, we opted to use the FALCON assembly because it was the most contiguous of the three, with the majority of the contigs (59 out of 78 primary contigs) bearing TTAGGG telomeric repeats at both ends. In addition, 14 of the remaining contigs had telomeric repeats at one end. While the presence of such conserved motives towards the end of the contigs suggests the complete assembly of most of the chromosomes and leads to an estimation of ~66 pairs of chromosomes in the *M. sphyraenae* nucleus, experimental evidence is needed to confirm the chromosome number in this species. Biallelic single nucleotide polymorphism (SNP) frequencies cluster around a ratio of 0.5/0.5 for each major/minor allele (Fig. 3a). This is indicative of a diploid genome, which was also supported by the statistical model of SNP frequency distribution (Table 3).

The *M. sphyraenae* genome contains 9,416 predicted protein coding sequences. Genes have an average length of 2,282 bp and are mostly mono-exonic (Fig. 3b). *De novo* characterization of repetitive elements indicates that around 12% of the genome is represented by transposable elements and other repeats. While some of these were classified into different known families of DNA transposons and long terminal repeat (LTR) retroelements, the vast majority comprises unclassified types (Fig. 3c). In comparison, the transcriptome assembly of *Mantamonas sphyraenae* contains 9,256 contigs from which we predicted 8,885 non-redundant proteins and the presence of 85.5% of BUSCO eukaryota_odb10 gene set (Fig. 2). 96% of these proteins are also found in the genome-based

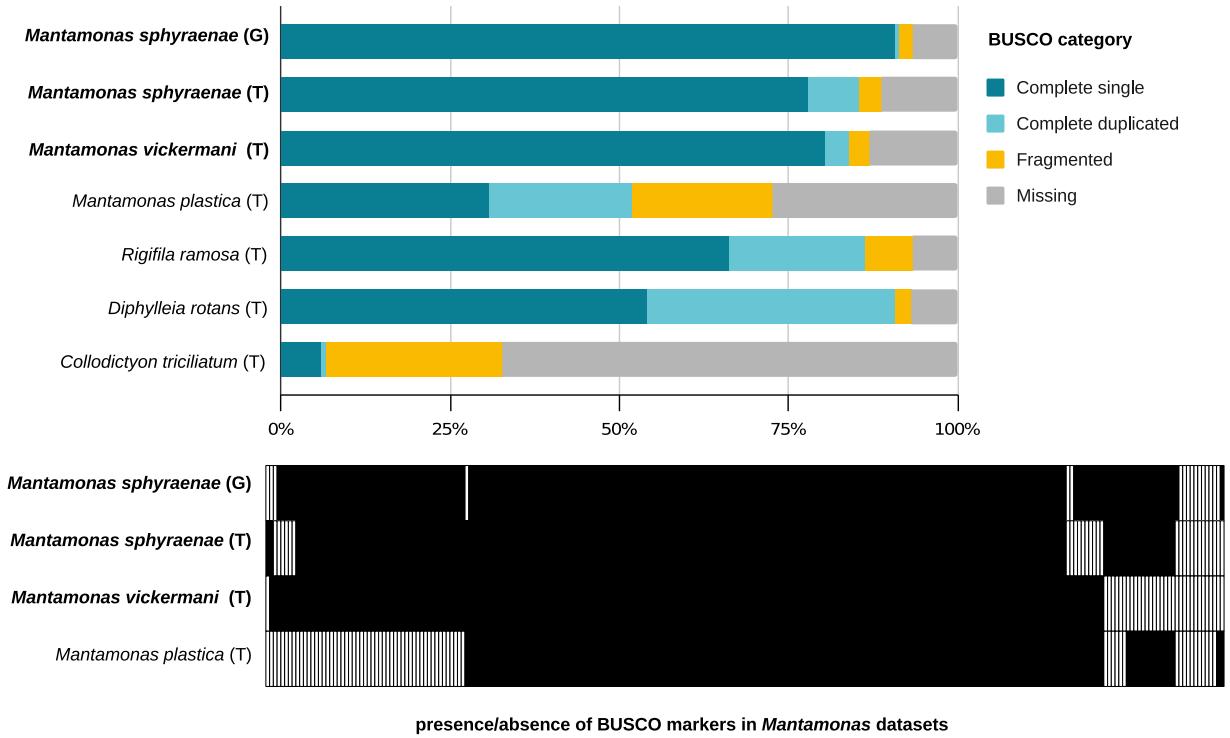

Fig. 2 Distribution of BUSCO orthologs (eukaryota_odb10) in inferred proteomes of mantamonad assemblies of this study (in bold) in comparison with those of other members of the CRuMs supergroup. Proteomes were inferred from genome (G) and transcriptome (T) assemblies. The top panel represents the BUSCO output for each CRuMs dataset, whereas bottom panel illustrates the patterns of presence/absence of each BUSCO gene (X axis) in the four Mantamonas predicted proteomes.

predicted proteome, suggesting that the genome assembly represents nearly the protein repertoire represented in the transcriptome (see details in the Technical validation section, Completeness analysis).

The *de novo* assembled transcriptome of *M. vickermani* had an average sequencing coverage of 80x and led to the inference of 9,561 non-redundant proteins. As for the genome and transcriptome of *M. sphyraenae*, the proteome inferred from this transcriptome resulted in a high BUSCO score, indicating a high completeness of the predicted gene complement for this species (Fig. 2). Some BUSCO genes are consistently missing in all the four *Mantamonas* predicted proteomes, suggesting a true absence of these genes in the genus.

We inferred the phylogenetic relationships of our species within the CRuMs clade using publicly available data to reconstruct a dataset of 182 conserved protein markers and recovered the monophyly of the *Mantamonas* genus and the placement of *M. sphyraenae* as sister to a clade containing *M. vickermanii* and *M. plastica* (details in Methods, Phylogenomics analyses).

To explore the gene content diversity of our new mantamonad species, we annotated the predicted proteomes genes with EggNOG mapper[9] and reconstructed the minimal core proteome for the genus Mantamonas and the CRuMs lineage (see details in Methods CRuMs orthologue analyses).

Finally, as an additional way of assessing the completeness of the *M. sphyraenae* and *M. vickermanii* sequence data and capturing a sense of the complexity of the cellular systems in these organisms, we interrogated the complement of one well-studied set of proteins, the membrane-trafficking system. This complex protein machinery underpins normal cellular function and is critical for feeding, cell growth, and interaction with the extracellular environment[10]. While some proteins are highly conserved across eukaryotic lineages, others have rarely been retained during evolution but were nonetheless present in the Last Eukaryotic Common Ancestor (LECA)[10]. Among them, the so-called "jotnarlogs" represent LECA proteins present in diverse extant eukaryotes but not in the major opisthokont model organisms.

The identification of homologs of the majority of the protein complement associated with the membrane trafficking system as well as some jortnalogs in the proteomes of the new *Mantamonas* species (details in Methods, Analysis of the conservation of the membrane-trafficking system complement) corroborated the high completeness of our genomic and transcriptomic datasets, and suggests that these datasets may provide interesting insights in the evolution of anciently originated protein machineries. Overall, our new *Mantamonas* nuclear genome and transcriptome sequences provide high quality data for a major, yet poorly known, eukaryotic supergroup. They will allow more comprehensive comparative studies of genetic diversity in microbial eukaryotes and a better understanding of deep eukaryotic evolution.

Genome ploidy	Delta log-likelihood values
Diploid	10,944
Triploid	161,437
Tetraploid	104,902

Table 3. nQuire Gaussian Mixture Model delta log-likelihood values for the *Mantamonas sphyraenae* genome.

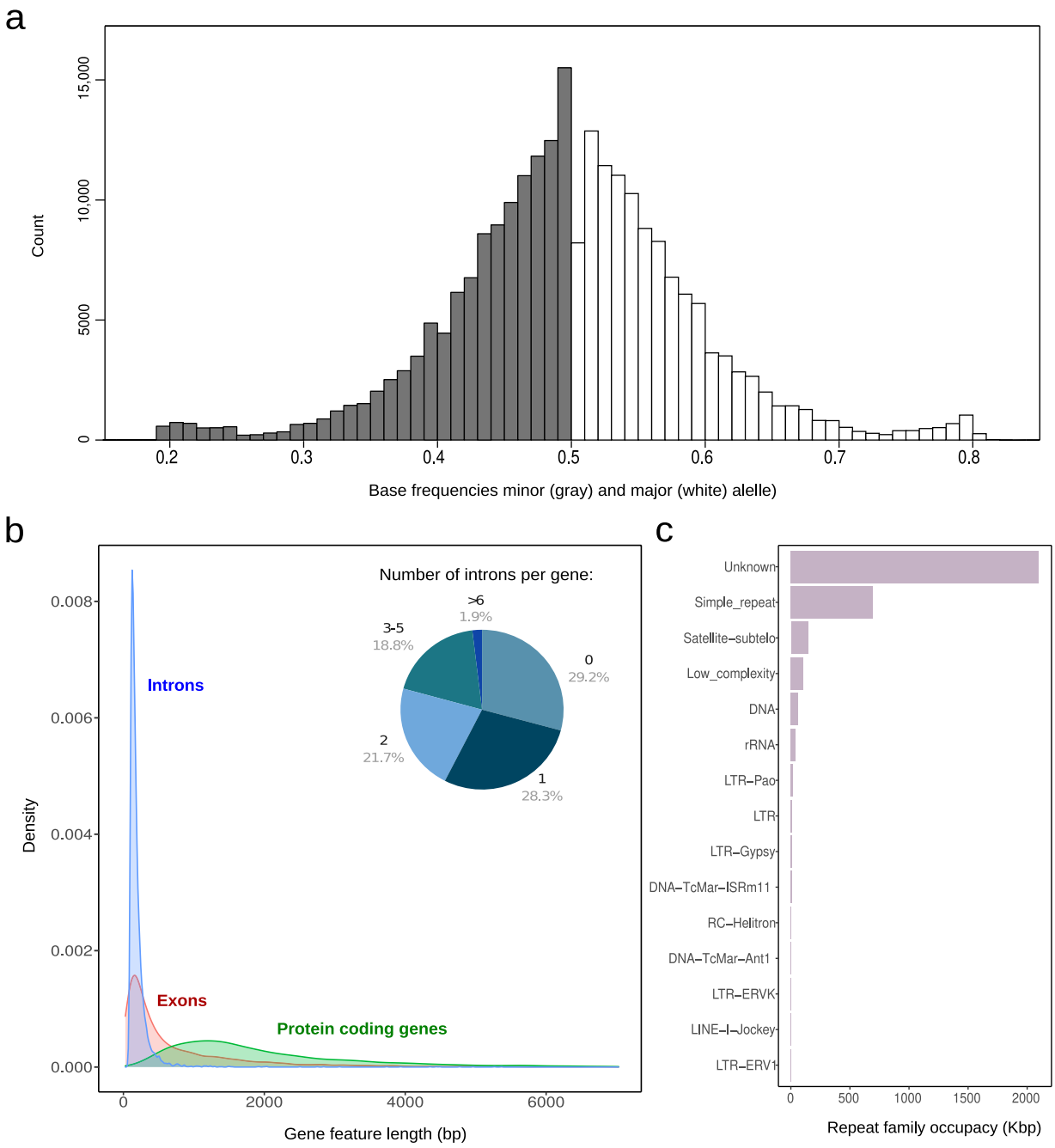

Fig. 3 Genomic features of *Mantamonas sphyraenae* sp. nov. (**a**) Biallelic SNP frequency distribution. (**b**) Length distribution and intron frequency of protein-coding genes. (**c**) Genomic occupancy of the families of repetitive elements identified *de novo*.

Methods

Isolation and microscopy of *Mantamonas sphyraenae* sp. nov. *Mantamonas sphyraenae* SRT-306 was collected on 26 Sep. 2013 from the surface of a barracuda caught in a lagoon on Iriomote Island, Taketomi,

Okinawa Prefecture, Japan (24° 23′ 36.762″ N, 123° 45′ 22.572″ E). It was isolated manually from the rough sample with a micropipette, and maintained in Erd-Schreiber medium[11] fortified with 2.5% (final volume) freshwater Cerophyl medium (ATCC 802). Stock cultures were kept in 8 ml volumes in 25 ml culture flasks at 16–18 °C, and transferred at three-week intervals. Bulk cultures were grown at room temperature in 10 cm Petri plates containing ~10 ml medium.

Live cells were observed on an Zeiss Axiovert 100 M inverted microscope equipped with DIC and phase contrast optics. Images were captured with an Olympus DP73 17.28-megapixel camera. Morphometric data were obtained at 1,000x magnification on 20 cells.

Morphological description of *Mantamonas sphyraenae* sp. nov. *Mantamonas sphyraenae* cells exhibited three general morphologies: 'balloons', which were typically ~5 µm long and ~3 µm wide, with a circularly curved anterior and a posterior end tapering to a point; 'kites', which were roughly diamond-shaped, about 3.5–4 µm long and wide; and 'mantas', which were 4–5 µm wide and ~3 µm long, having a broadly curved anterior end, a more tightly rounded right side, a bluntly rounded projection on the left side, and a posterior comprising either straight edges culminating in a point, two shallowly concave curves, or one of each. All three morphologies were plastic to some extent, although 'mantas' were noteworthy in that the left-side projection appeared rigid, and the curved right side frequently very plastic. Intermediates between the three morphologies were sometimes observed. In general, all cells in any given culture flask exhibited the same morphology, which often changed from one observation to the next, one to three weeks later. Exceptions to the prevalent morphology were almost always intermediate forms. We did not observe active transitions from one cell type to another, including to or from intermediate forms. Cells of all morphologies glided slowly and with constant speed, although occasionally stopping; the cell body frequently deformed when changing direction or colliding with other objects.

In all cases, a flagellum, 6–10 µm long, trailed behind the cell, always in a straight line except when the cell was turning, in which case it followed the cell's path. No movement of the flagellum was seen besides this. Under extremely favourable conditions, a second flagellum could be seen projecting from the anterior-left of 'manta' cells, at about a 45° angle. This second flagellum was invariably very thin, stiff, and 1–2 µm long. Very occasionally, we observed an additional protrusion, about the full width of a flagellum and about 1–2 µm long. This was always seen projecting from the posterior of the cell, immediately to the left of, and usually parallel to, the posterior flagellum. It appeared entirely static, and never appeared to change its length or orientation. Cysts were never observed at any stage of culture. Likewise, we never observed dividing cells.

***Mantamonas sphyraenae* nucleic acid extraction and genome/transcriptome sequencing.** To obtain nucleic acids, initially, five plates were inoculated with 500 µl from mature stock cultures. When these had reached high density (qualitatively determined), for each plate, the supernatant was discarded, cells were collected with the use of disposable cell scrapers, and the resulting 0.3–0.5 ml of concentrated cells were inoculated into 50 ml of fresh medium, which was then distributed into five new plates. This process was repeated, for a final count of 125 plates, for DNA extraction and 14 plates used for RNA extraction. For both preparations, cells were harvested with disposable cell scrapers and resuspended in sterile medium. The resuspension was prefiltered using 5.0-µm-pore polycarbonate filters, to remove bacterial flocs, and refiltered using 0.8-µm-pore filters, to remove individual bacteria.

For DNA extraction, filters were incubated in lysis buffer (50 mM Tris, 5 mM EDTA, 50 mM NaCl, pH 8), proteinase K (~300 µg/ml final concentration) and SDS (1% final concentration) for 1 hr on a rotator at 37°. The resulting solution was divided into two aliquots. From these, DNA was extracted in parallel using phenol/chloroform/isoamyl alcohol (25:24:1), extracted again using chloroform/isoamyl alcohol (24:1) and precipitated overnight in 95% EtOH at −20. The DNA was then pelletted in a centrifuge at 4°, washed with 80% EtOH, and resuspended in ddH_2O. The total yield was ~90 µg.

Long-read genomic sequences were obtained by using Single Molecule Real Time (SMRT) cell technology in a PacBio RSII system at the Cold Spring Harbor Laboratory. A total of 2,304,908 reads (18.7 Gbp) were acquired from 33 SMRT cells. Additional DNA samples were used to prepare two Illumina Nextera short-insert and mate-pair libraries following the manufacturer's protocols. The sequencing was done with the HiSeq 2500 System and a PE150 run option. A total of 62,929,978 read pairs (18.9 Gbp) and 53,901,870 read pairs (16.2 Gbp) were generated for the paired-end library and the mate pair library, respectively. For RNA extraction, the cell filters were incubated in TRI Reagent (Sigma-Aldrich) and RNA was isolated according to the manufacturer's instructions, using spin columns for elution. The total RNA sample was subjected to poly-A selection followed by Illumina TruSeq RNA library preparation and a total of 24,187,884 read pairs (7.3 Gbp) were sequenced using the Illumina HiSeq 2500 platform and a PE150 run option. All the genomic and transcriptomic Nextera library preparation and sequencing were conducted at the Weill Cornell's Genome Resources Core Facility.

***Mantamonas sphyraenae* genome assembly, gene prediction and ploidy analysis.** As the presence of co-cultured bacterial contamination in the sequencing data was expected, b oth the PacBio and Illumina reads were screened for contamination (see details in the technical validation section) and more than 60% of the original data identified as contaminant was discarded (see technical validation section). After this initial decontamination step, a total of 5.89 Gbp of long- read data was assembled using the Canu[5,12] and FALCON[6] pipelines.

The resulting genomic contigs from the Canu and FALCON approaches were then polished by aligning the screened PacBio reads to the draft genome using minimap2[13] and generating a consensus with Racon v1.3.1[14]. Subsequently, a second step of polishing was performed with the high quality Illumina reads by mapping them with bwa-0.7.15[15] and using Pilon v1.22[16] to correct for single base errors.

Additionally, MaSuRCA v3.2.6[7] was used to generate a hybrid assembly using the PacBio as well as the short-insert and mate-pair Illumina data that were retained after bacterial read filtering.

After these assembly efforts, any remaining bacterial contigs were identified by using a combination of homology searches and tetramer frequency-based binning (see details in the technical validation section). From the Canu assembly, a contig corresponding to mtDNA was identified and removed. Clean assemblies were then assessed based on their contiguity and completeness (Table 2) and the FALCON assembly was chosen for further analyses. Because of the specific parameter set utilized, our FALCON analysis did not assemble mtDNA due to its much higher sequence coverage compared to that for the nuclear DNA.

A custom library of repetitive elements was generated for the polished and cleaned nuclear genomic sequence by combining the results of RepeatModeler2[17] and Transposon-PSI (http://transposonpsi.sourceforge.net/) pipelines. The gathered repeat sequences from both analyse s were merged and clustered to generate a single consensus and refined repeat library that was further compared against the Dfam database[18] to classify the repetitive elements using RepeatModeler[17] refiner and classifier modules. Repetitive elements identified by this procedure were then masked out of the nuclear genome using RepeatMasker[17] before the prediction of protein-coding genes. Subsequently, the RNA-seq libraries were mapped against the genome sequence with HISAT-2[19] to generate spliced alignments, and BRAKER2[20] was employed to predict the nuclear protein coding genes integrating the extrinsic evidence from the RNA-Seq data.

Ploidy was inferred by assessing the distribution of allele frequencies at biallelic single nucleotide polymorphisms (SNPs) visually, and with modeling[21,22] using nQuire[22]. Briefly, the Nextera Illumina reads were mapped to the final genome assembly with Bowtie2 v2.3.5.1[23] and the resulting.bam file was used to calculate base frequencies for each biallelic site. These results were denoised using nQuire. The resulting frequencies were plotted in R version 3.3.3[24]. Finally, we ran the nQuire's Gaussian Mixture Model (GMM) command, which models the distribution of base frequencies at biallelic sites, and uses maximum likelihood to select the most plausible ploidy model (Table 3).

Isolation and microscopy of *Mantamonas vickermani* sp. nov. *Mantamonas vickermani* CRO19MAN was isolated from a sediment sample collected in July 2014 from the shallow marine lagoon Malo jezero (42°47′05.9"N 17°21′01.3"E) on the island of Mljet (Croatia, Mediterranean Sea). The sample was taken from the upper layer of the sediments at the shore of the lagoon with a sterile 15 ml Falcon tube at a depth of 10 cm below the water surface and stored at −20 °C. In September 2019, a small amount of sediment was inoculated in a Petri dish with 5 ml of sterile seawater supplemented with 1% YT medium (100 mg yeast extract and 200 mg tryptone in 100 ml distilled water, as in the protocol from the National Institute for Environmental Studies [NIES], Japan). After observation of some mantamonad cells, serial dilution was performed in a multiwell culture plate to further enrich the culture. We transferred 250 µl of culture to a well with 1 ml of fresh 1% YT seawater medium and then retransferred the same volume to a new well, repeating the process 5 times for a total of 24 wells. Single mantamonad cells were then isolated from one of the enriched cultures with an Eppendorf PatchManNP2 micromanipulator using a 65 µm VacuTip microcapillary (Eppendorf) and a Leica Dlll3000 B inverted microscope. This cell was inoculated into 1 ml of growth medium and after 48 hr incubation we confirmed an established monoculture of *M. vickermani* CRO19MAN.

Optical microscopy observations were performed with a Zeiss Axioplan 2 microscope equipped with oil-immersion differential interference contrast (DIC) and phase contrast objectives. Images were acquired with an AxiocamMR camera using the Zeiss AxioVision 4.8.2 SP1 suite. Videos were recorded using a Sony α9 digital camera. Morphometric data were obtained at 1,000x final magnification on 20 cells. Images were captured at multiple focal planes in order to visualise different cell parts. Measurements of flagella pertain to the visible parts, i.e., the posterior flagellar length is measured beginning from the point at which it emerges from underneath the cell at the body's posterior end.

Morphological description of *Mantamonas vickermani* sp. nov. *Mantamonas vickermani* cells are ~3 µm wide and ~3.5 µm long; thus noticeably smaller than those of *Mantamonas plastica* (~5 µm wide and ~5 µm long) (Glücksman *et al.*[1]) (Fig. 1g–l). Like *M. plastica, M. vickermani* also has a strongly flattened and plastic morphology. However, the characteristic blunt projection on the left-hand side of the cell observed in *M. plastica* is less conspicuous in *M. vickermani*, and not always observed in cells possessing an overall spherical to oval morphology (Fig. 1). The anterior flagellum of *M. vickermani* is ~2 µm long, rigid in all of its length, and vibrates with a small amplitude; its posterior flagellum is ~7 µm long and considerably thicker than the anterior one, having a very small acroneme that when seen is never longer than 1–2 µm. Both flagella are also shorter than those reported for *M. plastica* (~3 µm anterior and ~10 µm posterior).

Mantamonas vickermani glides in a smooth and continuous manner on the substrate with a similar speed and turning behavior to that observed for *M. plastica* (Glücksman *et al.*[1]; AAH, pers. obs.) (Movie 1 and Movie 2). As with *M. plastica, M. vickermani* is a bacterivore with a voracious appetite, engulfing bacteria at a high rate. Interestingly, and in contrast with Glücksman *et al.*[1], we did observe one cell possessing two posterior flagella, which strongly suggests that it was undergoing cellular division (Fig. 1).

***Mantamonas vickermani* RNA purification and transcriptome sequencing.** This new strain was grown for a week in 75 cm^2 cell culture flasks with ~10 ml of medium. Fully grown cultures were collected by gently scratching the bottom of the flasks with a cell scraper to resuspend the gliding flagellates and pooled in 50 ml Falcon tubes to be centrifuged at 10 °C for 15 minutes at 15,000 g. Total RNA was extracted from cell pellets with the RNeasy mini Kit (Qiagen) following the manufacturer protocol. Two cDNA Illumina libraries were constructed after polyA mRNA selection, and these were sequenced using the paired-end (2 × 125 bp) method with Illumina HiSeq 2500 Chemistry v4 (Eurofins Genomics, Germany).

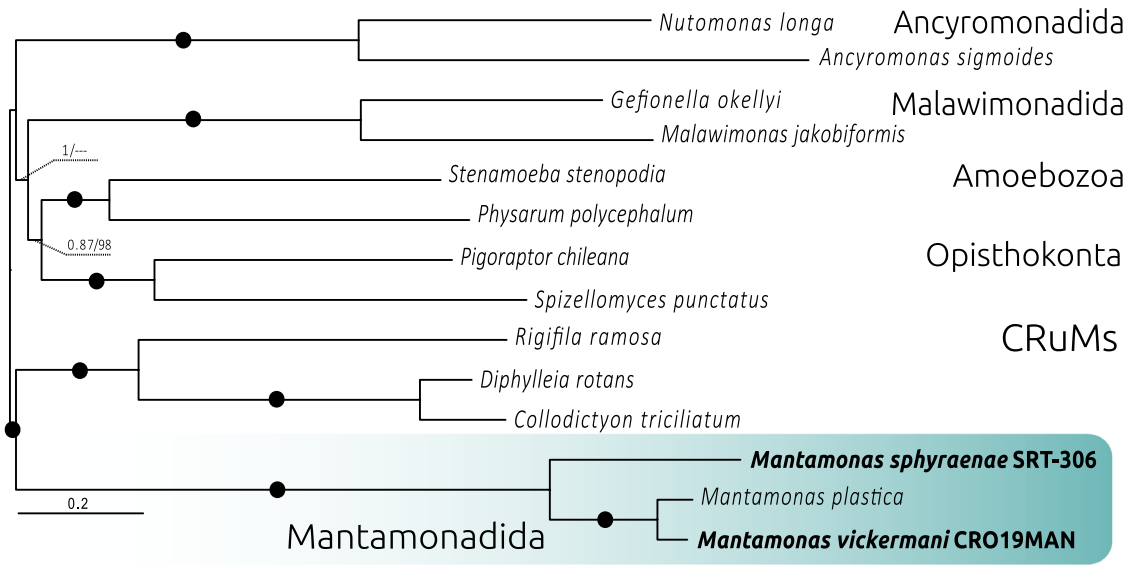

Fig. 4 Phylogenomic analysis of CRuMs clade. Bayesian inference (BI) phylogeny based on 182 conserved proteins from Lax *et al.*[3]. The tree was obtained using 62,088 amino acid positions with the CAT-GTR model. Statistical support at branches was also estimated using maximum likelihood (ML) under the LG+C60+F+R4 model with the PMSF approximation. Numbers at branches indicate BI posterior probabilities and ML bootstrap values, respectively; bootstrap values <50% are indicated by dashes. Branches with support values higher than or equal to 0.99 BI posterior probability and 95% ML bootstrap value are indicated by black dots. The tree was rooted between CRuMs and everything else.

Transcriptomes assembly and proteome prediction. The transcriptomic sequence of *M vickermani* and *M. sphyraenae* were assembled *de novo* using Spades v3.13.1[25] with the *rna* mode and default parameters specified. Transcripts were then screened to identify remaining contaminants using the Blobtools2[26] pipeline and homology searches against a custom database (see technical validation section). Predicted proteins were obtained from the clean transcripts using Transdecoder v2 (http:transdecoder.github.io) alowing for a single prediction by transcript (–single-bes-only option) and using the universat genetic code. Subsequently, CD-HIT[27] clustering was employed (with a threshold of $>=90\%$ of identity) to produce a non-redundant data set of proteins for each of the trancriptomes, and to eliminate falsely duplicated proteins stemming from alternatively spliced transcripts.

Phylogenomic analyses. The dataset of 351 conserved protein markers from Lax *et al.*[3] was updated by BLASTP searches[28] against the inferred proteomes of representatives of other eukaryotic lineages, including the proteomic data for our two new mantamonad strains. Each protein marker was aligned with MAFFT v.7[29] and trimmed using TrimAl[30] with the -automated1 option. Alignments were manually inspected and edited with AliView[31] and Geneious v6.06[32]. Single-protein trees were reconstructed with IQ-TREE v1.6.11[33] under the corresponding best-fitting model as defined by ModelFinder[34] implemented in IQ-TREE[33]. Each single-protein tree was manually inspected to discard contaminants and possible cases of horizontal gene transfer or hidden paralogy. At the end of this curation process, we kept a final taxon sampling of 14 species, including members of Ancyromonadida, Malawimonadida, Opisthokonta, and CRuMs (concatenated alignment and supplementary trees are available at Figshare[35]), and 182 protein markers that were present in all mantamonad species (with at least 80% of markers identified in each taxon). All proteins were realigned, trimmed as previously described, and concatenated, creating a final supermatrix with 62,088 amino acids.

A Bayesian inference tree was reconstructed using PhyloBayes-MPI v1.5a[36] under the CAT-GTR model[37], with two MCMC chains, and run for 10,000 generations, saving one of every 10 trees. Analyses were stopped once convergence thresholds were reached (i.e. maximum discrepancy <0.1 and minimum effective size >100, calculated using bpcomp). Consensus trees were constructed after a burn-in of 25%. Maximum likelihood (ML) analyses were done with IQ-TREE v1.6.11[33], first by calculating the ML tree under the LG+F+R4 model, which was used as guide tree for the PMSF approximation[38] run under the LG+C60+F+R4 model.

Consistent with previous studies, our maximum likelihood (ML) and Bayesian inference (BI) phylogenetic trees recovered the monophyly of CRuMs with high BI posterior probability (0.99) and ML bootstrap support (95%), although it is worth noticing that the outgroup is highly reduced since resolving the position of CRuMs in the tree of eukaryotes is outside the scope of this paper. The monophyly of *Mantamonas* received full support from both methods. We found Mantamonas sphyraenae to be sister to a maximally-supported clade containing *M. plastica* and *M. vickermani* (Fig. 4).

CRuMs orthologue analysis and protein functional annotation. Orthologous gene families were identified among the predicted proteomes of *Mantamonas sphyraenae*, *Mantamonas vickermani* and the publicly available proteomes of *Mantamonas plastica, Diphylleia rotans* and *Rigifila ramosa* as obtained from the EukProt

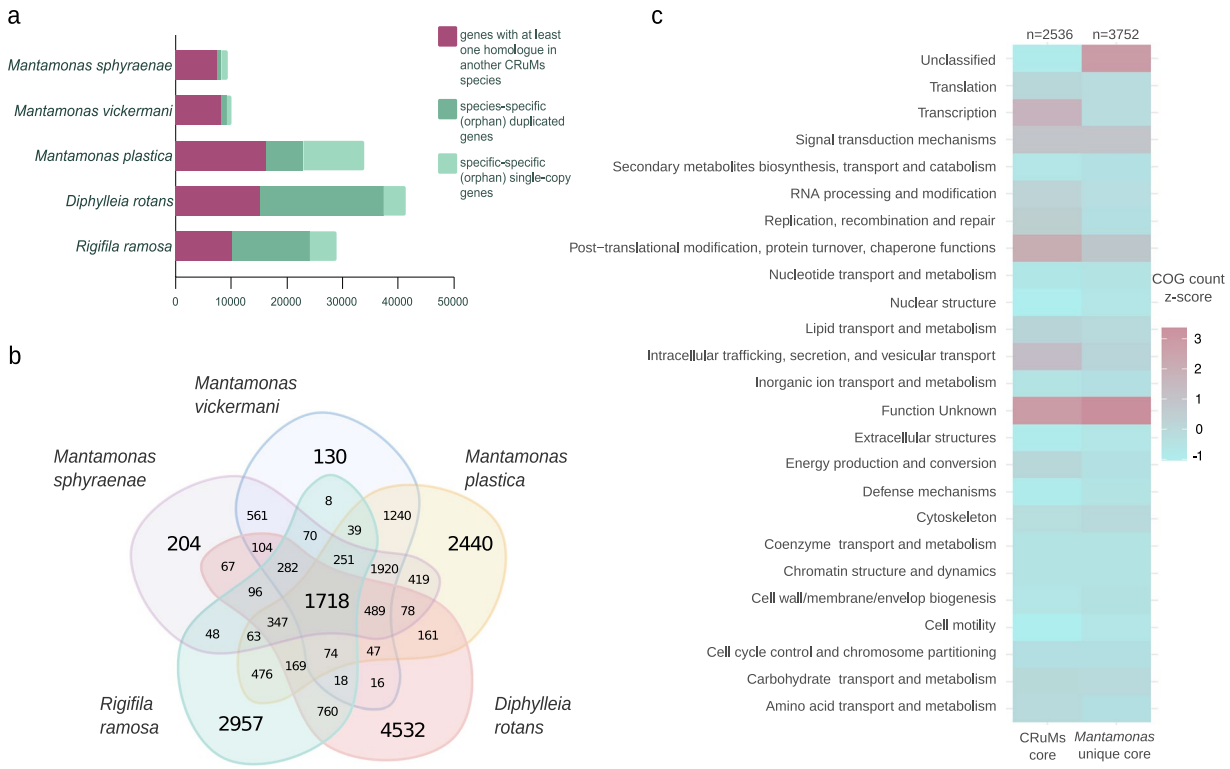

Fig. 5 Orthology analysis across the CRuMs supergroup. (**a**) Distribution of coding sequences shared among CRuMs representatives (magenta) or that are species-specific in one or several copies (dark and light green, respectively). Note that these numbers do not represent genes but open reading frames identified in assembled transcripts, except for *M. sphyraenae*. (**b**) Number of orthogroups shared among compared CRuMs species. (**c**) COG functional categories associated with orthogroups shared among all CRuMs, and those associated with orthogroups shared across *Mantamonas* species but absent in other CRuMs taxa. COG counts were scaled by column using z-score standardization.

v3 database[39] using OrthoFinder v2.5.4[40]. For this, we used DIAMOND[41] ("ultra-sensitive" mode, and query cover >= 50%), an inflation value of 1.5, and the MCL clustering algorithm (Fig. 5a).

Then, the predicted proteomes of *M. sphyraenae, M. vickermanii*, *M. plastica*, *D. rotans*, and *R. ramosa* were functionally annotated with the EggNOG-mapper pipeline[9], using DIAMOND ultra-sensitive mode and all domains of life as the target space. During this process, individual sequences composing the CRuMs orthogroups generated by OrthoFinder were assigned a COG functional category. This information was summarized at the orthogroup level by assigning to each orthogroup a single COG category corresponding to the most frequent annotation of its individual sequences, provided that it represented at least 50% of the sequences within the orthogroup.

A total of 1,718 orthogroups were found to be conserved among all CRuMs taxa (Fig. 5b), while 4,378 were identified as shared between the three *Mantamonas* species, representing the minimal core proteome of the genus *Mantamonas* as currently known, among which 2,161 orthogroups are not found in the other two CRuMs lineages. Our species also display a smaller number of unique proteins than the publically available proteomes likely due to the methodological strategy that we employed to assemble the transcriptomes and infer open reading frames that reduces the number of short and incomplete ORFs and sioforms when compared with the proteomes derived from the other CRuMs transcriptomes. However, beyond the absolute numbers of predicted coding sequences, the comparison between all these proteomes gives us an indication about the degree of the diversity of gene content in each of our two *Mantamonas* species.

Most of the proteins conserved among the CRuMs taxa (99.6%) were found to have an ortholog in the EggNOG database and to belong to at least one Cluster of Orthologous Groups (COG)[42,43] functional category, where the most highly represented were "Function unknown" and "Post-translational modificationand "Intracellular trafficking" (Fig. 5c). By contrast, a substantial amount of orthogroups conserved among mantamonads (12%), but absent in other CRuMs lineages, could not be assigned to any cluster in the EggNOG database. In addition, most orthogroups conserved in mantamonads but absent in other CRuMs that could be connected to an existing EggNOG cluster were annotated as "Function unknown" (Fig. 5c). Altogether, this large number of *Mantamonas*-specific genes of unknown function suggests that many genetic innovations occurred at the origin of this group.

Species and strain name	Type	Platform	Read type	SRA accession number
M. sphyraenae STR306	DNA	PacBio RS II	Single molecule	SRR21818797
M. sphyraenae STR306	DNA	Illumina HiSeq 2500	Paired	SRR21818798
M. sphyraenae STR306	DNA	Illumina HiSeq 2500	Mate pair	SRR22188164
M. sphyraenae STR306	RNA	Illumina HiSeq 2500	Paired	SRR21818794
M. vickermani CRO19MAN	RNA	Illumina HiSeq 2500	Paired	SRR21818793

Table 4. Summary of sequencing data records.

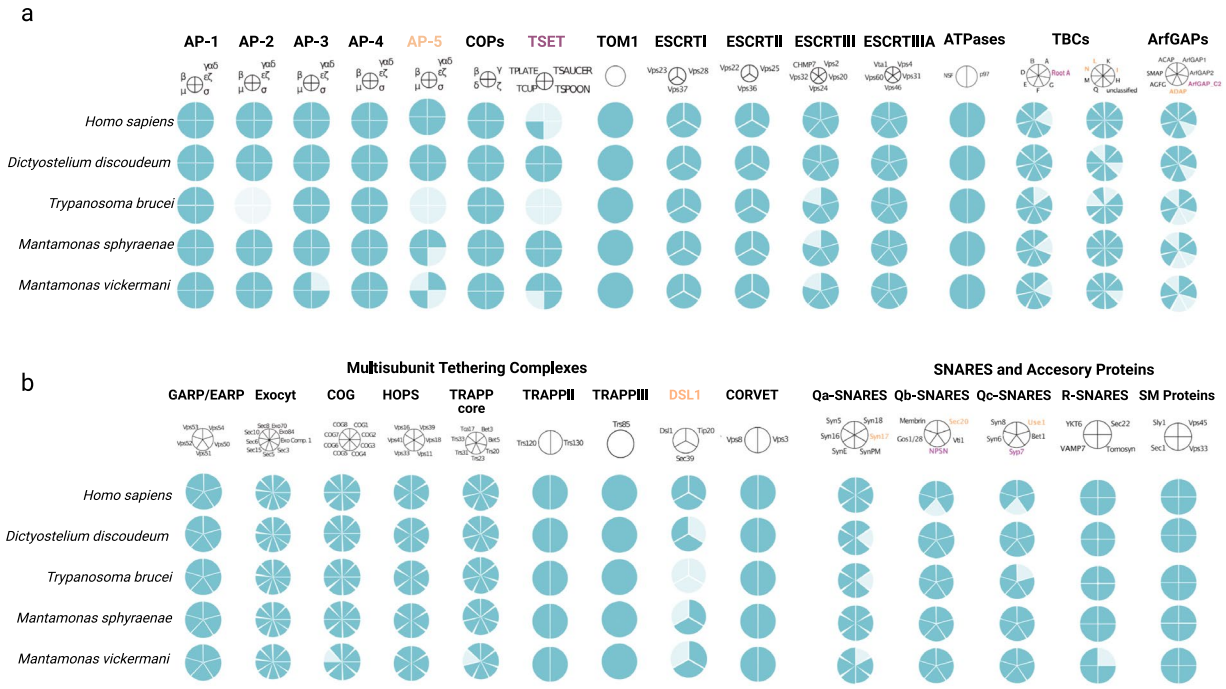

Fig. 6 Distribution of proteins associated with the membrane trafficking system in new *Mantamonas* species and other model organisms. (**a**) Selected vesicle formation machinery; (**b**) Selected vesicle fusion machinery. Names of proteins with jotnarlogs are in purple; those with patchy distribution are in orange.

Analysis of the conservation of the membrane-trafficking system complement. To assess the complement of the membrane trafficking system encoded in our *Mantamonas* genome and transcriptome datasets, we performed homologous searches of a selection of protein query sequences from the genomes of *Homo sapiens* (GCF_000001405.40), *Dictyostelium discoideum* (GCF_000004695.1), *Arabidopsis thaliana* (GCF_000001735.4) and *Trypanosoma brucei* (GCF_000002445.2) available at the GenBank of the NCBI (National Center for Biotechnology Information) database. These proteins included components of the machinery for vesicle formation (HTAC-derived coats, ESCRTs, and ArfGAPs) and vesicle fusion (SNAREs and SM proteins, TBC-Rab GAPs, and Multi-subunit tethering complexes)[10].

BLASTP and TBLASTN were used to search the predicted proteomes and nucleotide coding sequences, respectively, of *M. sphyraenae* and *M. vickermani*. The HMMER3 package was used to find more divergent protein sequences using the hmmsearch tool[44]. In cases in which only TBLASTN hits were retrieved, these were translated using Exonerate[45]. Potential orthologs (i.e., hits with an E-value below 0.05) were further analyzed by the Reciprocal Best Hit (RBH) approach, using the *Mantamonas* candidate orthologs as queries against the *H. sapiens*, *D. discoideum* and *A. thaliana* proteomes. If the best hit was the protein of interest and had an E-value two orders of magnitude lower than the next non-orthologous hit, this was considered as orthology validation. Forward and reverse searches were performed using the AMOEBAE tool[46].

We detected most proteins of the membrane-trafficking system in the two new *Mantamonas* species, making it one of the most complete known protein complements for this system. Notably, when compared to representatives of well-characterized model organisms from other supergroups (Fig. 6). *Mantamonas* encodes some rarely retained proteins, such as the AP5 complex[47] and syntaxin 17[48]. We also identified several jotnarlogs (Fig. 6), including a near-complete TSET complex, and the SNAREs NPSN and Syp7[35].

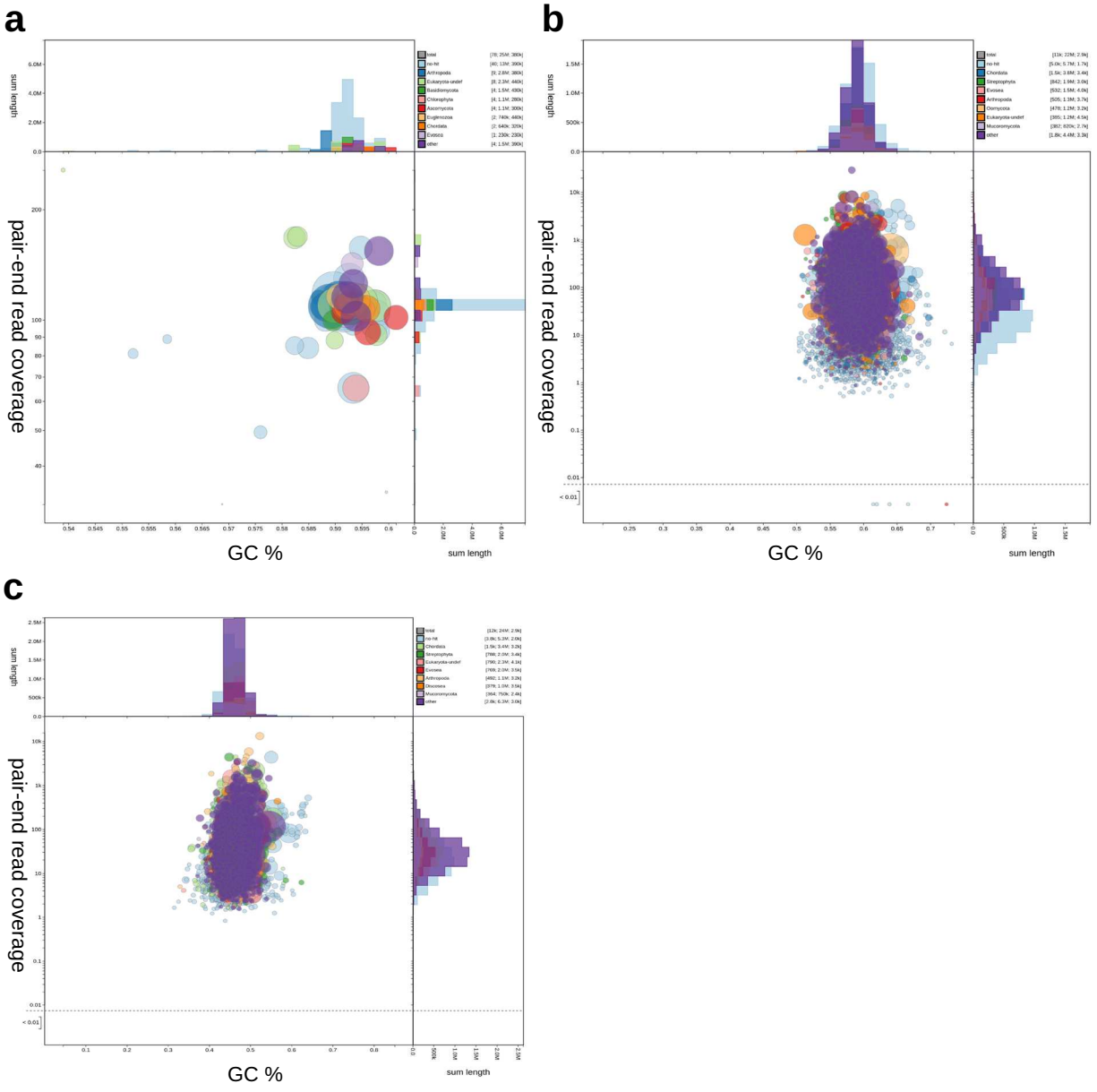

Fig. 7 Blob plot of read coverage against GC proportion in genome and transcriptomic contigs. (**a**) *M. sphyraenae* genomic sequences. (**b**) *M. sphyraenae* transcripts. (**c**) *M. vickermani* transcripts. Records are coloured according to their similarity to different phyla. Circles are sized in proportion to records cumulative length. The assembly has been filtered to exclude records whose taxonomic assignment matches "Bacteria". Histograms show the distribution of record length sums along each axis.

Data Records

The read data associated with the nuclear genome and transcriptomic datasets of *Mantamonas sphyraenae* and the transcriptome of *Mantamonas vickermani* have been submitted to the NCBI SRA database[49] (Table 4).

The Transcriptome Shotgun Assemblies have been deposited at DDBJ/EMBL/GenBank under the accessions GKLA00000000 and GKKZ00000000 for *M. vickermani* and *M. sphyraenae* respectively. The final nuclear genome assembly of *Mantamonas sphyraenae* has been deposited at GenBank under the accession GCA_026936335.1[50]. The versions described in this paper are the first versions. The prediction of protein-coding genes from the genome and transcriptome assemblies of *Mantamonas sphyraenae*, as well as from the transcriptome assembly of *M. vickermani* are available at Figshare[35].

Phylogenomic analysis alignments and trees, and membrane-trafficking predicted proteins table can be found on Figshare[35].

Technical Validation

Quality assessment of sequencing datasets. All Illumina paired-end raw reads used for genome polishing were quality-checked with FastQC v0.11.8[51] and trimmed using TRIMMOMATIC[52] to retain only reads with maximum quality scores. PacBio reads resulted in an N50 of 11,048 bp and an average coverage of 106x after filtering out the identified contaminant sequences (see below).

Identification and filtering of contaminant sequences. Mantamonads grow in non-axenic cultures with co-cultured prokaryotes on which they feed. Therefore, various methods were employed to ensure the correct identification and filtering of contaminant sequences in the genomic and transcriptomic datasets of *M. sphyraenae* and *M. vickermani*.

For the genomic dataset of *M. sphyraenae*, we first identified the main bacterial contaminants from the initial genome assemblies[53]. In addition, we established a custom database consisting of contigs assembled from Illumina sequencing data from bacteria only enrichment cultures derived from the lab's several xenic protist cultures. These were used to screen PacBio reads using BLASR v5.1[54] The Illumina reads were screened similarly using Bowtie2 v2.3.5.1[23]. Only Illumina reads in which neither pair aligned to the bacterial database were retained for further assembly recovering 57% and 49% of the pair-end and mate pair reads from the total libraries respectively.

After genome assembly using the filtered reads, remaining contaminant contigs were identified by using MyCC v1[55], which bins contigs based on their tetranucleotide frequencies and coverage. Clusters were formed using the affinity propagation (AP) algorithm and visualized in a 2-dimensional Barnes-Hut-SNE plot. BLASTN searches using default parameters were conducted against the 'nt' database from the NCBI to taxonomically classify the bins. Contigs were identified as contaminants if they contained no hits other than to prokaryotes, and if they were clustered away from the main eukaryotic bin. Finally mitochondrial sequences were screened out from the short and long read libraries of *M. sphyraenae* by mapping them against the mitochondrial genome using bwa-0.7.15[15] and minimap2[13] respectively.

The assembled transcriptomes of *M. sphyraenae and M. vickermani were* decontaminated with the Blobtools2 pipeline[26]. Briefly, this approach helps to identify contaminant sequences based on their biases in coverage and GC content, as well as on a taxonomic classification established by DIAMOND searches[41] against the 'nt' and Uniprot databases[56]. In addition, a second cleaning step was done by performing DIAMOND searches against a database containing all the proteins of the prokaryotic Genome Taxonomy Database (GTDB)[57] and the eukaryotic-representative EukProt v3 database[39]. A protein was considered as a probable contaminant and excluded from further analyses if its best hit corresponded to any protein from GTDB, with strict cutoffs of identity $\geq$50% and query coverage $\geq$50%. Finally, a blobplot was generated for the final genomic and transcriptomic contigs of *M. sphyraenae* and *M. vickermani*, respectively, to verify the absence of contaminant sequences (Fig. 7).

Completeness analysis. To assess the completeness of the decontaminated genome and transcriptome datasets, we employed the BUSCO v5.3.2 pipeline. We identified the percentage of near-universal single copy orthologs of the eukaryote_odb10 database[12] on the predicted proteomes of *M. sphyraenae* and *M. vickermani*, as well as those of other species belonging to the CRuMs supergroup available in the EukProt v3[39] database for comparison purposes (Fig. 2). Moreover, the comparison of the transcriptomic dataset and the genomic dataset of Mantamonas sphyraenae revealed that 96% of the proteins predicted in the transcriptome share similarity with the proteins derived from the genome (80% of these being identical) and 271 proteins were found to be present uniquely in the transcriptome. Adittionally, the mapping coverage from the clean transcriptomic reads to the genome sequence was of 97.38%, suggesting a near complete representation of the gene space in the genome-predicted proteins.

Data usage notes. Formal species descriptions

All taxonomic descriptions in this work were approved by all authors.

Eukarya: 'CRuMs'

Order Mantamonadida Cavalier-Smith 2011

Family Mantamonadidae Cavalier-Smith 2011

Genus *Mantamonas* Cavalier-Smith and Glücksman 2011

Mantamonas sphyraenae sp. nov. Description: Cells with varying morphologies: shaped as manta rays (as for genus in Glücksman *et al.*[1]), ~3 μm long and ~5 μm wide; diamonds, 4 ± 1 μm in both dimensions; or rounded anteriorly and tapering posteriorly, ~5 μm long and ~3 μm wide. Anterior flagellum stiff, 0.5–1.0 μm long. Other characters as for genus.

Type culture: SRT306

Type locality: Surface of barracuda caught in lagoon on Iriomote Island, Taketomi, Okinawa Prefecture, Japan (24° 23' 36.762″ N, 123° 45' 22.572″ E).

Isolator: Takashi Shiratori

Etymology: From Sphyraena, the genus name for barracuda, the fish from which the type strain was obtained.

Gene sequence: The nuclear genome and transcriptomic read sequencing data from *Mantamonas sphyraenae* (strain SRT306) were deposited in GenBank under BioProject accession number PRJNA886733.

Mantamonas vickermani sp. nov. Description: Cell size ~3 μm (2.5–4.3 μm) long, ~3.5 μm (3.0–4.0 μm) wide; cells almost perfectly round, although in some cases possessing a small projection to the left side of the cell;

without pseudopodia; anterior flagellum usually ≤2 μm long (1.2–2.7 μm), held forwards and to left ~40–50° to longitudinal axis, does not beat except for slight terminal vibration; posterior flagellum ~7 μm long (6–8.9 μm), conspicuous and sometimes acronematic. Other characters as for genus.

Type culture: CRO19MAN

Type locality: Specimen isolated from the sediments of the marine lake Malo jezero in the island of Mljet, Croatia.

Isolator: Luis Javier Galindo.

Etymology: The name vickermani honors work on heterotrophic protists by Keith Vickerman.

Gene sequence. The full transcriptome read data from *Mantamonas vickermani* (strain CRO19MAN) were deposited in GenBank under BioProject accession number PRJNA886733.

Code availability

All the employed software as well as their versions and parameters were described in the method section. If no parameters were specified, default settings were employed. Data visualization plots were generated using R v4.1.2 (https://cran.r-project.org/, R development core team) and https://bioinformatics.psb.ugent.be/webtools/Venn/.

Received: 19 December 2022; Accepted: 18 August 2023;
Published online: 09 September 2023

References

1. Glücksman, E. *et al*. The novel marine gliding zooflagellate genus Mantamonas (Mantamonadida ord. n.: Apusozoa). *Protist* **162**, 207–221 (2011).
2. Brown, M. W. *et al*. Phylogenomics Places Orphan Protistan Lineages in a Novel Eukaryotic Super-Group. *Genome Biol. Evol.* **10**, 427–433 (2018).
3. Lax, G. *et al*. Hemimastigophora is a novel supra-kingdom-level lineage of eukaryotes. *Nature* **564**, 410–414 (2018).
4. Burki, F., Roger, A. J., Brown, M. W. & Simpson, A. G. B. The New Tree of Eukaryotes. *Trends in Ecology & Evolution* (2020).
5. Koren, S. *et al*. Canu: scalable and accurate long-read assembly via adaptive -mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
6. Chin, C.-S. *et al*. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
7. Zimin, A. V. *et al*. Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Research* **27**, 787–792 (2017).
8. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution* (2021).
9. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution* **38**, 5825–5829 (2021).
10. More, K., Klinger, C. M., Barlow, L. D. & Dacks, J. B. Evolution and Natural History of Membrane Trafficking in Eukaryotes. *Curr. Biol.* **30**, R553–R564 (2020).
11. Okaichi, T. Collection and mass culture. *Yuudoku-Plankton-Hassei, Sayou-Kikou, Doku-Seibun: Toxic Phytoplankton-Occurrence, Made of Action, and Toxins* 23–34 (1982).
12. Kriventseva, E. V. *et al*. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Research* (2019).
13. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
14. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
15. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
16. Walker, B. J. *et al*. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).
17. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* **117**, 9451–9457 (2020).
18. Storer, J., Hubley, R., Rosen, J., Wheeler, T. J. & Smit, A. F. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob. DNA* **12**, 2 (2021).
19. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
20. Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom Bioinform* **3**, lqaa108 (2021).
21. Gompert, Z. & Mock, K. E. Detection of individual ploidy levels with genotyping-by-sequencing (GBS) analysis. *Molecular Ecology Resources* (2017).
22. Weiß, C. L., Pais, M., Cano, L. M., Kamoun, S. & Burbano, H. A. nQuire: a statistical framework for ploidy estimation using next generation sequencing. *BMC Bioinformatics* (2018).
23. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
24. R Core Team, R. R: A language and environment for statistical computing. https://www.R-project.org/ (2013).
25. Bushmanova, E., Antipov, D., Lapidus, A. & Prjibelski, A. D. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. *GigaScience* (2019).
26. Challis, R., Richards, E., Rajan, J., Cochrane, G. and Blaxter, M. BlobToolKit–interactive quality assessment of genome assemblies. *G3: Genes, Genomes, Genetics*, **10**(4), pp.1361–1374 (202).
27. Li, W., Jaroszewski, L. & Godzik, A. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* (2001).
28. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* (1990).
29. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
30. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
31. Larsson, A. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30**, 3276–3278 (2014).
32. Kearse, M. *et al*. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649 (2012).

33. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* (2015).
34. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
35. Eme, L., Blaz, J., Galindo, L. & Torruella, G. One high-quality genome and two transcriptome datasets for two new species of Mantamonas, a deep-branching eukaryote clade, *Figshare*, https://doi.org/10.6084/M9.FIGSHARE.22802432 (2023).
36. Lartillot, N., Lepage, T. & Blanquart, S. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25**, 2286–2288 (2009).
37. Lartillot, N. & Philippe, H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* **21**, 1095–1109 (2004).
38. Wang, H.-C., Minh, B. Q., Susko, E. & Roger, A. J. Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. *Syst. Biol.* **67**, 216–235 (2018).
39. Richter, D. J. *et al*. EukProt: a database of genome-scale predicted proteins across the diversity of eukaryotes. *Peer Community Journal*, 2 (2022)
40. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
41. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).
42. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278**, 631–637 (1997).
43. Galperin, M. Y., Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.* **43**, D261–9 (2015).
44. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–37 (2011).
45. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
46. Barlow, L. D. *et al*. Comparative genomics for evolutionary cell biology using AMOEBAE: Understanding the Golgi and beyond. In Golgi: Methods and Protocols (pp. 431-452). New York, NY: Springer US (2022).
47. Hirst, J. *et al*. Correction: The Fifth Adaptor Protein Complex. *PLoS Biol.* **10** (2011).
48. Arasaki, K. *et al*. A role for the ancient SNARE syntaxin 17 in regulating mitochondrial division. *Dev. Cell* **32**, 304–317 (2015).
49. Eme, L., Blaz, J. & Kim, E. *NCBI Sequence Read Archive*. https://identifiers.org/ncbi/insdc.sra:SRP401184.
50. Blaz, J., Galindo, L., Torruella, G. & Eme, L. Mantamonas sp. genome assembly ASM2693633v1. *Genbank* https://identifiers.org/insdc.gca:GCA_026936335.1 (2023).
51. FastQC. *FastQC: a quality control tool for high throughput sequence data.*, (2016).
52. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
53. Aponte, A. *et al*. The Bacterial Diversity Lurking in Protist Cell Cultures. *novi* **2021**, 1–14 (2021).
54. Chaisson, M. J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**, 238 (2012).
55. Lin, H.-H. & Liao, Y.-C. Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. *Sci. Rep.* **6**, 24175 (2016).
56. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).
57. Parks, D. H. *et al*. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* **50**, D785–D794 (2022).

Acknowledgements

The authors thank Drs. J.Z. Xiang, D. Xu, and H. Shang at the Weill Cornell's Genome Resources Core Facility for their assistance with Illumina sequencing. We also thank Dr. S. Goodwin in the NGS sequencing core at Cold Spring Harbor Laboratory for her help with PacBio sequencing, Drs. J.A. Burns and A.A. Pittis for their initial data analysis efforts and K. Lukacs for her help with maintaining the *M. sphyraenae* cultures. This work was funded by the Simons Foundation Grant awards to EK (SF-382790 & SF-876199). This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (ERC Starting Grant No 803151 to L.E., and ERC Advanced Grants No 322669 and 787904 to P.L.-G. and D.M., respectively). L.J.G. was funded by the Horizon 2020 research and innovation programme under the European Marie Skłodowska-Curie Individual Fellowship H2020-MSCA-IF-2020 (grant agreement no. 101022101 - FungEye). GT was supported by the 2019 BP 00208 Beatriu de Pinos-3 Postdoctoral Program (BP3; 801370). Work in the Dacks lab is supported by Discovery Grants from the Natural Sciences and Engineering Research Council of Canada (RES0043758, RES0046091).

Author contributions

E.K., L.E. conceived this project; T.S., K.I., L.J.G. isolated the *Mantamonas* strains; A.A.H., L.J.G. and G.T. maintained the cultures; A.A.H., A.Y. and L.J.G collected nucleic acid and generated image data; E.K., L.E., D.M., P.L.G. and J.B.D. designed various analyses. J.B., A.A.H., A.Y., L.A.T., A.F., G.T., L.J.G., A.T., H.K., S.W., A.N., J.B.D., E.K. and L.E. analyzed and interpreted the results; J.B., A.A.H., A.Y., S.W., H.K., J.B.D., E.K. and L.E. drafted the manuscript. All authors reviewed and approved the manuscript. E.K., L.E., P.L.G. D.M. and J.B.D. acquired funding.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to E.K. or L.E.

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

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

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

© The Author(s) 2023

5. The genome of *Ancyromonas sigmoides* the type species of Ancyromonadida

Context and results summary

Ancyrmonas sigmoides is a gliding biflagellate that was observed and described for the first time in 1882 (Kent 1882). Nowadays, *A. sigmoides* has been established as the type species of Ancyromonadida and while ultrastructural and molecular data exist for these species since the last decades (Heiss, Walker, and Simpson 2011; Atkins, McArthur, and Teske 2000), their biology and phylogenetic position remain enigmatic. In particular, ancyromonads have been considered an orphan lineage for long time because its lack of phylogenetic affinity to any other eukaryotic supergroup (Atkins, McArthur, and Teske 2000; Cavalier-Smith and Chao 2003; Heiss, Walker, and Simpson 2010), therefore probably ancient and key to resolve the deep structure of the eukaryotic tree.

In collaboration with Eunsoo Kim's lab, we combined PacBio and Illumina sequencing to generate a highly contiguous genome for this enigmatic orphan protist. This genome has an important proportion of repetitive elements and protein coding genes seemingly restricted to this species. The genome sequence assembled and described in the following short manuscript is the first genomic sequence for this major lineage. We have planned to submit this manuscript soon as a genome report to the journal *Genome Biology and Evolution*.

The nuclear genome of *Ancyromonas sigmoides*, the type species of the diverse and deeply divergent lineage of flagellates.

Jazmín Blaz, Aaron Heiss, Naoji Yubuki, John Burns, Maria Ciobanu, Luis J. Galindo, Guifré Torruella, Purificación López-García, David Moreira, Eunsoo Kim & Laura Eme

Abstract

Ancyromonads, a diverse and globally distributed group of heterotrophic flagellates, hold a pivotal deep position in the eukaryotic tree of life. Despite their evolutionary significance, no genomic data is available for this clade. Here, we present a contiguous assembly of the nuclear genome of *Ancyromonas sigmoides,* spanning 39 Mbp generated throughout the hybrid assembly of Illumina and PacBio reads. Remarkably, approximately 29% of the genome consists of repeats, primarily comprising unknown interspersed repeat families. Of the 11,138 protein-coding genes identified, only 56% exhibited detectable homology with genes found in other eukaryotes when compared to the non-redundant database. Intriguingly, 1,212 genes in *A. sigmoides* shared a close evolutionary relationship with prokaryotic and viral genes, underscoring a history of lateral gene transfer that likely contributed to the acquisition of novel functions in this species. This genome represents a critical foundation for future investigations aiming to unravel the molecular basis of the adaptations that enable ancyromonads to thrive in their environments and studies aiming to better understand the deep eukaryotic evolution.

Significance

The inadequate representation of various cultures and genomic data related to numerous microbial eukaryotic clades has significantly hindered our ability to comprehensively explore the diversity of key cellular and molecular traits across eukaryotes. Additionally, it has hampered our efforts to reconstruct phylogenetic relationships at the base of the eukaryotic tree of life. With the genome sequence of Ancyromonas sigmoides, a major orphan lineage, this gap in knowledge is substantially narrowed. This genomic data offers a new and unique perspective on the innovations within this lineage. Given its evolutionary distance from all existing eukaryotic supergroups, Ancyromonas sigmoides stands as a crucial reference for any comparative genomic analysis aiming to unravel the deep evolutionary history of eukaryotes.

Introduction

In recent years our view of the diversity and evolution of the eukaryotic tree of life (eToL) has been transformed by the discovery of diverse major lineages and the reevaluation of their phylogenetic relationships (Burki et al. 2020; Lax et al. 2018; Tikhonenkov et al. 2022). Although the position of the root and the precise order of divergence between these lineages remain debated, several phylogenomic analyses (Brown et al. 2018) have shed light on the deeply branching lineages, such as the Ancyromonadida clade, as emerging models to resolve this phylogeny and understand the deep evolution of the eukaryotic domain.

Ancyromonas sigmoides, the type species of the clade Ancyromonadida (Atkins et al. 2000), was described for the first time by Saville Kent in the 19th century (Saville-Kent 1882). This small biflagellate displays bean-like shapes and glides on their long posterior flagellum while flicking vigorously (Heiss et al. 2011).

The first molecular analyses of *Ancyromonas* revealed a large evolutionary distance to other eukaryotes (Atkins et al. 2000; Cavalier-Smith & Chao 2003; Heiss et al. 2010) and numerous phylogenetic studies have failed to robustly resolve their phylogenetic affiliation to any other eukaryotic crown lineage (Paps et al. 2013; Torruella et al. 2017; Brown et al. 2018).

Ancyromonads are globally distributed in benthic sediments from marine to freshwater ecosystems as well as soil (Yubuki et al. 2023; Tikhonenkov et al. 2006) and recent studies have demonstrated that this group encompasses a high diversity of cryptic species (Yubuki et al. 2023). As cosmopolitan phagotrophs ancyromonads are playing important roles in the nutrient cycling of these environments, however, little is known about their adaptations and molecular responses to the conditions of their habitat.

Here, we generated a highly continuous nuclear genome sequence for *Ancyromonas sigmoides* and compared it against taxonomic comprehensive databases to characterize its architecture and classify the origins of its protein coding genes. The further comparative analyses of this genomic sequence will refine our understanding of the major transitions during early eukaryotic diversification as well as provide insights on the ecological roles and adaptations of ancyromonads to their environments.

Results and discussion

The hybrid assembly of Illumina and PacBio data resulted in a nuclear genome sequence of almost 40 Mb (Table 1) after curation and decontamination. We were able to track the presence of 76.1% of the Benchmarking Universal Single-Copy orthologs (BUSCO) present in the eukaryota_obd10 database. From these markers, 11.8% were found to be fragmented in the genome, and the proteome derived from the transcriptome, suggesting some of these genes have a different structure in this species rather than being fragmented due to the lack of contiguity in the assembly.

The transcriptome-inferred proteome showed a lower BUSCO score (C:53.0%[S:31.0%,D:22.0%],F:13.7%,M:33.3%) and 94.01% reads from this transcriptome could be aligned to the genome, suggesting that the genomic sequence nearly represents the gene complement of the species.

Table 1. Genome assembly statistics

Ancyromonas sigmoides	
Genome assembly size (Mb)	39.76
Number of contigs	202
Contig mean length (Kb)	196.86
Longest contig (Kb)	891.69
Shortest contig (Kb)	10.30
N50 (Kb)	399
L50	35
GC content	58.22
Total repeat content	29 %
Protein coding genes	11,138
BUSCO eukaryota_odb10	C:64.3%[S:61.6%,D:2.7%], F:11.8%,M:23.9%

The genome of *Acyromonad sigmoides* contains a big proportion of interspersed repetitive elements clustered in families previously uncharacterised as well as simple repeats (Figure 1a). Moreover, ancyromonad protein coding genes have a mean length of ~2000 bp and 80% of the genes contain from 1 to 48 introns per gene.

Genes were found to use universal genetic code and up to 97.5% of their splicing junctions are canonical. Among functionally annotated proteins, those related to signal transduction were abundant as well as proteins involved in post translational modifications and chaperone functions (Figure 1b). Similarly to the non coding elements 34.6% of the protein coding genes could not be functionally annotated with eggNOG-mapper. Some of these contain structural protein domains characterized by InterproScan, for example 766 of these genes bear transmembrane domains.

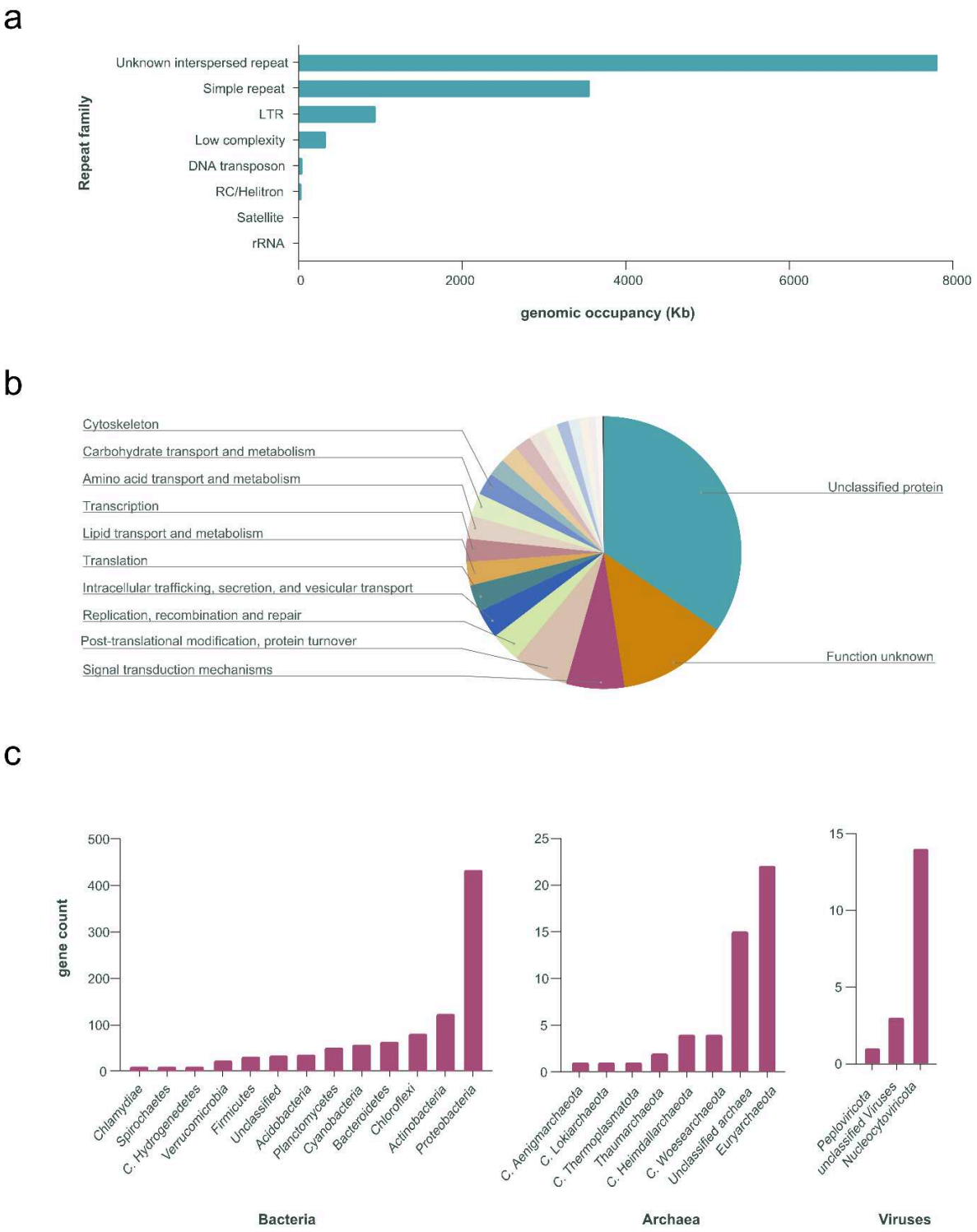

Figure 1. Genome features of *Ancyromonas sigmoides*. a) genomic occupancy of different classes of repetitive elements b) COG functional categories assigned by egg-NOG mapper to the inferred proteins c) taxonomic distribution of proteins with non-eukaryotic best hits in the *nr* database.

The most abundant domains found in these uncharacterized proteins corresponded to Leucine-rich repeat (PF13516) in 41 proteins (classified within the RNI-like and RAN GTPase activators families), and the Ankyrin domains (62 proteins) involved in protein-protein interactions and present in other proteins with diverse functions such as signal transduction, cell adhesion and cell-cycle regulation among others.

Another notably abundant Pfam domain within the genome of *Ancyromonas sigmoides* was EsV-1-7 cysteine-rich motif (PF19114) observed in 38 of the *A. sigmoides* predicted genes. This domain, whose name is derived from the *Ectocarpus* virus EsV-1 protein EsV-1-7, which possesses six EsV-1-7 repeats, has been observed to be distributed in brown algae, green algae, oomycetes and cryptophytes (Macaisne et al. 2017). Interestingly, when compared against the *nr*, the proteins *A. sigmoides* bearing this domain have as best hit hypothetical proteins of diverse algae and double-strand DNA giant viruses from the Phycodnaviridae family, suggesting these proteins could have been horizontally transferred via viral infections. Previous studies have pointed out the importance of viruses in the evolution of protist genomes and have shown that endogenous viral elements (EVEs) can make up a large proportion of protist genomes (Bellas et al. 2023).

Additionally, we identified 1,212 proteins with best hits to non-eukaryotic genes in the *nr* (Figure 1c), where bacteria represented the most abundant best hits, followed by archaea, and viruses from the Nucleocytoviricota family.

Bacterial hits were distributed across 50 phyla where the most abundant was Proteobacteria by far, followed by Actinobacteria and Chloroflexi. Sequences with best hits within Archaea were much less numerous, and distributed mostly among the Euryarchaeota or of unclear archaeal affiliation. These genes could either represent lateral gene transfers from these organisms to *Ancyromonas sigmoides* genome or, given the deep position of ancyromonads in the tree of eukaryotes, these could be ancestral eukaryotic genes that have been lost in other deep-branches of the tree.

Conclusion

The genome of *Ancyromonas sigmoides* represents the first one for the foundational resource for future comparative genomic studies. The proportion of likely lineage-specific proteins and repeated element families in the genomic sequence is coherent given the deep divergence of *A. sigmoides* from other sequenced species of eukaryotes and underpins the importance of a detailed molecular characterization of this emerging model species. Moreover, several of the proteins of this species might represent lateral gene transfers between eukaryotes, possibly driven by the introduction of viral sequences, as well as from prokaryotes. Further examination of the genomic diversity of ancyromonads is required to clarify the molecular basis of the adaptations of this species and the origin of its genomic diversity.

Methods

Ancyromonas sigmoides B70 **cell culturing, nucleic acids extraction and sequencing.**

Cultures of *Ancyromonas sigmoides* strain B70 (CCAP 1958/3) were obtained from Cavalier-Smith laboratory (Oxford). For DNA extraction cultures were grown in 50% Cerophyle solution and 50% of filtered sterile seawater in 25-ml culture flasks at room temperature (~18º C). Culture plates were inoculated with 500 µl derived from mature stock cultures. After 2-3 days once cultures reached a medium to high cell density, the supernatant was removed, and cells were gathered using disposable cell scrapers and distributed across new plates. This process was repeated several times up to reaching around 100 culture flasks. From these, DNA was extracted in parallel using phenol/chloroform/isoamyl alcohol (25:24:1), extracted again using chloroform/isoamyl alcohol (24:1), precipitated overnight in 95% EtOH at -20°, pelletted in a centrifuge at 4°, washed with 80% EtOH, and resuspended in ddH2O.

A total of 22.62 Gbp long read data was generated in a PacBio RS II platform of the Cold Spring Harbor Laboratory. Additionally 2 libraries of and of short pair-end reads and four libraries of mate-pair reads were generated in an Illumina Nextera platform at XX using a 2x150 configuration yielded to the generation of 24.76 Gbp and CC Gbp of sequencing data respectively in the Weill Cornell's Genome Resources Core Facility.

From another 8 flasks of cultures cells were harvested to extract a total of 4,186 ng of RNA (322 ng/µl in 13 µl) using the RNeasy mini Kit (Qiagen), following the manufacturer protocol. A cDNA Illumina library was constructed after polyA mRNA selection, and was sequenced using a paired-end (2 × 150 bp) configuration in a Illumina HiSeq 2500 (Chemistry v4) platform at *Eurofins Genomics*, Germany.

Short reads quality control

Short pair-end reads were filtered and trimmed based on their quality and presence of Illumina adapters using Trimmomatic (Bolger et al. 2014). Only sequences with a minimum of 28 PHRED scores in the 90% percent of bases and a minimum average quality of 30 PHRED scores were kept for further analyses. We estimated an assembly expected size of 81,595,175 bp based on the kmer frequencies and distribution of trimmed Illumina PE reads using *kmergenie* (Chikhi & Medvedev 2014). Moreover, MP reads were preprocessed using NxTrim (O'Connell et al. 2015) to eliminate adapters and identify true MP, PE and uncertain paired reads within our libraries.

Genome assembly and refinement

Two independent assembly strategies were tested in the long (PacBio) and the high quality short (Illumina PE) read sequencing data. The first one was based on generating a draft assembly with high continuity with the long reads and then improving the assembly by correcting the base calling with the short reads. Canu v1.9 (Kriventseva et al. 2019; Koren et al. 2017), and Flye v2.8.2 (Kolmogorov et al. 2019) long-read assemblers were compared with this purpose and then we performed an iterative

genome polishing step mapping the short pair-ended reads to the consensus unitigs to improve the base call of the resulting genomic sequence five times. The second strategy consisted of generating first a fragmented but accurate assembly with the short Illumina reads using SPAdes v3.15.3 (Vasilinetc et al. 2015; Prjibelski et al. 2020), then these accurate contigs were scaffolded in a backbone with the DBG2OLC (Ye et al. 2016) hybrid approach using the long-reads corrected by Canu. The backbone was then polished by an iterative cycle of mapping the long reads to the genomic sequence using minimap2 (Li 2018), generating a consensus sequence using Racon v1.3.1 (Vaser et al. 2017) and then a step of polishing with the short reads using bwa-mem v0.7.15 (Li & Durbin 2009) and Pilon v1.22 (Walker et al. 2014). The completeness of the genomic sequences was assessed with BUSCO v5.4 using the eukaryota_obd10 database. Based on the assessment of contiguity and completeness the SPADes+DBG2OLC assembled genome was chosen for its better overall quality.

Prediction of genomic features

Transposable elements (TEs) and other repeats were identified *de novo* using the high-scoring pair and k-mer frequency approaches implemented in the RepeatModeler2 pipeline (Flynn et al. 2020). The identified repeats were classified and masked before gene prediction using RepeatMasker v4.1.2-p2, using the sensitive mode. The genetic code of the genome sequence was inferred using Codetta (Shulgina & Eddy 2023). Protein coding genes prediction was then performed with Braker2 (Brůna et al. 2021; Hoff et al. 2016; Lomsadze et al. 2005) using both protein hints and the splicing hints generated from the RNA-seq data mapping (see hereafter). With this purpose, a custom protein database was built by concatenating the predicted proteins from all decontaminated *Ancyromonas sigmoides* transcriptome and the “protozoa'' set from the OrthoDB (Kriventseva et al. 2019) protein database. These proteins were then used as a reference for ProtHint (Brůna et al. 2020) to predict and score gene hints in the genome sequences. We also generated intron junction hints by mapping the RNA-seq data with STAR v2.5 (Dobin et al. 2013) against the genomic sequence in two rounds. In the first one, non-canonical splicing sites were identified and used as input for the second

mapping round from which 70.31% of the reads were uniquely mapped and 23.70% mapped to multiple loci of the genomic sequence.

The resulting protein predictions were annotated with eggNOG mapper v2 (Cantalapiedra et al. 2021) using the diamond search mode and all domains of life as target space. We implemented InterproScan v5 (Jones et al. 2014) analysis and BLASTP searches against the non-redundant (*nr*) database of the National Center of Biotechnology Information (NCBI) downloaded in February of 2022, to annotate functional domains and taxonomic affiliation of the proteins respectively.

Identification and filtering of contaminant sequences

Contaminant contigs were identified based on the taxonomic affiliation of their encoded proteins and their sequencing coverage biases. First, Prodigal was used to perform an initial protein prediction on the genomic contigs, these proteins were used as a query for diamond (Buchfink et al. 2015) searches against a custom database containing all the proteins from the GTDB release214 (Parks et al. 2022) and the Eukprot v3 (Richter et al. 2022) databases. A protein was classified as prokaryotic if its best hit was any protein from the GTDB database and had over 80% identity and coverage over 50% of the query length. Contigs encoding 50% or more prokaryotic proteins that represented more than the 10% of the contig length were classified as contaminants and discarded. Secondly, eukaryotic gene predictions were generated and refined following the Braker2 pipeline (see details below). The inferred proteins were then used as a query against the non-redundant nucleotide database and the GTDB-EukProt custom database. Contigs with putative contaminants were discarded if the contig did not contain eukaryotic proteins or any spliced-transcript. The genome sequence was then screened with Blobtools3 (Challis et al. 2020) to identify and inspect any contig with low coverage and biased GC content.

A similar pipeline was followed to decontaminate the transcriptomic dataset. First, transcriptome reads were quality-controlled and assembled using RNA-SPAdes (Prjibelski et al. 2020; Bushmanova et al. 2019). ORFs were then obtained using

Transdecoder (Haas & Papanicolaou 2017) and then screened for contamination with Blobtools3. Contaminant transcripts were discarded and a set of clean reads was retrieved by mapping the libraries against the clean transcripts with Hisat2 (Kim et al. 2019).

Acknowledgments

The authors thank the IFB-core cluster facilities and its support team for their help in the management of large datasets and software installation. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (ERC Starting Grant No 803151 to L.E. and ERC Advanced Grants No 322669 and 787904 to P.L.-G. and D.M.).

Literature cited

1. Atkins MS, McArthur AG, Teske AP. 2000. Ancyromonadida: a new phylogenetic lineage among the protozoa closely related to the common ancestor of metazoans, fungi, and choanoflagellates (Opisthokonta). J. Mol. Evol. 51:278–285.
2. Bellas C et al. 2023. Large-scale invasion of unicellular eukaryotic genomes by integrating DNA viruses. Proc. Natl. Acad. Sci. U. S. A. 120:e2300465120.
3. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 30:2114–2120.
4. Brown MW et al. 2018. Phylogenomics Places Orphan Protistan Lineages in a Novel Eukaryotic Super-Group. Genome Biology and Evolution. 10:427–433. doi: 10.1093/gbe/evy014.
5. Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. 2021. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genom Bioinform. 3:lqaa108.
6. Brůna T, Lomsadze A, Borodovsky M. 2020. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. NAR Genom Bioinform. 2:lqaa026.

7. Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nat. Methods. 12:59–60.
8. Burki F, Roger AJ, Brown MW, Simpson AGB. 2020. The New Tree of Eukaryotes. Trends in Ecology & Evolution. 35:43–55. doi: 10.1016/j.tree.2019.08.008.
9. Bushmanova E, Antipov D, Lapidus A, Prjibelski AD. 2019. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. Gigascience. 8. doi: 10.1093/gigascience/giz100.
10. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. 2021. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. bioRxiv. 2021.06.03.446934. doi: 10.1101/2021.06.03.446934.
11. Cavalier-Smith T, Chao EE-Y. 2003. Phylogeny of choanozoa, apusozoa, and other protozoa and early eukaryote megaevolution. J. Mol. Evol. 56:540–563.
12. Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. 2020. BlobToolKit--interactive quality assessment of genome assemblies. G3: Genes, Genomes, Genetics. 10:1361–1374.
13. Chikhi R, Medvedev P. 2014. Informed and automated k-mer size selection for genome assembly. Bioinformatics. 30:31–37.
14. Dobin A et al. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 29:15–21.
15. Flynn JM et al. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. Proc. Natl. Acad. Sci. U. S. A. 117:9451–9457.
16. Haas B, Papanicolaou A. 2017. TransDecoder.
17. Heiss AA, Walker G, Simpson AGB. 2010. Clarifying the taxonomic identity of a phylogenetically important group of eukaryotes: Planomonas is a junior synonym of Ancyromonas. J. Eukaryot. Microbiol. 57:285–293.
18. Heiss AA, Walker G, Simpson AGB. 2011. The ultrastructure of Ancyromonas, a eukaryote without supergroup affinities. Protist. 162:373–393.
19. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. 2016. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. Bioinformatics. 32:767–769.

20. Jones P et al. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics. 30:1236–1240.
21. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. 37:907–915.
22. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. Nat. Biotechnol. 37:540–546.
23. Koren S et al. 2017. Canu: scalable and accurate long-read assembly via adaptive -mer weighting and repeat separation. Genome Res. 27:722–736.
24. Kriventseva EV 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 Research. 47:D807–D811. doi: 10.1093/nar/gky1053.
25. Lax G et al. 2018. Hemimastigophora is a novel supra-kingdom-level lineage of eukaryotes. Nature. 564:410–414.
26. Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 34:3094–3100.
27. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 25:1754–1760.
28. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. 2005. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. 33:6494–6506.
29. Macaisne N et al. 2017. The Ectocarpus IMMEDIATE UPRIGHT gene encodes a member of a novel family of cysteine-rich proteins with an unusual distribution across the eukaryotes. Development. 144:409–418.
30. O'Connell J et al. 2015. NxTrim: optimized trimming of Illumina mate pair reads. Bioinformatics. 31:2035–2037.
31. Paps J, Medina-Chacón LA, Marshall W, Suga H, Ruiz-Trillo I. 2013. Molecular phylogeny of unikonts: new insights into the position of apusomonads and ancyromonads and the internal relationships of opisthokonts. Protist. 164:2–12.
32. Parks DH et al. 2022. GTDB: an ongoing census of bacterial and archaeal diversity

through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Res. 50:D785–D794.

33. Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. 2020. Using SPAdes De Novo Assembler. Curr. Protoc. Bioinformatics. 70:e102.
34. Richter DJ et al. 2022. EukProt: A database of genome-scale predicted proteins across the diversity of eukaryotes. Peer Community J. 2. doi: 10.24072/pcjournal.173.
35. Saville-Kent W. 1882. *A Manual of the Infusoria: plates*.
36. Shulgina Y, Eddy SR. 2023. Codetta: predicting the genetic code from nucleotide sequence. Bioinformatics. 39. doi: 10.1093/bioinformatics/btac802.
37. Tikhonenkov DV et al. 2022. Microbial predators form a new supergroup of eukaryotes. Nature. 612:714–719.
38. Tikhonenkov DV, Mazei YA, Mylnikov AP. 2006. Species diversity of heterotrophic flagellates in White Sea littoral sites. Eur. J. Protistol. 42:191–200.
39. Torruella G, Moreira D, López-García P. 2017. Phylogenetic and ecological diversity of apusomonads, a lineage of deep-branching eukaryotes. Environ. Microbiol. Rep. 9:113–119.
40. Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 27:737–746.
41. Vasilinetc I, Prjibelski AD, Gurevich A, Korobeynikov A, Pevzner PA. 2015. Assembling short reads from jumping libraries with large insert sizes. Bioinformatics. 31:3262–3268.
42. Walker BJ et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 9:e112963.
43. Ye C, Hill CM, Wu S, Ruan J, Ma ZS. 2016. DBG2OLC: Efficient Assembly of Large Genomes Using Long Erroneous Reads of the Third Generation Sequencing Technologies. Sci. Rep. 6:31900.
44. Yubuki N et al. 2023. Molecular and morphological characterization of four new ancyromonad genera and proposal for an updated taxonomy of the Ancyromonadida. J. Eukaryot. Microbiol. e12997.

6. Comparative genomics of ancyromonads since their divergence from other eukaryotic supergroups

Context and results summary

In a recent study by our team, it is shown that although morphologically similar, ancyromonad species diversity is large (Yubuki et al. 2023). Ancyromonads are distributed in diverse benthic and soil environments, suggesting that they suffered ecological transitions however their biology it's just started being understood.

In this chapter, we assembled and characterized the genomic sequence of six recently described species of ancyromonads isolated from diverse freshwater and marine sediments. The characterization of these new genomes has revealed that ancyromonads have evolved diverse genome architectures.

Furthermore, using phylogenetic reconciliation and a large dataset of publicly available eukaryotic proteomes we reconstructed the deep evolutionary history of the protein-coding gene content of the Ancyromonadida clade since their divergence from other major eukaryotic supergroups. In order to do this we reconstructed a species phylogeny in which ancyromonads are one of the first lineages diverging and form a sister group of Metamonada + *Gefionella okellyi* the only available representative for Malawimonadida. Furthermore the reconstruction of single gene families distributed across the compared species and its reconciliation with the species phylogeny revealed the patterns of gene gain and loss across eukaryotes. In particular, a high number or gene family originations and eukaryote to eukaryote gene transfers seem to have been specially important at the base of ancyromonadida clade.

Moreover, several ancyromonad exclusive genes (among eukaryotes) have homologues in archaea and bacteria. These genes represent putative lateral gene transfer into ancyromonads from prokaryotes that traces back to their diversification from their last common ancestor and involve proteins with diverse functions.

The detailed analysis of the functions of the gene families evolving across ancyromonads and acquired from prokaryotes is discussed in the following manuscript but is still ongoing. In particular, we need to be cautious when discussing losses of canonical eukaryotic components given the relative incompleteness of some of the genomes and the possibility that some of these gene families have evolved in ancyromonads beyond recognition. In addition, we are further investigating in the literature how the genes discussed represent common adaptations found in other organisms or are particular to ancyromonads. Therefore, a more in depth analysis and review of literature are needed to clarify these aspects and wrap-up this manuscript. The current results of this part of the project are presented in the following draft manuscript.

A window into the ancient genome evolution of ancyromonads, a deeply divergent clade of eukaryotes.

Jazmin Blaz, Naoji Yubuki, Maria Ciobanu, Brittany Baker, Luis J. Galindo, Guifré Torruella, Aaron Heiss, John Burns, Eunsoo Kim, Puri López-García, David Moreira and Laura Eme

Abstract

Ancyromonadida comprises a diverse group of free-living flagellates of profound evolutionary significance due to their divergence from all the known eukaryotic supergroups. Here we sequenced the nuclear genomes of six ancyromonad species unveiling their genomic architecture and gene content. Through a large-scale evolutionary analysis, we showed that a significant amount of gene family originations predate the diversification of ancyromonads. Important turnovers of gene families involved in signal transduction and cytoskeleton associated proteins further contributed to a great variation of the gene content among modern species. Moreover, some ancyromonads also acquired several genes from prokaryotes that could have facilitated their adaptation to benthic environments. Noteworthy among these acquisitions we found some proteins involved in the transport and metabolism of nitrate, which provide hints for the understanding of the ecological roles of ancyromonads in the nutrient cycles. Finally, we discussed the implications of our findings into the understanding of the earliest eukaryote radiations and the gene content of the last eukaryotic common ancestor. Our study provides the first insights into the intriguing genome diversity of ancyromonads and offers an unique view of the early evolution of the eukaryote domain.

Introduction

Since their diversification from a common ancestor, eukaryotes have split into several major lineages, also known as supergroups (Burki et al. 2020). The comparison of extant species from diverse supergroups has been crucial to reconstruct the features of the last eukaryotic common ancestor (LECA) (Vosseberg et al. 2021; Koumandou et al. 2013; Hampl et al. 2008; Speijer, Lukeš, and Eliáš 2015; Yubuki and Leander 2013; Richards and Cavalier-Smith 2005; Weiner et al. 2020) and to generate hypotheses about the mechanisms driving genome evolution across eukaryotes (Doolittle 1998; Schaack, Gilbert, and Feschotte 2010; Fritz-Laylin et al. 2010; López-García, Eme, and Moreira 2017; Vosseberg et al. 2021; Collens and Katz 2021). Furthermore, the discovery of lineages with a deep divergence within the eukaryotic tree of life (eToL) has dramatically expanded our understanding of the microbial diversity of this domain of life (Janouškovec et al. 2017; Brown et al. 2018; Schön et al. 2021; Lax et al. 2018; Galindo, López-García, and Moreira 2022; Tikhonenkov et al. 2022; Eglit et al. 2023). These discoveries raise new questions and play a central role in unraveling the early evolution of eukaryotes.

One of these lineages is the Ancyromonadida clade. Ancyromonads are phagotrophic and feed from prokaryotes that graze from their natural environments (Saville-Kent 1882; Heiss, Walker, and Simpson 2011). Species of this group have been previously observed thriving in benthic sediments of marine and freshwater environments or soil samples across the globe (Yubuki et al. 2023) suggesting they are capable of adapting to a wide variety of conditions. Although they display seemingly resembling morphologies, a recent analysis of new ancyromonad isolates has uncovered the existence of a previously neglected diversity of species (Yubuki et al. 2023), revealing ancyromonads are indeed a species rich high-ranking taxon. Despite their ecological role and evolutionary importance, little is known about the biology of these organisms and ever since they were described remain elusive to be robustly placed into global eukaryotic phylogenies and exhibit no affinity to any eukaryotic supergroup (Atkins, McArthur, and Teske 2000; Cavalier-Smith and Chao 2003; Paps et

al. 2013; Torruella, Moreira, and López-García 2017; Brown et al. 2018). Therefore ancyromonads have been considered for a long time an orphan branch of the eToL.

Here we sequenced the genome of six distantly related species of ancyromonads. Using a phylogenomic approach, we reconstructed the evolutionary processes shaping their genomic repertoires since their early diversification from other eukaryotes. We hypothesized how this genomic diversity has contributed to the ecological versatility of ancyromonad. Finally, we discuss how the inclusion of ancyromonads into the reconstruction of the ancestral genetic repertoires in eukaryotes can shed light into the genomic innovation that took place at the base of this domain.

Results and discussion

I. Ancyromonads exhibit diverse genome architectures and gene content.

We recently isolated and characterized several new species of ancyromonads (Yubuki et al. 2023). Here, we employed a custom strategy to couple cell sorting and whole genome amplification (WGA) to obtain the genomic sequence of six ancyromonad species representing the main branches within the clade, therefore filling an important gap into the genomic resources available for unicellular protists belonging to orphan clades. We detected the presence of 65 to 95% of the BUSCO markers of the eukaryota_odb10 (see Supplementary Material. *Genome quality assessment*) in the genomic sequences. The recovered genomic sequences display an important variation in their density of coding and non coding elements as well as their protein coding gene architecture (Figure 1 and Supplementary Material. *Genomes features*). Notably the genomes of *Ancyromonas sigmoides* and *Nutomonas limna* harbor the highest content of repetitive elements (occupying up to 8 and 4 Mbp of their genomes respectively) as well these species present larger genes bearing up to 53 and 67 introns, respectively. In contrast, *Planomonas micra, Ancyromonas mediterranea and Striomonas longa* exhibit more compact genomes with higher protein coding gene density.

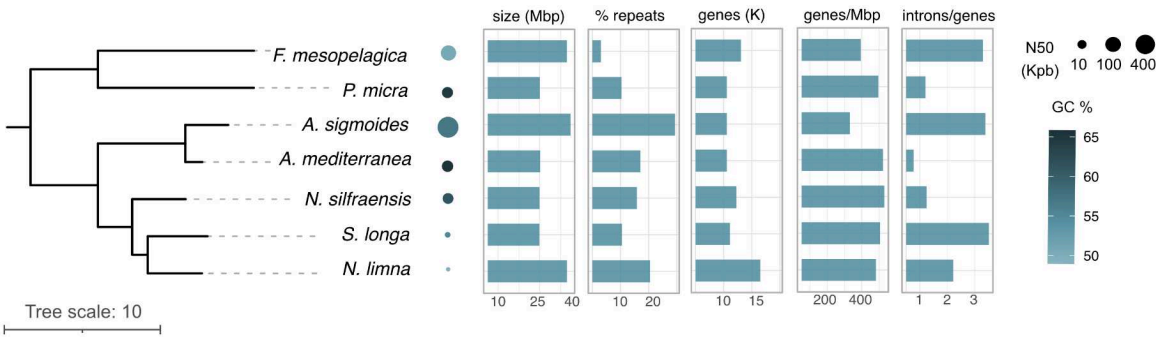

Figure 1. Ancyromonad phylogenetic relationships (left) and genome assemblies features (right). Maximum Likelihood phylogeny (based on 2,003 concatenated proteins selected by OrthoFinder) inferred with IQ-Tree under the LG+C60+G model. The tree was rooted between the clade with *F. mesopelagica* + *P. micra* and everything else, based on Yubuki et al. 2023.

Although most of the repetitive elements of these species were unknown interspersed elements, we observed a difference in the presence of classified families of elements (Figure S2) and number of proteins associated with transposable elements. For example, we identified 94 genes with integrases and 97 genes bearing reverse transcriptase domains in *A. sigmoides.* In comparison, *N. limna* which also exhibits a large and highly repetitive genome but had only 16 and 25 proteins with such domains respectively. This suggests that different mobile elements have contributed to the diversification of ancyromonad repetitive landscapes and the increase of genome sizes in these species. The distribution of Clusters of Orthologous genes (COG) categories was similar among all the genomes (Figure S6). However, a substantial percentage of the predicted proteins encoded in these genomes lacked detectable homologous proteins in the EggNOG v5 database (27-37%). When comparing the distribution of gene families across ancyromonads, we observed a high number of species specific gene families (Supplementary Material. Figure S7), although this extensive variation of gene content between species needs to be put in perspective with the relative incompleteness of some genomes (e.g. we detected only 63% of the eukaryota_odb10 BUSCO markers in *A. mediterranea*). Nevertheless, this suggests that despite their morphological similarities, ancyromonads have diverged greatly in terms of gene content since their last common ancestor.

II. Reconstruction of the evolutionary processes underlying the genomic divergence of ancyromonadida

To gain insight into the mechanisms giving rise to the diverse gene content of ancyromonads since their divergence from extant supergroups we conducted a large-scale evolutionary analysis using phylogenetic reconciliation of gene families and a species phylogeny. We collected a comprehensive database of eukaryotic proteomes that included the genomes of seven ancyromonads, two mantamonads (Blaz et al. 2023), and the inferred proteomes of 196 species belonging to the comparative set of the EukProtv3(Richter et al. 2022) database. We reconstructed 377,632 gene families encompassing 4,221,246 genes distributed across these species, for which we inferred maximum likelihood phylogenies. Using the single copy orthologs retrieved from this analysis we inferred a species phylogeny using a concatenation of 799 single copy orthologs and a phylogenetic constraint based on the current consensus of the eToL (Burki et al. 2020; Richter et al. 2022) (see Methods and Figure S8 at Supplementary Material). The resulting phylogeny was then rooted based on an updated version of the Opimoda and Diphoda split (Derelle et al. 2015).

In our phylogenomic reconstruction (Figure 2a and Figure S9), ancyromonads form a monophyletic group that branches as sister clade comprising Metamonada and *Gefionella okellyi* (the only representative of Malawimonadida in our dataset) with a support of 68%. Metamonada comprise the largest group of anaerobic protists and have evolved a diverse array of mitochondrial related organelles (MRO) and diverse lifestyles (Leger et al. 2017; Tachezy 2019; Stairs et al. 2021; S. K. Williams et al. 2023). Moreover, Malawimonada is a poorly sampled lineage of heterotrophic flagellates that exhibits ventral groove (O'kelly and Nerad 1999; Heiss et al. 2018). Malawimonada-Metamonada (MM) relationship has been previously observed (Heiss et al. 2018). Remarkably, Hemimastigophora places as sister branch of the clade containing Ancyromonads+(MM) with low support (69%), contrasting with recent studies which place them as a deep branching lineage within Diphoda (Tikhonenkov et al. 2022; Eglit et al. 2023). All these groups have very different morphologies, diverse lifestyles

and their phylogenetic positions have been historically hard to disentangle. The moderate support for Ancyromonads+(MM) and Hemimastigophora as sister-branch of them in our reconstruction could be attributed to the limitations of gene and species sampling in the vicinity of these lineages. An analysis with a richer taxon sampling and curated phylogenetic markers is needed to clarify the deep-level position of these clades. Our dataset and reconstruction can yield interesting comparisons with more standard datasets.

The reconstruction of the evolutionary history of gene families across our species dataset pinpointed the evolutionary trends of gene loss, gain and expansion of eukaryotic proteomes since their divergence from their last common ancestor (Figure 2b-c). This analysis has also revealed a high number of taxonomically restricted genes in ancyromonads (Supplementary Material. Figure S10). Accordingly, our reconciliation results suggest that the last ancyromonad common ancestor has undergone an important number of gene family originations in comparison to the overall events of gene family gain observed at the base of other eukaryotic clades (Figure 2b-c). These gene families could have originated by domain fusion and shuffling, as well as horizontal gene transfers from species outside the dataset (i. e. viruses, prokaryotes, unsampled eukaryotes, this possibility is explored in a further section). Additionally, many of them (n= 4,773 gene families) do not have tractable homologs in other species from our dataset, the *nr* or the prokaryotic database GTDB and therefore they could either be ancient genes that have evolved beyond recognition or the result of *de novo* gene birth events in ancyromonads. Diverse authors have proposed that gene birth, in combination with the rewiring of conserved regulatory networks of a genome, might contribute to the phenotypic diversity and adaptation to particular niches of organisms (Khalturin et al. 2009; Tautz and Domazet-Lošo 2011; Arendsee, Li, and Wurtele 2014). Therefore, the existence of these types of genes in ancyromonads could underlie particular phenotypes and raise the following questions: Do they express differently under certain conditions?, Are they under selection pressures? Is the proportion of taxonomically restricted genes similar to other orphan lineages of protists?

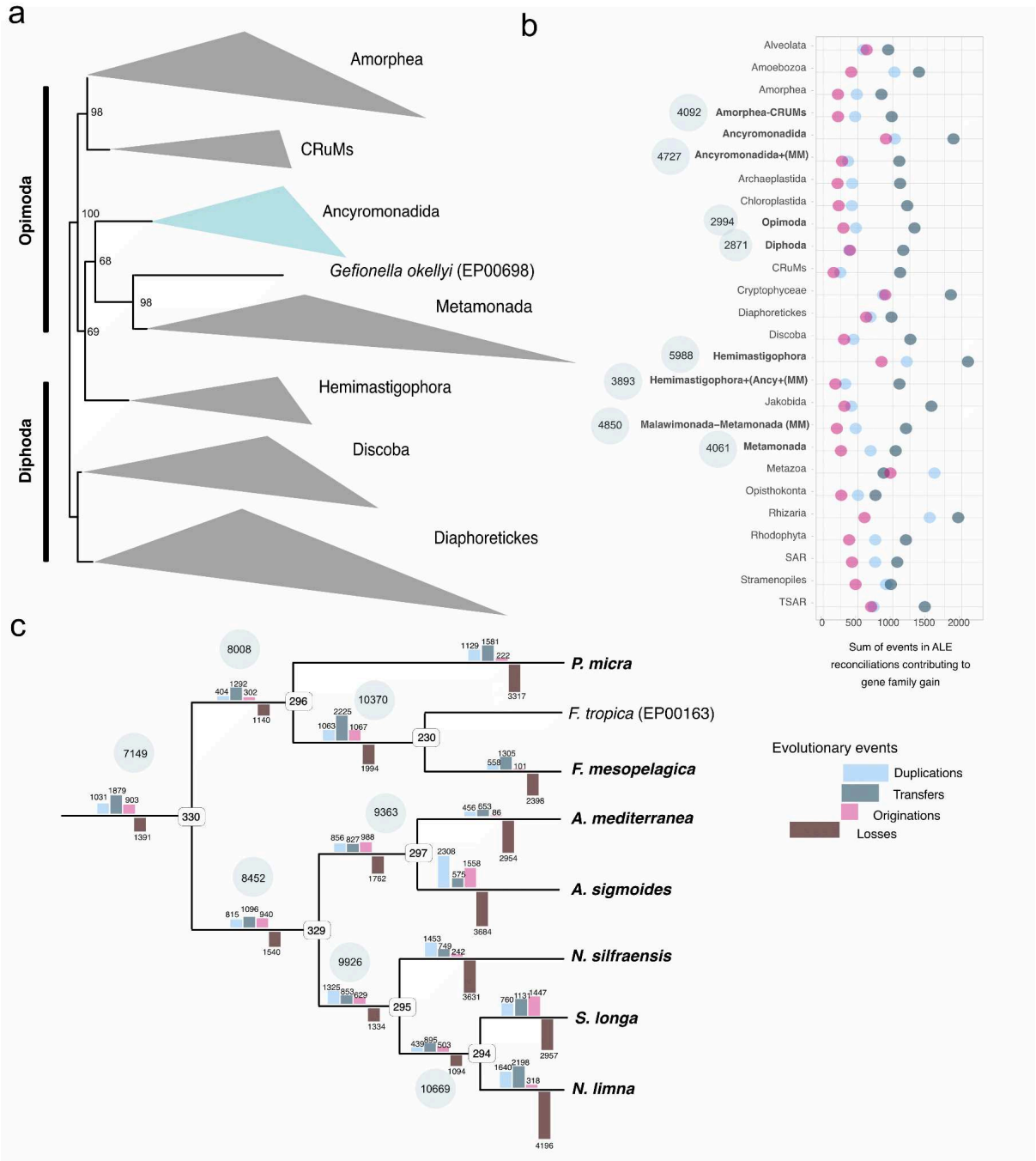

Figure 2. Ancyromonadida phylogenetic position of new ancyromonad species within the eToL and evolution of their genomic repertoires. a) Eukaryotic phylogeny based on the analysis of 766 markers, 205 taxa and 47,611 sites and inferred using IQ-Tree under the LG+C60+G model. Support at branches were estimated using 1,000 bootstrap replicates. b) Numbers and types of evolutionary events inferred by ALE to explain gene family gains at the base of different eukaryotic lineages. Numbers represent the sums of all evolutionary events inferred by ALE (see color code) and number of gene copies (within circles) for each node. c) Patterns of gene content evolution across the ancyromonad clade. The numbers in the cladogram splits correspond to the name of the internal node.

Although most of the genes acquired into the ancestral and modern proteomes were inferred to have been vertical inherited (Figure S11), we observed a striking contribution of eukaryote to eukaryote lateral gene transfer (eukLGT) at the base of of ancyromonads and into particular points of their diversification (Figure 2c). Similarly, a high amount of eukLGT was inferred at early splits within the species phylogeny (Figure 2b). This pattern is more prevalent at early splits than across the clades, where duplication was inferred to be the main mechanism of gene acquisition (Figure S12). SIngle gene family phylogenies could meet more difficulty to solve ancient events than recent ones and this could affect the transfer frequency inferred by ALE, and can also be affected by taxon density (T. A. Williams et al. 2023), therefore disentangling punctual transfer events will require a manual inspection of these phylogenies. However, to our knowledge, there there is no previous estimations of lateral transfer rate across eukaryotes at this evolutionary scale, therefore, these results add evidence to the debate about the importance of eukLGT during the early evolution of eukaryotes (see further discussion in the section VI).

III. Evolution of key eukaryotic components across ancyromonads

The last ancyromonad common ancestor was predicted to have a genome repertoire of around 7.5K genes in our ALE evolutionary reconstruction. The functional characterization of the gene families under change at key points of ancyromonad evolution (Figure 3 and Supplementary Material Figures S13-S14) revealed that the branch leading to this ancestor has undergone an important turnover (acquisitions and losses) in gene families involved in key eukaryotic components such as the cytoskeleton and signal transduction mechanisms.

For example, 44 gene families with microtubule motor and actin binding activity and 9 gene families widely distributed in the dataset and with a probable ancient origin were lost in ancyromonads. In contrast, several gene family originations in the base of ancyromonads included tetratricopeptide-domain proteins, proteins belonging to the TRAFAC class myosin-kinesin ATPase superfamily, and Filamin-like proteins bearing

immunoglobulin domains. Similarly, 152 gene families predicted to have a protein serine/threonine kinase activity and widely distributed across the species of our dataset were not found in ancyromonads. In contrast, gained gene families involved in signal transduction included several calcium ion binding proteins, cGMP and cAMP binding proteins, proteins bearing EGF domains and G proteins bearing Rho domain.

Interestingly, some of the families gained at the base of ancyromonads also included several genes involved in the semaphorin-plexin signaling pathway. This signalization pathway regulates cell morphology and motility by triggering changes to the cytoskeletal and adhesive machinery that regulate cellular morphology in many different mammal cell types (Alto and Terman 2017). Other gained gene families that could be involved in a related function were found to harbor phosphatidylinositol binding, BAR, EFC/F-BAR, and the IMD/I-BAR domains which have been reported to be involved in the change of membrane shape by interacting with the actin polymerization machinery and phosphoinositide (Takenawa 2010). These changes suggest that ancyromonads have evolved specialized gene families involved in cytoskeleton and signal transduction.

Other gene family originations at the base of ancyromonads conserve domains shared with proteins that are part of other canonical eukaryotic systems. This could mirror the result of the evolutionary tinkering of this preexisting machineries in ancyromonads. An example of these are the gene families bearing the Regulator of Chromosome Condensation 1 (RCC1) repeat (OG0091149 and OG0074833), often present in proteins involved in the microtubule coordination during the cell cycle, (Bischoff and Ponstingl 1991; Shields et al. 2003) and a gene family with the Mu homology domain (OG0157779, Pfam PF10291) present in endocytic adaptors such as Sip1 (H.-D. Li, Liu, and Michalak 2011; Reider et al. 2009).

Moreover, although the COG categories involved in the processing and storage of genetic information displayed overall less changes than the others (Figure 3), our analyses revealed a surprising retraction of gene families involved in translation and

traduction, at the basal node of ancyromonads (Supplementary Material. Figure S13). The losses in the category of translation included 12 gene families classified in other organisms as translation initiation factors, proteins with GTP hydrolysis activity and several proteins from the aminoacyl-tRNA synthetase superfamily. Moreover, the losses in transcription included 37 gene families of RNA polymerase II transcription factors. We observed also the loss of very conserved gene families involved in the modification of mRNA such as the mRNAs m6A methyltransferase complex (OG0037821 and OG0014797) and several rRNA methyltransferases. Further retractions of these gene families were also inferred in the genomes of *P. micra, A. meditarranea and N. silfraensis* (also the most streamlined genomes across the clade). This could imply that ancyromonads have evolved divergent RNA processing factors.

Considering the high variation among ancyromonad species we were also interested in the processes shaping the gene content inside the Ancyromonadida clade. The branches leading towards the two main clades of ancyromonads (node 296 and node 329 in the Figure 3) and the subsequent evolutionary trajectories towards modern species (Supplementary Material. Figure S13) displays different patterns of duplication and turnover of genes associated to different functional categories.

The evolutionary trajectory of the *Fabomonas* is mainly characterized by eukLGTs, gene family originations and losses. Moreover, although most of the families in this genome remain small, the biggest gene families in the *F. mesopelagica* genome consisted in recently duplicated proteins of Zinc finger C2H2-type domains and membrane bound predicted regions that could be involved in the response to environmental stimuli. Putative eukLGTs into this lineage included mainly gene families within function unknown, however 28 gene families predicted to have telomerase activity were inferred to be transferred from Opisthokonta. Similarly to *F. mesopelagica*, *P. micra* displayed a lower number of gene family duplications than other ancyromonads and several putative eukLGT. *P. micra* displays a higher number of species specific proteins than *F. mesopelagica.* The genomes of these lineages were also the ones with more BUSCO markers conserved. They share more genes with other

eukaryotes, this could be related with the fact that many more eukLGTs happened during their evolution, or that this lineage has retained more ancestral features than other ancyromonads.

Figure 3. Change in gene family copy numbers inferred in the branches leading to the ancestors and modern species of ancyromonads (splitted by associated COG category). Protein occurrences at a node were calculated by adding the number of copies inferred by ALE at this node for all gene families annotated as belonging to a given COG category. Node 330: Last ancyromonad common ancestor. Node 296: *Fabomonas* genus + *P. micra.* Node 230: *Fabomonas* genus. Node 329: *Ancyromonas* genus + *(N. silfraensis* + *(S. longa* + *N. limna).* Node 297: *Ancyromonas* genus. Node 295: *N. silfraensis* + *(S. longa* + *N. limna).* Node 294: *N. limna* + *S. longa.*

In comparison, there were more gene duplications predating the divergence of the modern species within the genera *Ancyromonas,* and the fresh-water species of *Nyramonas, Striomonas,* and *Nutomonas* (Supplementary Material. Figure S13). Cytoskeleton protein families were expanded at the base of this group, as well as several proteins involved in signal transduction and carbohydrate transport and

metabolism (node 329 in Figure 3). This could suggest that the species within this clade have evolved different strategies to respond to their environments in comparison with *Planomonas* and *Fabomonas* genera.

Among all ancyromonads, *A. sigmoides* is the species with more duplications during its evolutionary trajectory. Besides the functional categories shown in Figure 3, the most duplicated gene families within this genome consisted in uncharacterized genes harboring gypsy retrotransposon domains, as well as uncharacterized proteins with EsV-1-7 cysteine-rich motifs of a possible viral origin. Uncharacterized gene families harboring CalX-like domains involved in calcium binding and regulation and transmembrane regions are also abundant in the genome of this species.

Furthermore, we observed important events of origination and duplication in transmembrane ion transport category, as well as aminoacid and carbohydrate metabolism and transport predating the diversification of *N.silfraensis + (S. longa +N. limna*) clade (node 295 in Figure3 and Supplementary Material Figure S13). Particularly ABC, and Major Facilitator family transporters were duplicated in this node. *S. longa* has further expanded its repertory of genes with predicted metallopeptidases activity and solute carrier families of transporters of Sodium-calcium exchange.

Moreover, although we observed rampant losses on annotated genes in most COG categories in *N. silfraensis,* this species has expanded several gene families with unknown functions belonging to the Quinoprotein alcohol dehydrogenase-like superfamily (OG0006040 with 91 copies) as well as uncharacterized proteins originated within ancyromonads and lacking homologous in the *nr*, GTDB and Interpro databases. These families included OG0006447 and OG0006040 with 59 and 87 copies in this species respectively.

Finally, *N. limn*a is by far the species with more species specific gene families. This genome also has other particularities such as high proportion of unknown interspersed repeat elements and genes with a high average number of introns. Our

reconstruction suggests that there was an overall expansion of genes associated to metabolism in this genome (Figure 3). Other recent expansions included RCC1- domain harboring genes, genes bearing the DEAD helicase C domain involved in the mRNA surveillance pathway, and OG0014972 predicted to belong to the bacterial porins OmpA family.

IV. Footprint of gene transfer beyond eukaryotes in ancyromonad unique families

A frequent source of genomic novelty into the genomes of unicellular eukaryotes are the lateral gene transfers (LGTs) from Bacteria and Archaea (Andersson et al. 2003; Becker, Hoef-Emden, and Melkonian 2008; Van Etten and Bhattacharya 2020; Woehle et al. 2022), often involving the acquisition of metabolic capacities. As we have seen, several of the proteins in the ancyromonad genomes share domains with known prokaryotic components. To characterize the putative ancestry of non eukaryotic LGT donors we screened the ancyromonad originations against the *nr* database using a highly sensitive threshold to detect even distantly homologous proteins. We have detected homologous sequences for 391 ancyromonad-conserved gene families in different prokaryotic proteomes (Figure 4). Most of the genes with homologues in bacteria had as best hits proteins from organisms belonging the phyla Pseudomonadota, Myxococcota, Actinomycetota and Planctomycetota (Figure 2a). Moreover, few orthogroups were homologous to proteins in Euryarchaeota.

Most of the putative LGTs were protein families with unknown functions as well as proteins involved in signal transduction (for example an abundant and highly duplicated gene family OG0004079 in ancyromonads of Von Willebrand Factor D And EGF Domain-Containing Proteins) (Figure 2b). Other proteins included Peptidoglycan-binding gene families, transporters and enzymes with diverse activities such as glycerophosphodiester-phosphodiesterases, thioesterases, a Isochorismatase, a fructofuranosidase, a carbonic anhydrase and proteins with predicted gluconolactonase activity among others.

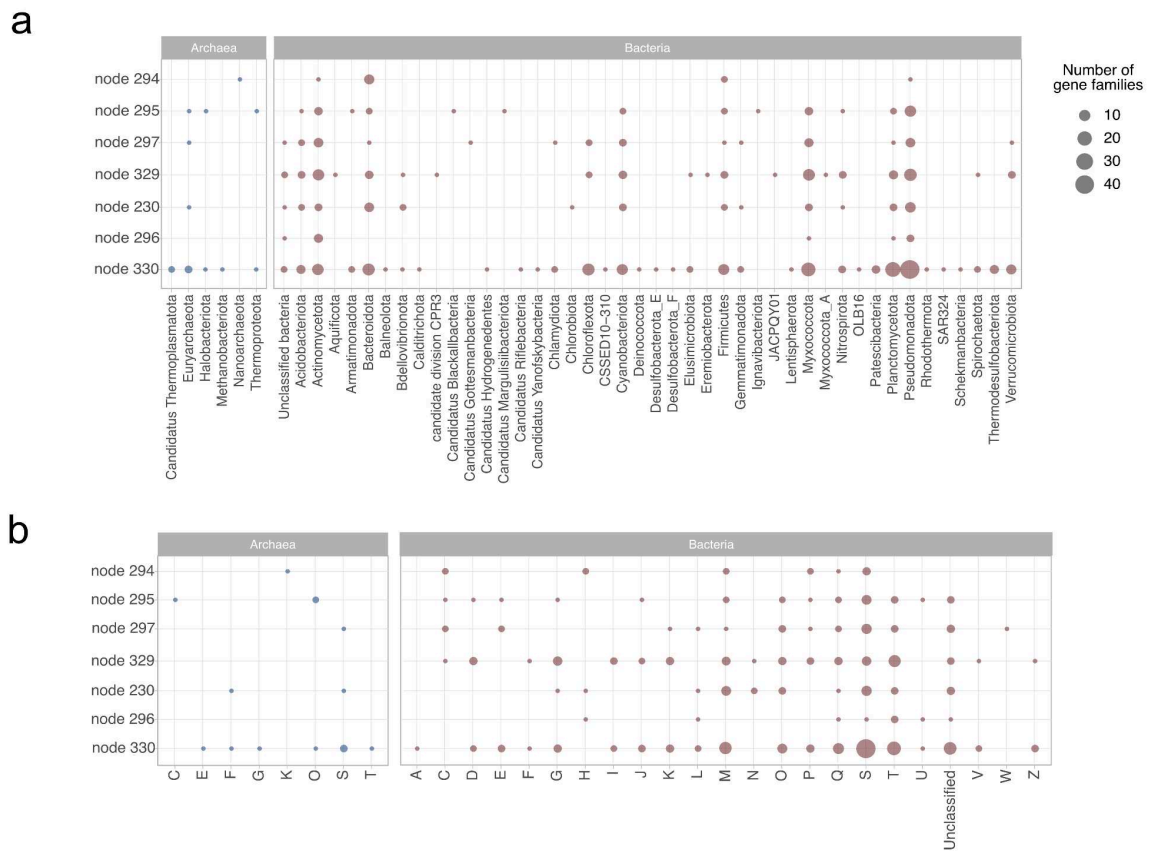

Figure 4. Prokaryote Lateral Gene Transfer footprint in ancyromonad genomes. a)Taxonomic affiliation of the putative LGT donors gain in each ancyromonad ancestor. Node 296: *Fabomonas* genus + *P. micra.* Node 230: *Fabomonas* genus. Node 329: *Ancyromonas* genus + *(N. silfraensis* + *(S. longa* + *N. limna).* Node 297: *Ancyromonas* genus. Node 295: *N. silfraensis* +*(S. longa* + *N. limna).* Node 294: *N. limna* + *S. longa.* b) COG categories associated with the putative LGT. A: RNA processing and modification, C: Energy production and conversion, D: Cell cycle control and chromosome partitioning, E: Amino acid transport and metabolism, F: Nucleotide transport and metabolism, G: Carbohydrate transport and metabolism, H: Coenzyme transport and metabolism, I:Lipid transport and metabolism, J:Translation, K: Transcription, L: Replication, recombination and repair, M: Cell wall/membrane/envelope biogenesis, N: Cell motility. O: Post-translational modification, protein turnover, chaperone functions, P: Inorganic ion transport and metabolism, Q: Secondary metabolites biosynthesis, transport and catabolism, S: Function unknown, T: Signal transduction mechanisms, U: Intracellular trafficking, secretion, and vesicular transport, V: Defense mechanisms, W: Extracellular structures, Z: Cytoskeleton.

Notably, we identified a gene family (OG0144218) homologous to the bacterial McrBC restriction endonuclease part of a prokaryotic Restriction Modification (RM) system acquired in the node 329. RM system components play important roles in prokaryotes against the insertion of foreign DNA into their genomes (Hampton, Watson, and Fineran 2020; Leão et al. 2023). Furthermore, several RM components have been

transferred to eukaryotes multiple times giving rise to diverse DNA modifications, among them the canonical 5' methyl cytosine epigenetic mark (de Mendoza, Lister, and Bogdanovic 2020; Arkhipova, Yushenova, and Rodriguez 2023). The low identity of the detected Ancytomycetota bacterial homologues (25-31%) suggest this represent a relatively ancient LGT, in agreement with the fact that we map it back to node 329.

V. Insights into the environmental adaptations and metabolic versatility of ancyromonads

Benthic sediments are characterized by the stratification of abiotic conditions such as oxygen and temperature as well as the availability of nutrients given the reduction of light and primary production (Osuna-Cruz et al. 2020). We identified some gene families putatively involved in environment sensing and response. These families included the aerotaxis receptor (OG0000242) harboring PAS domains which in bacteria is responsible for triggering the chemotaxis of cells towards air bubbles (Rebbapragada et al. 1997). Other sensing proteins highly conserved across ancyromonad species were YPD1 (OG0005290) involved in osmosensing, and the membrane lipid desaturase acyl-lipid omega-6 desaturase (OG0001085) tied to the response to low temperature.

Interestingly, all ancyromonads species encode nitrate/nitrite sensing proteins bearing the NIT (OG0000196, Pfam PF08376) and NarX-like (OG0010333, Pfam PF13675) domains. In bacteria, proteins harboring these domains trigger signaling pathways tied to nitrate assimilation, chemotaxis and enzyme activity in response to nitrate/nitrite concentration (Shu, Ulrich, and Zhulin 2003; Matilla et al. 2022). Furthermore, we detected nitrate transporters ABC type (OG0000152) and major facilitator superfamily (MFS) type (OG0001767 and OG0008653) encoded in ancyromonads.

Transport and nitrate assimilation is widely distributed across eukaryotes although seemingly restricted to autotroph and osmotroph lineages such as plants, fungi, diatoms and oomycetes (Ocaña-Pallarès et al. 2019). The phylogenetic histories of the gene set involved in nitrate assimilation of these organisms suggest that in fungi,

this metabolic capacity has been acquired through lateral transfer between eukaryotes (Ocaña-Pallarès et al. 2019). Considering the high conservation of nitrate sensors and transporters across ancyromonads being heterotrophs is striking. A possible explanations could be that nitrate is being imported into the cells of these species and being used by endosymbionts.

Moreover, *N. limna* encodes a respiratory nitrate reductase (jg14674.t1) bearing a Molybdopterin oxidoreductase domain (Pfam PF00384). The progressive reduction of nitrate takes place in the denitrification pathway, which is performed by a wide range of in facultative anaerobic organisms (Frostegård et al. 2022). The nitrate reductase of *N. limna* has as best hit a nitrate reductase from the bacteria *Arcicella rosea,* suggesting this gene was horizontally transferred from this species.

The use of nitrate as a terminal electron acceptor in the absence of oxygen has been previously observed in eukaryotes such as such as ciliates (Finlay, Span, and Harman 1983), fungi (Takaya et al. 2003), foraminifera of the order Rotaliida (Woehle et al. 2018; Glock et al. 2019) and benthic diatoms (Kamp et al. 2011, 2015). The genetic toolkit and completeness of these pathways is modular and variates widely among these organisms, resulting in different final products such as gaseous nitrogen compounds like nitrous oxide (N_2O) in the fungus *Fusarium oxysporum* or dinitrogen gas in the rotalid *Globolumilina* (N_2). Moreover, symbiotic bacteria are proposed to play a complementary role in rotalids by performing the steps for which foraminifera do not harbor the enzymes (Woehle et al. 2022). The conservation or Nap-like protein in *Nutomonas limna* suggest this species could perform some steps of the denitrification pathway.

More refined analysis on the genes of ancyromonads are needed to gain insight into the evolutionary history of these proteins in ancyromonads and the role of these organisms in the nitrogen cycle of their environments. The examination of such hypotheses will also further require biochemical experiments.

VI. Perspectives to improve our inferences of the early evolution of eukaryotes

The improved taxon sampling of this study gives us an unprecedented opportunity to explore the deep evolution of the eukaryotic domain as a whole. ALE reconciliation returns a probability presence for each of the analyzed gene families at a given internal node of the species phylogeny. Taking this probability into account, it is possible to explore the distribution, and thus loss and retention of gene families with a pre-LECA origin across our species phylogeny (Figure 5a). This distribution hints the retention of gene families anciently originated across the extant supergroups. The presence/absence patterns of those gene families in metamonada appear similar to those inDiphoda members. In contrast, the distribution of these genes in ancyromonads diverges from those found both in Diphoda and Opimoda, highlighting their deep divergence within eukaryotes. Malawimonads and Hemimastigophorids show the same behavior; however, the sampling for these lineages is still poor and the inclusion of more representatives could change this picture.

As a way to ponder our ALE results, we inferred gene family evolution using Dollo parsimony, in which gene families are assumed to have originated only once. These inferences suggest that 68% of the gene families distributed across ancyromonads were inferred already present in LECA (Figure 5b), while only few gene families in ancyromonads appear to have been originated at the base of Opimoda, and even less in HAMM and AMM. This points towards a lack of synapomorphies for these clades, suggesting that this part of the tree might be incorrectly resolved, or that these lineages diverged rapidly from one another. Interestingly, 2,387 out of 23,653 orthogroups inferred to be present in the LECA by Dollo parsimony were identified to be shared uniquely between ancyromonads and the parent node of Diaphoretickes and/or Jakobids and/or Discoba members. They could either represent previously unidentified LECA gene families or the result of ancient eukLGTs. This set of gene families is mostly associated with functions as intracellular trafficking, signal transduction, ubiquitin-based protein degradation, and, to a lesser extent, cytoskeletal and

RNA-processing genes (Supplementary Material. Figure S13). Studying eukLGT is particularly challenging when comparing lineages for which genomic data has variable quality, therefore a further analysis of gene family phylogenies as well as the inclusion of more taxa could help us to discern true eukLGTs.

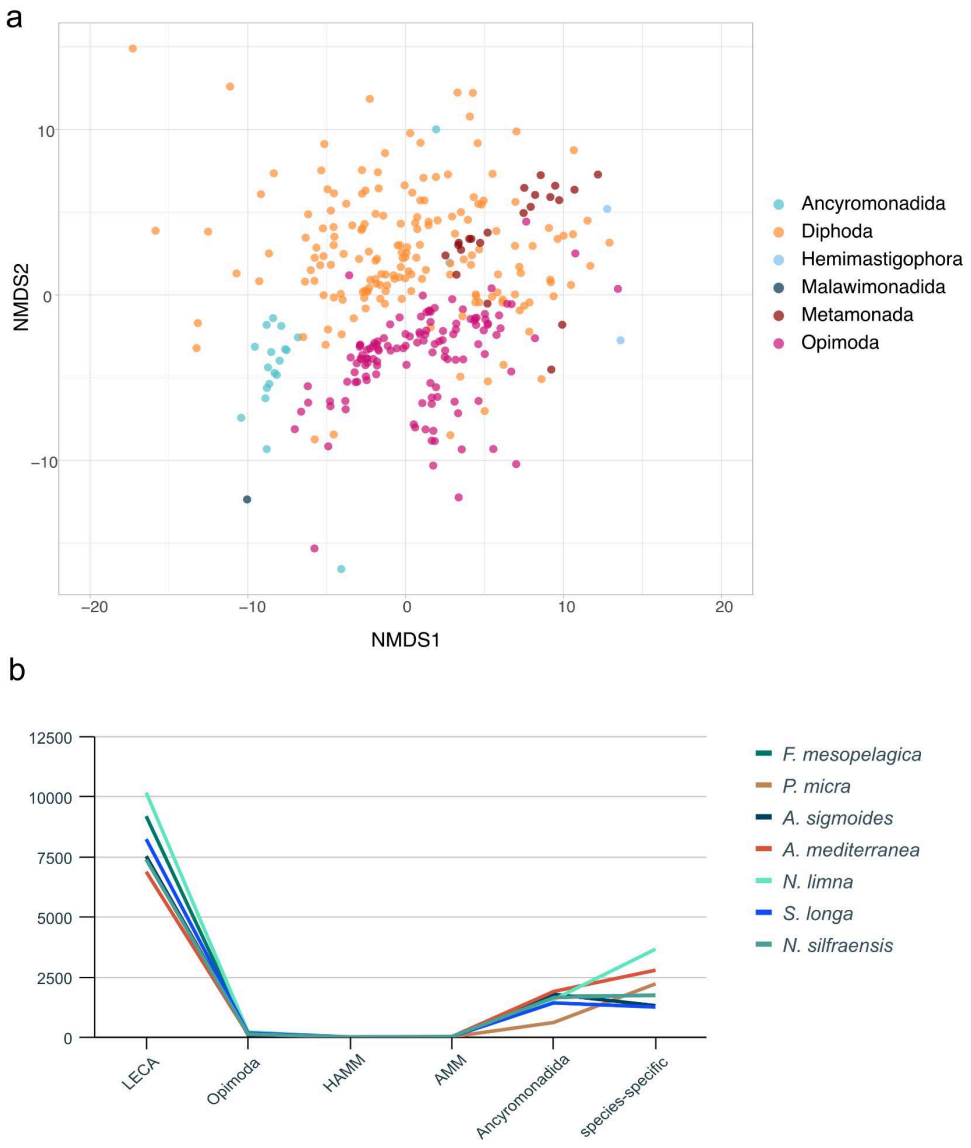

Figure 5. a) Non-metric multidimensional scaling (NMDS) analysis of the probability presence (PP) of gene families with an ancient origin. The (PP) from 8,500 gene families with ancient origin was retrieved from the ALE analysis for each terminal and internal node in the phylogeny. This matrix was then used to calculate the Euclidean distances between the nodes of the phylogeny, which were then ordinated using NMDS. b) Timing of emergence of ancyromonads gene families along the species tree as inferred by Dollo parsimony.

Conclusions

Ancyromonads have evolved diverse genome architectures and gene content. A high proportion of their protein coding sequences originated in their lineage and are restricted to them. These genes are of great biological interest because they could define exclusive ecological adaptations of ancyromonads to their environments. Different mechanisms of gene acquisition, expansion and selfish elements have contributed to structural changes and increased the genomic diversity across these species.

Furthermore, lateral gene transfers, from bacteria and archaea have allowed the acquisition of several signalization components, transporters, enzymes, and defense mechanisms into this lineage. These capacities could be associated to the adaptation of ancyromonads and the environmental versatility exhibited by some species. In addition of ancyromonads reveal the conservation of more than 2K gene families previously only found in species belonging to the Diphoda clade. These families could represent previously undescribed LECA genes retained uniquely in Diphoda and ancyromonads, providing a new window into the features of the Last Eukaryotic Common Ancestor. Alternatively, they could represent eukLTGs. Additional investigations are needed to disentangle those possibilities.

Altogether, our results improve the genomic sampling of deeply divergent lineages of protists and add new evidence that refine our understanding of the evolution of the gene content across eukaryotes. These results also raise questions about the universality of these evolutionary patterns, the relative importance of the mechanism for gene acquisition and common adaptations across other lineages of protists.

Finally, the evolutionary history of gene families with ancient origin suggests overall an ancient divergence of ancyromonads from major eukaryotic supergroups, however the lack of specific sinapomorphies of HAMM and AMM could suggest that the phylogenetic placement of ancyromonads within the eToL is even deeper or is associated to unsampled taxa. As more genomic studies of deeply divergent lineages become available, it is important to include ancyromonads alongside high quality marker data in order to finally decipher their enigmatic position as well as the backbone structure and root of the eToL.

Methods

Ancyromonad cell culturing , DNA and RNA isolation and sequencing.

Six ancyromonad species representing the main branches within the Ancyromonadida clade (Yubuki et al. 2023) were selected from our culture collection for genome sequencing: *Fabomonas mesopelagica, Striomonas longa, Planomonas micra, Nutomonas limna , Ancyromonas mediterranea* and *Nyramonas silfraensis.*

Culture flasks of 50mL were inoculated with 500 microliters of stock culture and 8mL of media containing 1:1000 of yeast extract and sterile sea or freshwater (depending on the origin of the strain) and maintained at 18ºC until reaching maximal cell density monitored by optical microscope. The cells were then harvested using a cell scraper and redistributed into two new culture flasks. This procedure was repeated until obtaining 7-10 flasks per species. Passes were done every three days until reaching an abundant cell density with the exception of *Nutomonas limna,* which was grown in a medium containing fresh water and 1:100 of soil extract prepared in culture plates of 12 wells, making weekly passes due to its slow growing rate.

For each species respectively, 7 to 10 total flasks were decantated from the media and washed gently with sterile media. Subsequently, cells were harvested from each flask and centrifuged at 4ºC at 1000g. Total DNA was isolated from the cell pellets using the PowerBioFilm DNA extraction kit (Qiagen). Additionally, a similar procedure was employed to obtain cells from well grown cultures to isolate total RNA with RNeasy micro kit (Qiagen). A DNA and a cDNA library a library of cDNA was prepared after polyA mRNA selection were sequenced at Eurofins Genomics (Germany) in an Illumina HiSeq 2500 sequencing platform using pair-end configuration (2x150) for each of the species.

In order to produce enough material for the nanopore sequencing and reduce the presence contaminant sequences we implemented a custom sequencing workflow that included the sorting of ancyromonad cell samples and Whole Genome Amplification protocol (WGA). We used SYBR green DNA dying to isolate 10 samples of ~200 ancyromonad cells for each of the species using a Flow cytometer–cell sorter BD FACSAria™ III. These samples were inspected at the optical microscope to verify the integrity of eukaryotic cells and were further used as input for two different WGA cycles with the REPLI-g and True-prime QIAGEN kits respectively. Both amplifications were sequenced using Illumina HiSeq (2x150 bp) as previously described. Based on our assessment of the Illumina data from the previous workflow (see details in Supplementary Material. *Genome sequencing and assembly*), genomic DNA of each species was amplified using the REPLI-G protocol from the remaining sorted cell samples of each of the species. The resulting DNA was purified and treated by a digestion of T7 Endonuclease to eliminate the branching DNA structures produced by the WGA. This material was then sequenced on a minIT model MK 1B Oxford Nanopore Technologies (ONT) with a SQK-LSK109 kit. The sequence basecalling of these datasets was performed with Guppy V.5 (Oxford Nanopore Technologies Ltd.) using the *super accuracy* model obtaining from 3.4 to 32.1 basecalled Gbp for each species (Supplementary Material. *Genome sequencing and assembly*).

Genome sequences assembly and curation

We combined the sequencing datasets generated from the bulk and cell sorting + WGA sequencing strategies (Supplementary Material. *Genome sequencing and assembly*). The approach yielding better results in terms of contiguity and completeness consisted in separately generating short reads and long reads genome assemblies using Spades v3.15.3 (Prjibelski et al. 2020) and metaFlye v2.9 (Kolmogorov et al. 2020) respectively, and them scaffold the short-read derived contigs using the long read assembly as backbone with RagTag v2 (Alonge et al. 2022). To correct potential assembly errors and refine the genomic sequences we performed an iterative cycle of consensus and polishing steps. Long reads were mapped against the genomic sequence using minimap2(H. Li 2018) and then a consensus sequence was generated using Racon v1.3.1 (Vaser et al. 2017) and Medaka v5 (Oxford Nanopore Technologies) three times. Subsequently, short reads were mapped against the genomic sequence with bwa-mem v0.7.15 (H. Li 2013) and Pilon v1.22 (Walker et al. 2014) for 3 times.

Contaminant contigs were identified based on the taxonomic affiliation of their proteins and their coverage biases in tw rounds. First, Prodigal V2.6.3 (Hyatt et al. 2010) was used to perform an initial protein prediction on the draft genome sequences. Protein predictions were then used as query to perform diamond (Buchfink, Reuter, and Drost 2021) searches against a custom database containing all the proteins of the Genome Taxonomy Database (GTDB)(Parks et al. 2022) and the Eukprot v3 (Richter et al. 2022) databases. A protein was classified as a contaminant if matched as best hit any protein from the GTDB database and had identity ≥ 80% and query coverage ≥ 50%. Contigs were discarded if 50% or more of their proteins were classified as contaminants if these proteins represented more than the 10% of the contig length. During the second round, after inferring the eukaryotic genes (see details below), the proteins of each genome were used as a query against the non-redundant nucleotide database (NCBI) and the GTDB-EukProt custom database.

Contigs with putative contaminants (classified with the same cutoffs as previously mentioned) were discarded if the contigs did not contain eukaryotic proteins identified based on the presence of introns and or Eukaryotic hits. BlobToolKit (Challis et al. 2019) was used on the final genomic sequences to verify the absence of contaminant sequences based on their taxonomic affiliation, GC content and coverage biases.

Transcriptomic datasets processing

The transcriptome read libraries of each species were assembled *de novo* using rnaSPAdes (Prjibelski et al. 2020; Bushmanova et al. 2019). BlobToolKit was used to screen the assembled transcripts for contamination as previously described. Contaminant transcripts were discarded, and reads libraries were mapped against the clean transcripts using STAR v2.5 (Dobin et al. 2013) to generate a contaminant free set of reads. Open Reading Frames (ORFs) sequences were extracted from the cleaned transcripts using Transdecoder v5 (Haas and Papanicolaou 2017) allowing a single protein prediction by transcript and using the universal genetic code. Subsequently, CD-HIT (Fu et al. 2012) was employed to cluster the proteins with a threshold of >=90% of identity and produce a non-redundant data set of proteins for each of the transcriptomes. These proteins were then pooled and used as a training set for the eukaryotic gene prediction pipeline.

Genome features prediction and annotation

A custom library of repetitive elements was generated for each ancyromonad genomic sequence by combining the results of RepeatModeler2 (Flynn et al. 2020), including the LTRharvest algorithm, and Transposon-PSI (Haas, n.d.). The repeat libraries were then clustered and annotated against the Dfam database using HMMer and the identified repetitive regions were sofmasked in the genomic sequences using RepeatMasker (Tarailo-Graovac and Chen 2009) prior protein coding genes prediction.

The protein coding gene prediction was conducted with Braker2 (Brůna et al. 2021; Hoff et al. 2016; Lomsadze et al. 2005), using protein hints and RNA-seq data as extrinsic evidence. With this purpose, we concatenated the predicted proteins from all decontaminated ancyromonads transcriptomes and the "protozoa" protein set from the OrthoDB protein database (Kriventseva et al. 2019). These proteins were then used as a reference for ProtHint (Brůna, Lomsadze, and Borodovsky 2020) to predict and score gene hints in the genome sequences. Additionally, the clean RNA-seq libraries were mapped against the genome sequences with STAR v2.5 (Dobin et al. 2013) with two passes to generate intron junction hints.

The completeness of decontaminated genomic and transcriptomic sequences as well as the inferred proteomes was evaluated using BUSCO v5.3.2 (Manni et al. 2021) and the eukaryota_obd10 database as a query. The percentage of transcriptomic reads from the cleaned libraries that mapped to the genome sequences was also used as a proxy to evaluate the completeness of the genomic sequences (Supplementary Material. *Genome quality assessment*).

Gene family reconstruction and annotation

To reconstruct gene families we generated a dataset of eukaryotic proteomes, including the inferred proteins of the six sequenced ancyromonads, the genome-derived proteome of *Ancyromonad sigmoides* (Blaz et al., *in preparation*), the inferred proteomes of *Mantamonas sphyraenae* and *Mantamonas vickermani* (Blaz et al. 2023) and the protein sets of 196 species representing a wide taxonomic diversity from *The Comparative Set* of the EukProt v3 (Richter et al. 2020). We employed OrthoFinder2 (Emms and Kelly 2019) using diamond (Buchfink, Reuter, and Drost 2021) with the *ultra-sensitive* mode and OrthoMCL (L. Li, Stoeckert, and Roos 2003) clustering with an inflation value of 1.2. In other to reduce the spurious clustering of proteins with shared

small domains only diamond hits with a minimum identity of 20% and a minimum in overlap over the hits of 40%, respectively were kept.

For each gene family, sequences were aligned using MAFFT v.7.427 (Katoh and Standley 2013) and the alignments were trimmed using BMGE (Criscuolo and Gribaldo 2010) using the default parameters. Maximum Likelihood phylogenies were then computed for the gene families containing at least four sequences (n=132,929) with IQtree v2.0.3 (Nguyen et al. 2015) with 1000 ultrafast bootstrap replicates (Hoang et al. 2018). These phylogenies were inferred under the mixture model of sequence evolution LG+C60+G except for the gene families of more than 1,500 sequences which were run using the LG+C20+G. The ultra-fast bootstrap samples were recovered from each phylogeny and additional mock phylogenies were generated for the gene families of two and tree sequences (n= 244,703) to be included in the phylogenetic reconciliation analysis (see below).

To assign functional information to the reconstructed gene families we generated individual annotations for each of the protein datasets using eggNOG-mapper v2 (Cantalapiedra et al. 2021) and InterproScan v5 (Blum et al. 2021; Jones et al. 2014). Gene families were then assigned several signatures (COG, Pfam, KO and Interpro accessions) when the signature was represented by at least 20% of the proteins belonging to the corresponding gene family.

Investigation of the patterns of genome evolution across a global eukaryotic phylogeny

To place the ancyromonad species within the eukaryotic tree of life (eToL), we used a custom approach to reconstruct a phylogeny for our species dataset based on the current eToL model solving the deepest polytomies of that tree. From the Orthofinder pipeline, we retrieved a dataset of 766 single copy gene families. These markers were

concatenated in a supermatrix and aligned using MAFFT and processed by BMGE to remove the sites with up to 90% of missing data, resulting in an alignment of 47,611 positions.

A Maximum Likelihood phylogenetic reconstruction was conducted using IQTREE v2.0.3 (Nguyen et al. 2015) with 1000 ultra-fast bootstrap replicates under the LG+C60+G sequence evolution model and the partially solved backbone of the eToL provided by the Eukprot v3 (Richter et al. 2022) was provided as a constraint tree (Supplementary Material. Evolutionary analyses). The resulting phylogeny was then rooted at the Opimoda-Diphoda split (Derelle et al. 2015). An alternative species tree topology was also tested by repositioning the clade Hemimastigophora within Diphoda, according to recent phylogenomic results (Tikhonenkov et al. 2022; Eglit et al. 2023), see comparison at Supplementary Material. Evolutionary analyses.

A phylogenetic reconciliation approach was implemented to infer the evolutionary history of the gene content across the diversification of ancyromonads from the eukaryotic supergroups in the eToL. With this purpose we used Amalgamated Likelihood Estimation (ALE) suite (Szöllõsi et al. 2013) that takes samples gene-family phylogenies accounting for their uncertainty as well as to take into account the proportion of expected missing data given to the estimated incompleteness of the compared datasets (calculated as the proportion of missing eukaryota_odb10 BUSCO markers). We employed ALEml_undated to estimate the DTL trends across each node of the species phylogeny by adding the frequencies of these events for all the gene families. Verticality by node was estimated as the proportion of singletons out of the total acquisitions of gene families at a particular node (Williams et al. 2023).

Moreover, the origin of a gene family was considered to be the node with the maximum value of Origination for each gene family; however gene families with lower values than 0.3 were considered to have an uncertain ancient origin. Families with an ancient origin for the NMDS analysis of Figure 5 correspond to this group, as well as families with predicted presence at the root. Phylogenies were processed with ETE3

(Huerta-Cepas, Serra, and Bork 2016) and visualized with ITOL (Letunic and Bork 2021) and ggtree (Yu 2022) R package (R Core Team 2021). An additional scenario has been tested by considering a species phylogeny in which Hemimastigophora branches within Diphoda (see Supplementary Material. Figure S15). The overall ancestral gene counts (copies) and trends of DTLO across the tree are maintained with few exceptions at early splits (see Supplementary Material Figures S16-17).

Analysis of Horizontal Gene Transfers from prokaryotic donors

We investigated the putative prokaryotic origin of Ancyromonad specific proteins that were distributed in at least two species. A representative sequence for each of these gene families meeting these conditions (n=3992 gene families) was screened for prokaryotic homology by diamond searches against the *nr*. Thresholds of e-value (max. 1xe-5) and identity percentage (min. 25%) were employed to retrieve up to 500 homologous proteins for which taxonomy was assigned.

Distribution of gene families with ancient origin

Considering the maximum origination value per gene family obtained from the ALEml results as previously described, we considered all the gene families with unclear ancient origin as well as gene families with copy number higher than 0 at the ancestral node of the species phylogeny. Then we extracted the presence probability of these 8,500 gene families across all the nodes of the species phylogeny (representing ancestral and extant proteomes) and used that matrix to calculate euclidean distances between the nodes. These distances were then projected in a Non-metric multidimensional scaling (NMDS) analysis and shown in Figure 5a. To contrast these results we employed Dollo Parsimony using COUNT which designates a binary presence/absence value to each of the gene families at each internal node of the phylogeny. This was used to estimate a minimum timing of emergence for each gene family distributed in ancyromonads as shown in Figure 5b.

Supplementary Material

https://docs.google.com/document/d/1Dn6KCyVruYIK6EIW23OS24PkSY-f8ymiWD7gZaHCXNQ/edit?usp=sharing

Acknowledgements

We thank Philippe Deschamps for his essential guidance and assistance in the management and processing the data of this project in our local computing cluster. We also thank Jon Jerlström-Hultqvist for his guidance on the assembly and annotation pipeline and Kelsey Williamson for sharing with us her insights about the structure of the eToL. Several analyses were conducted on the Core Cluster of the Institut Français de Bioinformatique (IFB) (ANR-11-INBS-0013), we thank the IFB support team for their help in software installation. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (ERC Starting Grant No 803151 to L.E. and ERC Advanced Grants No 322669 and 787904 to P.L.-G. and D.M.).

References

Alonge, Michael, Ludivine Lebeigle, Melanie Kirsche, Katie Jenike, Shujun Ou, Sergey Aganezov, Xingang Wang, Zachary B. Lippman, Michael C. Schatz, and Sebastian Soyk. 2022. "Automated Assembly Scaffolding Using RagTag Elevates a New Tomato System for High-Throughput Genome Editing." *Genome Biology* 23 (1): 258.

Alto, Laura Taylor, and Jonathan R. Terman. 2017. "Semaphorins and Their Signaling Mechanisms." *Methods in Molecular Biology* 1493: 1–25.

Andersson, Jan O., Asa M. Sjögren, Lesley A. M. Davis, T. Martin Embley, and Andrew J. Roger. 2003. "Phylogenetic Analyses of Diplomonad Genes Reveal Frequent Lateral Gene Transfers Affecting Eukaryotes." *Current Biology: CB* 13 (2): 94–104.

Arendsee, Zebulun W., Ling Li, and Eve Syrkin Wurtele. 2014. "Coming of Age: Orphan

Genes in Plants." *Trends in Plant Science* 19 (11): 698–708.

Arkhipova, Irina R., Irina A. Yushenova, and Fernando Rodriguez. 2023. "Shaping Eukaryotic Epigenetic Systems by Horizontal Gene Transfer." *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 45 (7): e2200232.

Atkins, M. S., A. G. McArthur, and A. P. Teske. 2000. "Ancyromonadida: A New Phylogenetic Lineage among the Protozoa Closely Related to the Common Ancestor of Metazoans, Fungi, and Choanoflagellates (Opisthokonta)." *Journal of Molecular Evolution* 51 (3): 278–85.

Becker, Burkhard, Kerstin Hoef-Emden, and Michael Melkonian. 2008. "Chlamydial Genes Shed Light on the Evolution of Photoautotrophic Eukaryotes." *BMC Evolutionary Biology* 8 (July): 203.

Bischoff, F. R., and H. Ponstingl. 1991. "Catalysis of Guanine Nucleotide Exchange on Ran by the Mitotic Regulator RCC1." *Nature* 354 (6348): 80–82.

Blaz, Jazmin, Luis Javier Galindo, Aaron A. Heiss, Harpreet Kaur, Guifré Torruella, Ashley Yang, L. Alexa Thompson, et al. 2023. "High Quality Genome and Transcriptome Data for Two New Species of Mantamonas, a Deep-Branching Eukaryote Clade." *bioRxiv*. https://doi.org/10.1101/2023.01.20.524885.

Brown, Matthew W., Aaron A. Heiss, Ryoma Kamikawa, Yuji Inagaki, Akinori Yabuki, Alexander K. Tice, Takashi Shiratori, et al. 2018. "Phylogenomics Places Orphan Protistan Lineages in a Novel Eukaryotic Super-Group." *Genome Biology and Evolution*. https://doi.org/10.1093/gbe/evy014.

Brůna, Tomáš, Katharina J. Hoff, Alexandre Lomsadze, Mario Stanke, and Mark Borodovsky. 2021. "BRAKER2: Automatic Eukaryotic Genome Annotation with GeneMark-EP+ and AUGUSTUS Supported by a Protein Database." *NAR Genomics and Bioinformatics* 3 (1): lqaa108.

Brůna, Tomáš, Alexandre Lomsadze, and Mark Borodovsky. 2020. "GeneMark-EP+: Eukaryotic Gene Prediction with Self-Training in the Space of Genes and Proteins." *NAR Genomics and Bioinformatics* 2 (2): lqaa026.

Buchfink, Benjamin, Klaus Reuter, and Hajk-Georg Drost. 2021. "Sensitive Protein Alignments at Tree-of-Life Scale Using DIAMOND." *Nature Methods* 18 (4): 366–68.

Burki, Fabien, Andrew J. Roger, Matthew W. Brown, and Alastair G. B. Simpson. 2020. "The New Tree of Eukaryotes." *Trends in Ecology & Evolution* 35 (1): 43–55.

Bushmanova, Elena, Dmitry Antipov, Alla Lapidus, and Andrey D. Prjibelski. 2019. "rnaSPAdes: A de Novo Transcriptome Assembler and Its Application to RNA-Seq Data." *GigaScience* 8 (9). https://doi.org/10.1093/gigascience/giz100.

Cavalier-Smith, Thomas, and Ema E-Y Chao. 2003. "Phylogeny of Choanozoa, Apusozoa, and Other Protozoa and Early Eukaryote Megaevolution." *Journal of Molecular Evolution* 56 (5): 540–63.

Challis, Richard, Edward Richards, Jeena Rajan, Guy Cochrane, and Mark Blaxter. 2019. "BlobToolKit – Interactive Quality Assessment of Genome Assemblies." *bioRxiv*. https://doi.org/10.1101/844852.

Collens, Adena B., and Laura A. Katz. 2021. "Opinion: Genetic Conflict With Mobile Elements Drives Eukaryotic Genome Evolution, and Perhaps Also Eukaryogenesis." *The Journal of Heredity* 112 (1): 140–44.

Criscuolo, Alexis, and Simonetta Gribaldo. 2010. "BMGE (Block Mapping and Gathering with Entropy): A New Software for Selection of Phylogenetic Informative Regions from Multiple Sequence Alignments." *BMC Evolutionary Biology* 10 (July): 210.

Derelle, Romain, Guifré Torruella, Vladimír Klimeš, Henner Brinkmann, Eunsoo Kim, Čestmír Vlček, B. Franz Lang, and Marek Eliáš. 2015. "Bacterial Proteins Pinpoint a Single Eukaryotic Root." *Proceedings of the National Academy of Sciences of the United States of America* 112 (7): E693–99.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. "STAR: Ultrafast Universal RNA-Seq Aligner." *Bioinformatics* 29 (1): 15–21.

Doolittle, W. F. 1998. "You Are What You Eat: A Gene Transfer Ratchet Could Account for Bacterial Genes in Eukaryotic Nuclear Genomes." *Trends in Genetics: TIG* 14 (8): 307–11.

Eglit, Yana, Takashi Shiratori, Jon Jerlström-Hultqvist, Kelsey Williamson, Andrew J. Roger, Ken-Ichiro Ishida, and Alastair G. B. Simpson. 2023. "Meteora Sporadica, a Protist with Incredible Cell Architecture, Is Related to Hemimastigophora." *bioRxiv*.

https://doi.org/10.1101/2023.08.13.553137.

Emms, David M., and Steven Kelly. 2019. “OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics.” *Genome Biology* 20 (1): 238.

Finlay, B. J., A. S. W. Span, and J. M. P. Harman. 1983. “Nitrate Respiration in Primitive Eukaryotes.” *Nature* 303 (5915): 333–36.

Flynn, Jullien M., Robert Hubley, Clément Goubert, Jeb Rosen, Andrew G. Clark, Cédric Feschotte, and Arian F. Smit. 2020. “RepeatModeler2 for Automated Genomic Discovery of Transposable Element Families.” *Proceedings of the National Academy of Sciences of the United States of America* 117 (17): 9451–57.

Fritz-Laylin, Lillian K., Simon E. Prochnik, Michael L. Ginger, Joel B. Dacks, Meredith L. Carpenter, Mark C. Field, Alan Kuo, et al. 2010. “The Genome of Naegleria Gruberi Illuminates Early Eukaryotic Versatility.” *Cell* 140 (5): 631–42.

Frostegård, Åsa, Silas H. W. Vick, Natalie Y. N. Lim, Lars R. Bakken, and James P. Shapleigh. 2022. “Linking Meta-Omics to the Kinetics of Denitrification Intermediates Reveals pH-Dependent Causes of N2O Emissions and Nitrite Accumulation in Soil.” *The ISME Journal* 16 (1): 26–37.

Fu, Limin, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. 2012. “CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data.” *Bioinformatics* 28 (23): 3150–52.

Galindo, Luis Javier, Purificación López-García, and David Moreira. 2022. “First Molecular Characterization of the Elusive Marine Protist Meteora Sporadica.” *Protist* 173 (4): 125896.

Glock, Nicolaas, Alexandra-Sophie Roy, Dennis Romero, Tanita Wein, Julia Weissenbach, Niels Peter Revsbech, Signe Høgslund, David Clemens, Stefan Sommer, and Tal Dagan. 2019. “Metabolic Preference of Nitrate over Oxygen as an Electron Acceptor in Foraminifera from the Peruvian Oxygen Minimum Zone.” *Proceedings of the National Academy of Sciences of the United States of America* 116 (8): 2860–65.

Haas, B. n.d. “TransposonPSI: An Application of PSI-Blast to Mine (retro-) Transposon ORF Homologies.” *Broad Institute, Cambridge, MA, USA*.

Haas, B., and A. Papanicolaou. 2017. “TransDecoder.”

Hampl, Vladimir, Jeffrey D. Silberman, Alexandra Stechmann, Sara Diaz-Triviño, Patricia J. Johnson, and Andrew J. Roger. 2008. "Genetic Evidence for a Mitochondriate Ancestry in the 'Amitochondriate' Flagellate Trimastix Pyriformis." *PloS One* 3 (1): e1383.

Hampton, Hannah G., Bridget N. J. Watson, and Peter C. Fineran. 2020. "The Arms Race between Bacteria and Their Phage Foes." *Nature* 577 (7790): 327–36.

Heiss, Aaron A., Martin Kolisko, Fleming Ekelund, Matthew W. Brown, Andrew J. Roger, and Alastair G. B. Simpson. 2018. "Combined Morphological and Phylogenomic Re-Examination of Malawimonads, a Critical Taxon for Inferring the Evolutionary History of Eukaryotes." *Royal Society Open Science* 5 (4): 171707.

Heiss, Aaron A., Giselle Walker, and Alastair G. B. Simpson. 2011. "The Ultrastructure of Ancyromonas, a Eukaryote without Supergroup Affinities." *Protist* 162 (3): 373–93.

Hoang, Diep Thi, Olga Chernomor, Arndt von Haeseler, Bui Quang Minh, and Le Sy Vinh. 2018. "UFBoot2: Improving the Ultrafast Bootstrap Approximation." *Molecular Biology and Evolution* 35 (2): 518–22.

Hoff, Katharina J., Simone Lange, Alexandre Lomsadze, Mark Borodovsky, and Mario Stanke. 2016. "BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS." *Bioinformatics* 32 (5): 767–69.

Hyatt, Doug, Gwo-Liang Chen, Philip F. Locascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. 2010. "Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification." *BMC Bioinformatics* 11 (March): 119.

Janouškovec, Jan, Denis V. Tikhonenkov, Fabien Burki, Alexis T. Howe, Forest L. Rohwer, Alexander P. Mylnikov, and Patrick J. Keeling. 2017. "A New Lineage of Eukaryotes Illuminates Early Mitochondrial Genome Reduction." *Current Biology: CB* 27 (23): 3717–24.e5.

Kamp, Anja, Dirk de Beer, Jana L. Nitsch, Gaute Lavik, and Peter Stief. 2011. "Diatoms Respire Nitrate to Survive Dark and Anoxic Conditions." *Proceedings of the National Academy of Sciences of the United States of America* 108 (14): 5649–54.

Kamp, Anja, Signe Høgslund, Nils Risgaard-Petersen, and Peter Stief. 2015. "Nitrate Storage and Dissimilatory Nitrate Reduction by Eukaryotic Microbes." *Frontiers in*

Microbiology 6 (December): 1492.

Katoh, Kazutaka, and Daron M. Standley. 2013. “MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.” *Molecular Biology and Evolution* 30 (4): 772–80.

Khalturin, Konstantin, Georg Hemmrich, Sebastian Fraune, René Augustin, and Thomas C. G. Bosch. 2009. “More than Just Orphans: Are Taxonomically-Restricted Genes Important in Evolution?” *Trends in Genetics: TIG* 25 (9): 404–13.

Kolmogorov, Mikhail, Derek M. Bickhart, Bahar Behsaz, Alexey Gurevich, Mikhail Rayko, Sung Bong Shin, Kristen Kuhn, et al. 2020. “metaFlye: Scalable Long-Read Metagenome Assembly Using Repeat Graphs.” *Nature Methods* 17 (11): 1103–10.

Koumandou, V. Lila, Bill Wickstead, Michael L. Ginger, Mark van der Giezen, Joel B. Dacks, and Mark C. Field. 2013. “Molecular Paleontology and Complexity in the Last Eukaryotic Common Ancestor.” *Critical Reviews in Biochemistry and Molecular Biology* 48 (4): 373–96.

Kriventseva, Evgenia V., Dmitry Kuznetsov, Fredrik Tegenfeldt, Mosè Manni, Renata Dias, Felipe A. Simão, and Evgeny M. Zdobnov. 2019. “OrthoDB v10: Sampling the Diversity of Animal, Plant, Fungal, Protist, Bacterial and Viral Genomes for Evolutionary and Functional Annotations of Orthologs.” *Nucleic Acids Research* 47 (D1): D807–11.

Lax, Gordon, Yana Eglit, Laura Eme, Erin M. Bertrand, Andrew J. Roger, and Alastair G. B. Simpson. 2018. “Hemimastigophora Is a Novel Supra-Kingdom-Level Lineage of Eukaryotes.” *Nature* 564 (7736): 410–14.

Leão, Pedro, Mary E. Little, Kathryn E. Appler, Daphne Sahaya, Emily Aguilar-Pine, Kathryn Currie, Ilya J. Finkelstein, Valerie De Anda, and Brett J. Baker. 2023. “Asgard Archaea Defense Systems and Their Roles in the Origin of Immunity in Eukaryotes.” *bioRxiv*. https://doi.org/10.1101/2023.09.13.557551.

Leger, Michelle M., Martin Kolisko, Ryoma Kamikawa, Courtney W. Stairs, Keitaro Kume, Ivan Čepička, Jeffrey D. Silberman, et al. 2017. “Organelles That Illuminate the Origins of Trichomonas Hydrogenosomes and Giardia Mitosomes.” *Nature Ecology & Evolution* 1 (4): 0092.

Li, Hao-Dong, Wen-Xin Liu, and Marek Michalak. 2011. “Enhanced Clathrin-Dependent Endocytosis in the Absence of Calnexin.” *PloS One* 6 (7): e21678.

Li, Heng. 2013. “Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM.” *arXiv [q-bio.GN]*. arXiv. http://arxiv.org/abs/1303.3997.

Li, Heng. 2018. “Minimap2: Pairwise Alignment for Nucleotide Sequences.” *Bioinformatics* 34 (18): 3094–3100.

Li, Li, Christian J. Stoeckert Jr, and David S. Roos. 2003. “OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes.” *Genome Research* 13 (9): 2178–89.

Lomsadze, Alexandre, Vardges Ter-Hovhannisyan, Yury O. Chernoff, and Mark Borodovsky. 2005. “Gene Identification in Novel Eukaryotic Genomes by Self-Training Algorithm.” *Nucleic Acids Research* 33 (20): 6494–6506.

López-García, Purificación, Laura Eme, and David Moreira. 2017. “Symbiosis in Eukaryotic Evolution.” *Journal of Theoretical Biology* 434 (December): 20–33.

Manni, Mosè, Matthew R. Berkeley, Mathieu Seppey, Felipe A. Simão, and Evgeny M. Zdobnov. 2021. “BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes.” *Molecular Biology and Evolution* 38 (10): 4647–54.

Matilla, Miguel A., Félix Velando, David Martín-Mora, Elizabet Monteagudo-Cascales, and Tino Krell. 2022. “A Catalogue of Signal Molecules That Interact with Sensor Kinases, Chemoreceptors and Transcriptional Regulators.” *FEMS Microbiology Reviews* 46 (1). https://doi.org/10.1093/femsre/fuab043.

Mendoza, Alex de, Ryan Lister, and Ozren Bogdanovic. 2020. “Evolution of DNA Methylome Diversity in Eukaryotes.” *Journal of Molecular Biology* 432 (6): 1687–1705.

Nguyen, Lam-Tung, Heiko A. Schmidt, Arndt von Haeseler, and Bui Quang Minh. 2015. “IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies.” *Molecular Biology and Evolution* 32 (1): 268–74.

Ocaña-Pallarès, Eduard, Sebastián R. Najle, Claudio Scazzocchio, and Iñaki Ruiz-Trillo. 2019. “Reticulate Evolution in Eukaryotes: Origin and Evolution of the Nitrate Assimilation Pathway.” *PLoS Genetics* 15 (2): e1007986.

O’kelly, Charles J., and Thomas A. Nerad. 1999. “Malawimonas Jakobiformis N. Gen., N.

Sp. (Malawimonadidae N. Fam.): A Jakoba-like Heterotrophic Nanoflagellate with Discoidal Mitochondrial Cristae." *The Journal of Eukaryotic Microbiology* 46 (5): 522–31.

Osuna-Cruz, Cristina Maria, Gust Bilcke, Emmelien Vancaester, Sam De Decker, Atle M. Bones, Per Winge, Nicole Poulsen, et al. 2020. "Author Correction: The Seminavis Robusta Genome Provides Insights into the Evolutionary Adaptations of Benthic Diatoms." *Nature Communications* 11 (1): 5331.

Paps, Jordi, Luis A. Medina-Chacón, Wyth Marshall, Hiroshi Suga, and Iñaki Ruiz-Trillo. 2013. "Molecular Phylogeny of Unikonts: New Insights into the Position of Apusomonads and Ancyromonads and the Internal Relationships of Opisthokonts." *Protist* 164 (1): 2–12.

Parks, Donovan H., Maria Chuvochina, Christian Rinke, Aaron J. Mussig, Pierre-Alain Chaumeil, and Philip Hugenholtz. 2022. "GTDB: An Ongoing Census of Bacterial and Archaeal Diversity through a Phylogenetically Consistent, Rank Normalized and Complete Genome-Based Taxonomy." *Nucleic Acids Research* 50 (D1): D785–94.

Prjibelski, Andrey, Dmitry Antipov, Dmitry Meleshko, Alla Lapidus, and Anton Korobeynikov. 2020. "Using SPAdes De Novo Assembler." *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]* 70 (1): e102.

Rebbapragada, A., M. S. Johnson, G. P. Harding, A. J. Zuccarelli, H. M. Fletcher, I. B. Zhulin, and B. L. Taylor. 1997. "The Aer Protein and the Serine Chemoreceptor Tsr Independently Sense Intracellular Energy Levels and Transduce Oxygen, Redox, and Energy Signals for Escherichia Coli Behavior." *Proceedings of the National Academy of Sciences of the United States of America* 94 (20): 10541–46.

Reider, Amanda, Sarah L. Barker, Sanjay K. Mishra, Young Jun Im, Lymarie Maldonado-Báez, James H. Hurley, Linton M. Traub, and Beverly Wendland. 2009. "Syp1 Is a Conserved Endocytic Adaptor That Contains Domains Involved in Cargo Selection and Membrane Tubulation." *The EMBO Journal* 28 (20): 3103–16.

Richards, Thomas A., and Thomas Cavalier-Smith. 2005. "Myosin Domain Evolution and the Primary Divergence of Eukaryotes." *Nature* 436 (7054): 1113–18.

Richter, Daniel J., Cédric Berney, Jürgen F. H. Strassert, Yu-Ping Poh, Emily K. Herman, Sergio A. Muñoz-Gómez, Jeremy G. Wideman, Fabien Burki, and Colomban de

Vargas. 2020. “EukProt: A Database of Genome-Scale Predicted Proteins across the Diversity of Eukaryotes.” *bioRxiv*. bioRxiv. https://doi.org/10.1101/2020.06.30.180687.

Richter, Daniel J., Cédric Berney, Jürgen F. H. Strassert, Yu-Ping Poh, Emily K. Herman, Sergio A. Muñoz-Gómez, Jeremy G. Wideman, Fabien Burki, and Colomban de Vargas. 2022. “EukProt: A Database of Genome-Scale Predicted Proteins across the Diversity of Eukaryotes.” *Peer Community Journal* 2 (e56). https://doi.org/10.24072/pcjournal.173.

Saville-Kent, William. 1882. *A Manual of the Infusoria: Plates*.

Schaack, Sarah, Clément Gilbert, and Cédric Feschotte. 2010. “Promiscuous DNA: Horizontal Transfer of Transposable Elements and Why It Matters for Eukaryotic Evolution.” *Trends in Ecology & Evolution* 25 (9): 537–46.

Schön, Max E., Vasily V. Zlatogursky, Rohan P. Singh, Camille Poirier, Susanne Wilken, Varsha Mathur, Jürgen F. H. Strassert, et al. 2021. “Single Cell Genomics Reveals Plastid-Lacking Picozoa Are Close Relatives of Red Algae.” *Nature Communications* 12 (1): 6651.

Shields, Christina M., Rachel Taylor, Tara Nazarenus, Joseph Cheatle, Ann Hou, Audrey Tapprich, Alexis Haifley, and Audrey L. Atkin. 2003. “Saccharomyces Cerevisiae Ats1p Interacts with Nap1p, a Cytoplasmic Protein That Controls Bud Morphogenesis.” *Current Genetics* 44 (4): 184–94.

Shu, Chengyi J., Luke E. Ulrich, and Igor B. Zhulin. 2003. “The NIT Domain: A Predicted Nitrate-Responsive Module in Bacterial Sensory Receptors.” *Trends in Biochemical Sciences* 28 (3): 121–24.

Speijer, Dave, Julius Lukeš, and Marek Eliáš. 2015. “Sex Is a Ubiquitous, Ancient, and Inherent Attribute of Eukaryotic Life.” *Proceedings of the National Academy of Sciences of the United States of America* 112 (29): 8827–34.

Stairs, Courtney W., Petr Táborský, Eric D. Salomaki, Martin Kolisko, Tomáš Pánek, Laura Eme, Miluše Hradilová, et al. 2021. “Anaeramoebae Are a Divergent Lineage of Eukaryotes That Shed Light on the Transition from Anaerobic Mitochondria to Hydrogenosomes.” *Current Biology: CB* 31 (24): 5605–12.e5.

Tachezy, Jan. 2019. *Hydrogenosomes and Mitosomes: Mitochondria of Anaerobic Eukaryotes*. Springer.

Takaya, Naoki, Seigo Kuwazaki, Yoshiaki Adachi, Sawako Suzuki, Tomoko Kikuchi, Hiro Nakamura, Yoshitsugu Shiro, and Hirofumi Shoun. 2003. "Hybrid Respiration in the Denitrifying Mitochondria of Fusarium Oxysporum." *Journal of Biochemistry* 133 (4): 461–65.

Takenawa, Tadaomi. 2010. "Phosphoinositide-Binding Interface Proteins Involved in Shaping Cell Membranes." *Proceedings of the Japan Academy. Series B, Physical and Biological Sciences* 86 (5): 509–23.

Tarailo-Graovac, Maja, and Nansheng Chen. 2009. "Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences." *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]* Chapter 4 (March): 4.10.1–4.10.14.

Tautz, Diethard, and Tomislav Domazet-Lošo. 2011. "The Evolutionary Origin of Orphan Genes." *Nature Reviews. Genetics* 12 (10): 692–702.

Tikhonenkov, Denis V., Kirill V. Mikhailov, Ryan M. R. Gawryluk, Artem O. Belyaev, Varsha Mathur, Sergey A. Karpov, Dmitry G. Zagumyonnyi, et al. 2022. "Microbial Predators Form a New Supergroup of Eukaryotes." *Nature* 612 (7941): 714–19.

Torruella, Guifré, David Moreira, and Purificación López-García. 2017. "Phylogenetic and Ecological Diversity of Apusomonads, a Lineage of Deep-Branching Eukaryotes." *Environmental Microbiology Reports* 9 (2): 113–19.

Van Etten, Julia, and Debashish Bhattacharya. 2020. "Horizontal Gene Transfer in Eukaryotes: Not If, but How Much?" *Trends in Genetics: TIG* 36 (12): 915–25.

Vaser, Robert, Ivan Sović, Niranjan Nagarajan, and Mile Šikić. 2017. "Fast and Accurate de Novo Genome Assembly from Long Uncorrected Reads." *Genome Research* 27 (5): 737–46.

Vosseberg, Julian, Jolien J. E. van Hooff, Marina Marcet-Houben, Anne van Vlimmeren, Leny M. van Wijk, Toni Gabaldón, and Berend Snel. 2021. "Timing the Origin of Eukaryotic Cellular Complexity with Ancient Duplications." *Nature Ecology & Evolution* 5 (1): 92–100.

Walker, Bruce J., Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel,

Sharadha Sakthikumar, Christina A. Cuomo, et al. 2014. “Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement.” *PloS One* 9 (11): e112963.

Weiner, Agnes K. M., Mario A. Cerón-Romero, Ying Yan, and Laura A. Katz. 2020. “Phylogenomics of the Epigenetic Toolkit Reveals Punctate Retention of Genes across Eukaryotes.” *Genome Biology and Evolution* 12 (12): 2196–2210.

Williams, Shelby K., Jon Jerlström Hultqvist, Yana Eglit, Dayana E. Salas-Leiva, Bruce Curtis, Russell Orr, Courtney W. Stairs, Alastair G. B. Simpson, and Andrew J. Roger. 2023. “Extreme Mitochondrial Reduction in a Novel Group of Free-Living Metamonads.” *bioRxiv*. https://doi.org/10.1101/2023.05.03.539051.

Williams, Tom A., Adrián A. Davín, Benoit Morel, Lénárd L. Szánthó, Anja Spang, Alexandros Stamatakis, Philip Hugenholtz, and Gergely J. Szöllősi. 2023. “Parameter Estimation and Species Tree Rooting Using ALE and GeneRax.” *Genome Biology and Evolution* 15 (7). https://doi.org/10.1093/gbe/evad134.

Woehle, Christian, Alexandra-Sophie Roy, Nicolaas Glock, Jan Michels, Tanita Wein, Julia Weissenbach, Dennis Romero, et al. 2022. “Denitrification in Foraminifera Has an Ancient Origin and Is Complemented by Associated Bacteria.” *Proceedings of the National Academy of Sciences of the United States of America* 119 (25): e2200198119.

Woehle, Christian, Alexandra-Sophie Roy, Nicolaas Glock, Tanita Wein, Julia Weissenbach, Philip Rosenstiel, Claas Hiebenthal, Jan Michels, Joachim Schönfeld, and Tal Dagan. 2018. “A Novel Eukaryotic Denitrification Pathway in Foraminifera.” *Current Biology: CB* 28 (16): 2536–43.e5.

Yubuki, Naoji, and Brian S. Leander. 2013. “Evolution of Microtubule Organizing Centers across the Tree of Eukaryotes.” *The Plant Journal: For Cell and Molecular Biology,* 75 (2): 230–44.

Yubuki, Naoji, Guifré Torruella, Luis Javier Galindo, Aaron A. Heiss, Maria Cristina Ciobanu, Takashi Shiratori, Ken-Ichiro Ishida, et al. 2023. “Molecular and Morphological Characterization of Four New Ancyromonad Genera and Proposal for an Updated Taxonomy of the Ancyromonadida.” *The Journal of Eukaryotic Microbiology*, August, e12997.

7. Exploring the genome regulation systems of ancyromonads and their role in shifting environments

Context and results summary

Epigenetic mechanisms are key for the emergence of the complexity and physiological plasticity that we observe in eukaryotes. In particular, microbial eukaryotes employ epigenetic systems for their responses to environmental signals and they might be also an important speciation driver in the long term (Weiner and Katz, 2021). However, epigenetics and genome expression has been scarcely explored in non parasitic protists being the availability of genomic data an important bottleneck. Ancyromonads, a diverse group of benthic flagellates, represent fascinating study models because of their key deep position within the tree of eukaryotes.

In the previous chapters, we have observed that although morphologically similar, ancyromonads exhibit a wide genomic diversity, for which we can hypothesize that they employ different molecular strategies in their adaptation to their environments. In this last part of the project, harnessing the genomic data and the cultures of these organisms from our previous efforts, we explored the potential role of DNA methylation in the genome regulation of ancyromonads.

We used Bisulfite sequencing to have a first glance of the 5-methylcitosine (5mC) marks across the seven species for which we have sequenced the genome growing under control conditions. All the ancyromonads species that we analyzed displayed globally low 5mC levels (<2%). More detailed analysis of different classes of repeats and genes is still pending, however a preliminary analysis of methylation in the transposable elements and protein coding genes of *Ancyromonas sigmoides* revealed methylated sites present a higher variation within TEs and that there is a slightly higher methylation in gene bodies than in adjacent regions. Furthermore, we looked for conserved domains associated to DNA methyltransferases (DNMTs) in the genomes of these organisms. We

detected 5mC specific DNMTs in only tree of the four compared species. Furthermore, we detected proteins bearing domains of 4mC and 6mA DNA modification proteins.

In addition we designed an experiment to study the gene expression profiles of the type species *Ancyromonas sigmoides* in response to different growing conditions. These conditions included shifts in the salinity, temperature and oxygen. We identified around 6K differentially expressed genes in at least one condition of our experiment.

Notably, genes with integrase domains and putatively related to gypsy-type transposons were differentially expressed across multiple conditions, suggesting a potential role of these elements in genome regulation of this species. This manuscript is in preparation as the analysis of the methylation and expression data are still ongoing.

(Epi)genomic diversity in ancyromonads, a benthic and early divergent lineage of microeukaryotes.

Author list:

Jazmin Blaz, Naoji Yubuki, Maria Ciobanu, Luis J. Galindo, Guifré Torruella, Puri López-García, Aaron Heiss, Eunsoo Kim, David Moreira and Laura Eme

Abstract

Research on epigenetics and genome regulation across eukaryotic diversity is extremely limited, and shows wide variation, even between model organisms. Ancyromonads, a diverse group of benthic flagellates, present an interesting study model because of their deep position in the eukaryotic tree of life, suggesting they might have kept features that are more like the ones of the Last Eukaryotic Common Ancestor than more derived lineages such as animals and plants. In this study, we investigated DNA methylation diversity across seven ancyromonads species, making it one of the first works on this topic in free-living protists. Our results showed that ancyromonads generally exhibit low methylation levels (<2%) in CpG, CpHpG, and CpHpH contexts, a pattern fairly distinct to the ones described across eukaryotes so far. The type species of the clade, *Ancyromonas sigmoides* displayed different methylation profiles in the body of genes and transposable elements. Additionally, we conducted an experiment to study the gene expression profiles of *Ancyromonas sigmoides*, the type species of the clade, under shifts in salinity, temperature, and oxygen levels. We identified approximately 6,000 genes differentially expressed across different conditions, out of 11,138 genes. Notably, genes with integrase domains putatively belonging to gypsy-type retrotransposons showed differential expression across multiple conditions, hinting at a potential role of these elements in the genome regulation of this species. This study is ongoing, and currently in preparation, nevertheless, the preliminary findings shed light on the epigenetic diversity and genomic dynamics within this enigmatic group of protists.

Introduction

The epigenetic modifications across a genome (the epigenome) impacts the transcriptional activity of the genome through their interplay with regulation systems (Katz 2006). The role of the epigenome in the emergence of physiological diversity and complexity in models of plants, fungi, and animals is widely established (Lowdon, Jang, and Wang 2016; Madhani 2021; Lloyd and Lister 2022). Recent studies have pointed to the ubiquity of these mechanisms across diverse eukaryotes (Weiner et al. 2020; Grau-Bové et al. 2022; Weiner and Katz 2021). In particular, recent investigations in microbial eukaryotes have unveiled a remarkable diversity of 5-methyl-cytosine (5mC) DNA methylation landscapes and pathways (de Mendoza et al. 2018; de Mendoza, Lister, and Bogdanovic 2020; Hoguin et al. 2023). However, the exploration of 5mC patterns and their associated pathways across diverse lineages in the eukaryotic tree of life remains limited.

The orphan eukaryotic clade Ancyromonadida, holds a pivotal position within the global eukaryotic phylogeny and therefore represents an interesting model to expand our understanding of the true diversity of epigenetic mechanisms across eukaryotes. Ancyromonads are phagotrophic and graze on prokaryotes that live in their environments (Heiss, Walker, and Simpson 2011; Glücksman, Snell, and Cavalier-Smith 2013). These organisms have a cosmopolitan distribution and have been previously isolated from benthic sediments and soil environments (Tikhonenkov, Mazei, and Mylnikov 2006; Yubuki et al. 2023). These environments are characterized by a conspicuous stratification of available resources. Molecular adaptations to such conditions are starting to be understood (Osuna-Cruz et al. 2020) but remain poorly explored in across protist diversity.

Here, we generated DNA methylation data for seven species of ancyromonads in order to analyze 5mC patterns across their genomes. Since the role of DNA methylation on transcription levels is still poorly understood, we performed an experiment to study the changes in the methylation landscapes and gene expression profiles of *Ancyromonas sigmoides,* the type species of the clade, under five environmental conditions. This first offers a better understanding of the effect of methylation on transcription, but also

provides insights into the putative functions of the currently many unannotated genes in ancyromonads.

Preliminary results and discussion

DNA methylation diversity across ancyromonads

The genome-wide 5mC landscapes of ancyromonads species were explored using Whole Genome Bisulfite (WGB) sequencing. Preliminary analysis of this data showed that methylome landscapes were characterized by an overall low level of methylation distributed in CpG, PpHpG, and CpHpH contexts (where H can be Thymine/Adenine/Cytosine) (Figure 1).

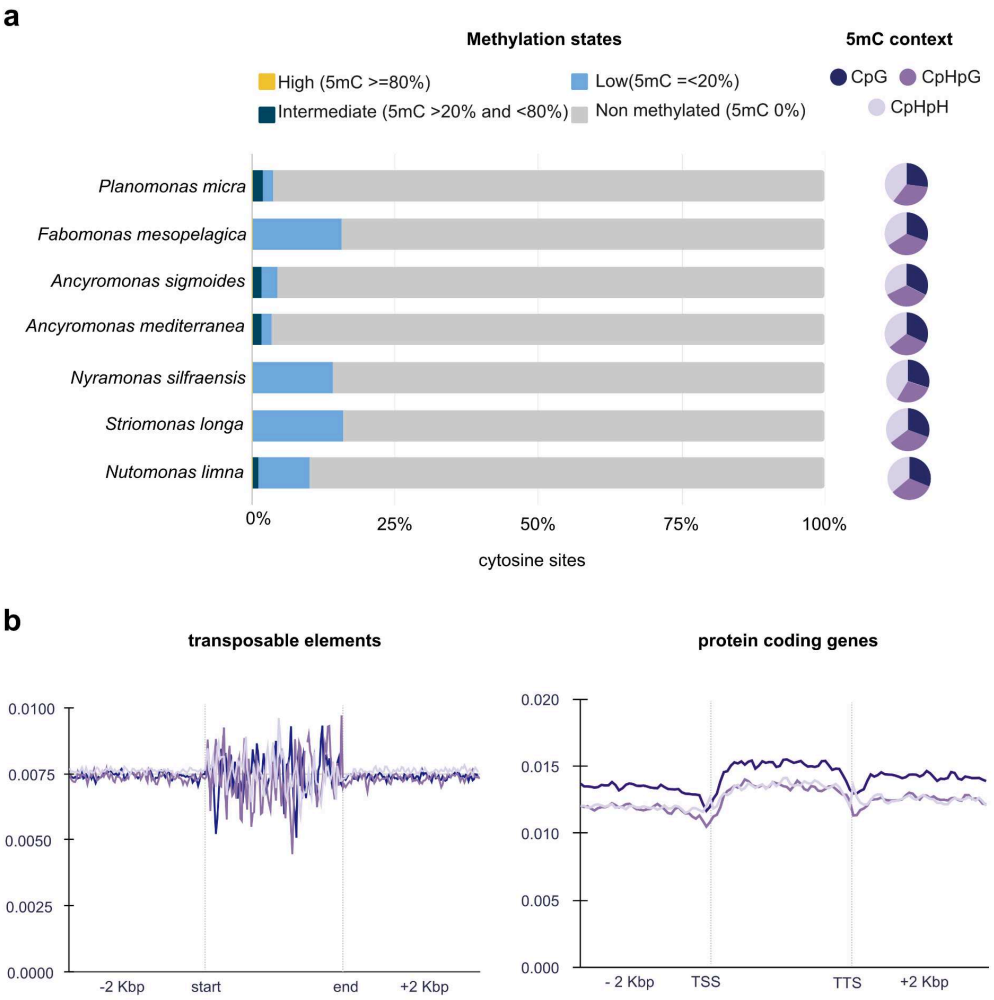

Figure 1. Genome-wide methylation in ancyromonads. a) Global levels of DNA methylation across genome cytosine sites of seven ancyromonad species and the proportion of contexts of methylated sites. b) Average methylation levels across transposable elements and protein-coding genes inferred in the genome of *Ancyromonas sigmoides*.

From all the cytosine sites with a coverage of at least 10x, we calculated the level of methylation by dividing the number of 5mC at that site on reads mapped to any of the strands by the sum of methylated and non-methylated cytosines at the same site in reads mapping to either strand. We arbitrarily classified sites as weakly methylated (<20% of reads at that site were methylated), intermediate (20-80% 5mC levels) and highly methylated (>80%) (Figure 1a). We observed that the methylated sites across ancyromonad genomes were mostly weakly methylated, with only a few sites classified as intermediate or extremely few as highly methylated. Although found at low levels, 5mC marks have been suggested to regulate tissue differentiation throughout the life-cycle of *S. japonica (Fan et al. 2020)* and to be associated with the transcriptional silencing of TEs (and, therefore, their accumulation in genomes) in several *Neurospora* species (Fan et al. 2020; Hosseini et al. 2020).

Among the predicted genomic features of *Ancyromonas sigmoides*, TEs and genes displayed different methylation profiles (Figure 1b). Protein coding genes display a higher methylation level in CpG contexts and a slight decrease of methylation levels around the transcription start and termination sites. Similar profiles have been observed in other species referred to as “gene body methylation” (P. A. Jones 2012). Although the role of gene body methylation remains unclear, it is widely conserved across diverse eukaryotes investigated so far (Bewick and Schmitz 2017; Veluchamy et al. 2013; Neri et al. 2017). In contrast, the methylation levels are much more variable and show a much broader range of methylation levels across TEs than in their adjacent regions, irrespective of the sequence context, and are overall lower than gene body methylation. A more detailed analysis of the methylation levels of different classes of TEs might shed light into this variation. TE silencing by methylation is a prevalent feature of the methylation landscape across diverse eukaryotes (Schmitz, Lewis, and Goll 2019). We will investigate in the near future the transcription level of these TEs to see if they indicate a similar role for TE methylation in *A. sigmoides*.

Moreover, the eukaryotic writers of 5mC marks are a family of proteins known as DNA methyltransferases (DNMTs), for which several homologues are differentially

distributed across eukaryotes (de Mendoza, Lister, and Bogdanovic 2020; Hoguin et al. 2023). DNMT proteins share a homologous 5mC MTase domain (Lyko 2018), which we identified in the proteomes of three of the seven ancyromonad species (Table 1). Non-significant hits were also detected in the other genomes. The lack of clear homologues of these proteins is striking given that all the genomes exhibited DNA methylation to some extent. These proteins could have been lost or evolved beyond recognition in certain ancyromonads species; we need to investigate this further, for example searching for them using tblastn. It is also possible that some of the current gene predictions of ancyromonads are inaccurate.

Table 1. Hidden Markov Model searches on the PF00145 profile (C-5 cytosine-specific DNA methylase) using hmmsearch against ancyromonads predicted proteins (*e-value < 0.1).

Genome	target gene	E-value	score	bias
F. mesopelagica	jg4792.t1	0.28	10	0
F. mesopelagica	jg8454.t1	0.56	9	0.1
P. micra	jg5067.t1	0.37	9.4	0
P. micra	jg7029.t1	0.39	9.3	0
A. sigmoides	jg338.t1	0.21	10.2	0.2
A. mediterranea	jg8386.t1	2.60E-13*	49.5	0
A. mediterranea	jg7635.t1	0.63	8.7	1.8
A. mediterranea	jg7747.t1	0.84	8.3	0
A. mediterranea	jg9922.t1	0.86	8.3	1.1
A. mediterranea	jg10678.t1	0.93	8.2	0.8
N. limna	jg14665.t1	1.30E-93*	313.9	0
N. limna	jg7806.t1	3.60E-06*	26.4	0
N. limna	jg13369.t1	0.094*	11.9	0
N. limna	jg13259.t1	0.21	10.7	0.4
N. limna	jg9611.t1	0.29	10.3	0
N. limna	jg16163.t1	0.59	9.3	0
S. longa	jg11205.t1	0.051*	12.3	0.1
S. longa	jg11020.t1	0.58	8.8	0
S. longa	jg72.t1	0.76	8.4	0.2
N. silfraensis	jg7680.t1	0.11	11.3	0.1
N. silfraensis	jg4495.t1	0.5	9.1	2.1

Other proteins with domain architecture related to DNMTs were also found. For example *Fabomonas mesopelagica* encodes two proteins (jg125.t1 and jg508.t19) bearing the cytosine-specific DNA methyltransferase replication foci domain (PF12047). These proteins have a high identity to DNMT1 proteins, whose most conserved activity is the maintenance of methylation in CpG contexts acting on hemimethylated sites (Svedružić 2011). These proteins in *F. mesopelagica* were predicted to be localized in the nucleus, however none of them were found to bear a detectable 5mC MTase domain.

In addition, proteins harboring the DNA methylase domain PF01555 were also found in *Ancyromonas sigmoides, Nutomonas limna* and *Fabomonas mesopelagica*. This domain is found across prokaryotic N-4 cytosine-specific and N-6 Adenine-specific DNA methyltransferases (Cheng and Blumenthal 1999) and recently found also bdelloid rotifers in which this protein has been co-opted by an expression regulatory system (Arkhipova, Yushenova, and Rodriguez 2023). Finally, all ancyromonad genomes encode proteins bearing the PF05063 MT-A70 domain and putative homologs to N(6)-adenine-specific methyltransferase (METTL4), which is capable of methylating RNA and DNA (Chen et al. 2020; Hao et al. 2020). The presence of these proteins across ancyromonads suggests the existence of some of these modifications in the genomes of these organisms.

Ancyromonas sigmoides gene expression under environmental shifts

To provide insights into the molecular responses of *Ancyromonas sigmoides* to changes in its natural environment, we have designed an experiment to test the effect of five growth conditions. *Ancyromonas sigmoides* was chosen due to its contiguous genome and faster growth. The experiment spanned seven days, including culture scaling and transfer to the tested conditions (high and low temperature, low oxygen, and high and low salinity). Our qualitative observations of the cultures indicated that while some cells can survive up to 30ºC, there was a reduction in the number of cells in the cultures at this temperature. In contrast, abundant cells were observed under low oxygen conditions and low temperatures, although they appear less motile under this latter condition. No significant differences were observed in response to changes in salinity,

which aligns with the frequent shifts between marine and freshwater environments across ancyromonads observed using metabarcoding in Yubuki et al. 2023. On the seventh day, after three days on the shifted conditions, cells were harvested from each replicate, and DNA and RNA were extracted for RNA-seq and WGB sequencing.

Currently, we have only analyzed the expression data from this experiment (Figure 2). We identified 6,359 genes being differentially expressed in at least one of the tested conditions. The most divergent gene expression patterns were displayed by the low and high temperatures, followed by salinity, being low oxygen the condition with less DEGs (Figure 2ab).

In the high-temperature conditions, genes classified in the DNA integration, DNA repair, ubiquitin transferase and protein catabolic processes GO terms were enriched among the overexpressed genes (Figure 2c). Interestingly, several histones were also found to be significantly overexpressed under high temperature. Histones are structural components of the chromatin, and have been previously proposed to act as a part of the gene regulation in response to temperature change in diverse organisms (Deal and Henikoff 2010). In contrast, the categories of translation and protein folding were significantly under-expressed, as well as proteins involved in cilium assembly and transport. This might indicate that the ancyromonad cells are dealing with DNA damage and reducing their growth. This is coherent with the qualitative observations of the cultures under such conditions in which we observe fewer cells.

Across all conditions, the low temperature yielded the largest change in the number of overexpressed genes compared to the control (Figure 2b). These genes spanned several functional categories, and RNA binding and nucleus were among the most enriched categories among the genes overexpressed in this condition, these included genes involved in ribosome biogenesis messenger RNA biogenesis and nuclear exosome complex. Proteosome complex, proteolysis, vesicle mediated transport and membrane coat proteins that are involved in the endocytic-vacuolar pathway were also enriched in this condition. Protein misfolding can occur under thermal stress (Feller 2018), therefore, the enrichment of these pathways under low temperature could suggest ancyromonad cells were adjusting the protein homeostasis. The

regulation of these pathways in response to decreased temperatures has been previously reported in yeast (Isasa et al. 2016). In addition, the enrichment of DNA replication might indicate that cells under these conditions are dividing. Ancyromonads have been recovered from deep sea sediments with temperatures of 3.5 °C and high pressure (150 bar) (Živaljić et al. 2018), Therefore these results could suggest that ancyromonad are commonly able to grow at low temperatures.

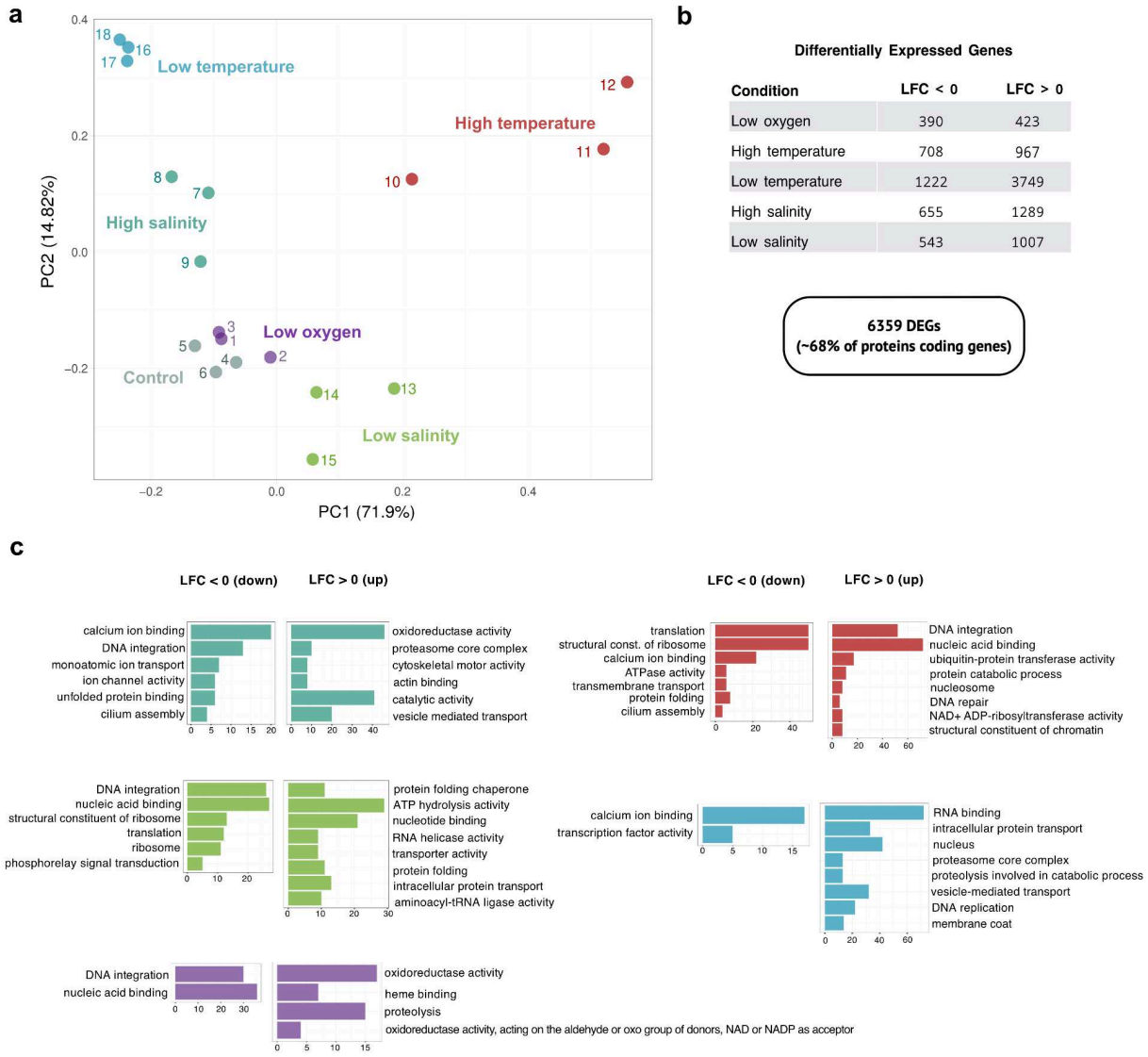

Figure. 2. Gene expression patterns of *Ancyromonas sigmoides* under shifting conditions. a) Clustering and PCA sample projection according to gene expression profile of each sample. The expression counts were normalized by DESeq2's median of ratios transformation. b) Differentially expressed genes statistics (*padj* < 0.01) LFC: log2FoldChange. c) Gene Ontology (GO) terms significantly enriched (*padj* < 0.01) in each of the gene sets. X-axis represents the number of DEGs in each enriched category; the color code is the same as for the PCA plot.

Proteins involved in ion transport were over-expressed in low salinity and under-expressed in high salinity, which is an expected response under ion imbalance. Moreover, genes related to cytoskeleton assembly were more highly expressed in the high salinity condition, while protein folding was more important with low salinity, pointing to the role of these cellular processes in the salinity shift response.

Among the DEGs with a higher expression during the exposure to low oxygen, we identified significantly enriched categories that suggest a metabolic adjustment of *Ancyromonas sigmoides* under anoxia. Among the genes overexpressed, there were mitochondrial enzymes such as Isocitrate/isopropylmalate dehydrogenase, succinate dehydrogenase/fumarate reductase flavoprotein, and malate dehydrogenase (these three involved in the tricarboxylic acid (TCA) cycle), the electron-transferring flavoprotein-ubiquinone oxidoreductase and the enzyme 3-hydroxyacyl-CoA dehydrogenase which produces NADH and participate in the beta-oxidation of fatty acids. The TCA cycle is a crucial source of reducing equivalents (NADH and FADH2) that feed into the electron transport chain (ETC) for ATP production during oxidative phosphorylation. The over-expression of TCA components might indicate that this process remains active even under these conditions, suggesting the capacity of this species of switching a different terminal electron acceptor than oxygen. Alternatively, a reverse TCA cycle has been observed during hypoxia in melanoma cells (Filipp et al. 2012), as well as in some (facultative) anaerobic protists (e.g. Gawryluk et al. 2016).

Other genes with increased expression under low oxygen encoded proteins with heme-binding domains include the cytochromes P450, globin-like proteins and heme-dependent peroxidases. Moreover, the persulfide dioxygenase ETHE1, a sulfide regulator in the mitochondria, and tryptophan 2,3-dioxygenase were also overexpressed in this condition compared to the control. These enzymes use dioxygen for their activities, their overexpression could respond to the need for a more efficient usage of available oxygen within the cell.

Interestingly, proteins with predicted integration activity were differentially expressed in several conditions. These proteins were under-expressed in high salinity, low salinity and low oxygen and overexpressed during high temperature. When

compared to the non-redundant database, these genes are homologous to gypsy retrotransposons and uncharacterized proteins in other eukaryotes.

Retrotransposons are genetic elements that can move within a genome. The expression of retrotransposons can be influenced by various factors, for example, changes in the epigenetic landscape of the genome influencing the accessibility of retrotransposon sequences (Jachowicz et al. 2017) or the activation of signaling pathways, for instance, stress-related signaling pathways may modulate the activity of transcription factors that regulate retrotransposon expression (de la Vega et al. 2007; Miousse et al. 2015). Retrotransposon activity might be regulated to prevent excessive genome instability under certain environmental challenges (Miousse et al. 2015). In contrast, under specific environmental pressures, the activation of retrotransposons may facilitate the generation of genetic variation that could be beneficial for adaptation to new environmental conditions and well as influencing the gene expression of adjacent genes (Kashkush, Feldman, and Levy 2015; Conte, Dastugue, and Vaury 2002; Li et al. 2018). The differential expression of proteins belonging to mobile elements under environmental shifts in *Ancyromonas sigmoide*s is interesting. Therefore the investigation of the genomic context of these elements or the potential coexpression with other genes will help us to gain insight into the role and impact of active transposition in this species.

Finally, it's important to note that 22% of the DEGs identified in this experiment consisted of genes without tractable homologues in non-ancyromonad species. The expression patterns associated with these under various conditions provides a weak but potentially exploitable information on the function of ancyromonad gene innovations.

Preliminary conclusions and perspectives

Previous works have shown that ancyromonads are cosmopolitan and ecologically diverse (Yubuki et al. 2023). To shed light into the genomic regulatory strategies conserved in these organisms as well as their molecular responses to changes in their environment we generated several lines of evidence.

We first aimed to explore the prevalence and diversity of DNA methylation marks into the genomes of previously sequenced ancyromonad species. The 5mC MTase domain was only found in three of the seven compared species: *A. mediterranea, N. silfraensis* and *N. limna*. Considering this, it was striking that all ancyromonads display 5mC, with low global methylation levels, more or less equally distributed in the three possible contexts CpG, CpHpG and CpHpH. However, the analysis of the average methylation levels focused on repetitive elements and protein-coding genes of *Ancyromonas sigmoides* indicated different methylation patterns. We observed that this species displays gene body methylation, being higher in CpG contexts.

Moreover, through our experiment of shifting conditions we could observe that *Ancyromonas sigmoides* can thrive in hypoxic conditions, low temperatures, and changes in salinity, but its growth is restricted by high temperatures. We identified around 6K genes (~60% from the total of the genes) in *A. sigmoides* that were differentially expressed in at least one of these conditions. The conditions with more contrasting expression profiles consisted of low and high temperature, followed by the salinity, and lastly low oxygen. The genes differentially expressed in these conditions provide hints into the metabolic and cellular responses of ancyromonads to their environments. Interestingly, among the genes that were differentially expressed in several conditions, we found genes bearing integrase domain and putatively belonging to retrotransposons, pointing to a possible role of these elements in the genome regulation of this organism.

Based on the candidate sequences retrieved by the sensitive search of epigenetic proteins, I will perform a refined characterization of the evolutionary history of these

proteins by reconstructing phylogenies for them and their homologs in prokaryotes and eukaryotes.

Moreover I will detail the characterization of the DNA methylation profiles by exploring the DNA methylated sites and levels in different regions of the seven ancyromonad genomes, and genes categorized by annotation and level of expression.

Similarly, the analysis of the *Ancyromonas sigmoides* methylation data from the experiment is still pending, but it could reveal potential differences between the methylation of genomic features under different conditions as well as their potential correlation with the expression of these features, in particular, the transposon that we observed to have a dynamic expression. With the integration of these datasets, we aim to gain insight into the genome regulation mechanisms of ancyromonads and their adaptations to the environment.

Material and methods

Cell culturing, nucleic acid isolation and DNA methylation profiles characterization

Ancyromonad cultures from seven species were grown as previously described (Yubuki et al. 2023). Total DNA was extracted from the cell pellets of the well grown cultures for each species. The purified gDNA was sent to Eurofins Genomics, Germany for Whole Genome Bisulfite Sequencing (WGBS). Bisulfite conversion was carried out with the EZ-96 DNA Methylation-Lightning MagPrep kit (zymo Research) having a conversion rate of >99,5%. Novaseq libraries were sequenced in a PE configuration (2x150) in an Illumina HiSeq platform.

Quality control and adapter trimming of the WGBS reads was performed using Trimmomatic and BBmap (Bolger, Lohse, and Usadel 2014; Bushnell 2014). Only sequences with a minimum of 28 PHRED scores in the 90% percent of bases, and a minimum average quality of 30 PHRED scores were kept for further analyses. Methylated cytosines were identified by mapping the high quality bisulfite-transformed reads to the reference genomes of the ancyromonad species using Bismark (Krueger and Andrews 2011). Furthermore, methylation calling from the resulting sorted bam mapping files was performed with BathMet2 (Zhou et al. 2019). Only positions with a coverage ≥ 10x were considered. Methylation levels of single cytosine positions were calculated as follows: 5mC/(5mC + C), where 5mC and C correspond to the read count aligned to any of both strands of the DNA sequence. Average methylation levels across transposable elements and protein coding genes were calculated using 50 pb sliding windows.

DNMTs searches

In order to fetch putative DNA methyltransferases (DNMTs) we employed HMM searches (Johnson, Eddy, and Portugaly 2010) using as query the Pfam profiles PF00145 (conserved in 5-methyl-cytosine specific (5mC) DNMTs), PF01555, and PF02384 conserved in N-4 cytosine specific (4mC) and N-6 Adenine specific (6mA) DNA methyl-transferases respectively.

Environmental variation experiment

Ancyromonas sigmoides, the type species of Ancyromonadida (Heiss, Walker, and Simpson 2010) was selected due to its faster growth rate and the contiguity of its genomic sequence to perform an experiment to test the effect of different conditions on the transcriptome and methylome profile of this species.

This species was reactivated through 1:5 serial dilutions from our collection of cryopreserved protists and growth in 12-wells culture plates. After a few generations the cultures were scaled up to obtain a high quantity of cells by inoculating nine 12-wells culture plates. The cells were collected after four days and pooled in a single container to homogenize and further inoculate 200 cell culture flasks of 50 mL. 24 flasks were randomly picked to be transferred to the each tested condition after four days while the other 24 remained in the normal (control) conditions. The normal growing conditions of this species are 1% YT medium consisting of 50% natural sea water and 50% mineralised water, with an approximate of 3% of marine salts at 18ºC. The treatment conditions consisted of low temperature (4ºC), high temperature (30º), low salinity (1.5% salinity), high salinity (7% salinity) and low oxygen (culturing flasks full of media with closed caps).

The cultures were maintained in the treatment conditions and during the 4th day the ancyromonad cells were harvested using sterile cell scrapers, pooled by centrifugation in 15mL tubes and snap frozen in liquid nitrogen. Three cell pellets were obtained from each treatment and control from which we extracted DNA and RNA with the *All prep DNA-RNA extraction* kit (QIAGEN). The purified nucleic acid material was sent to Eurofins Genomics (Germany) to be sequenced with WGBS and RNA-seq after a poly-A selection respectively.

Characterization of gene expression patterns of Ancyromonas sigmoides

Paired unassembled reads from the 18 RNAseq libraries were quality controlled with Trimmomatic as previously described. Each library was then mapped to the reference genome of *Ancyromonas sigmoides* using two-pass alignment with STAR v2 (Dobin et al. 2013). We used the DESeq2 R package (Love, Huber, and Anders 2014; R Core Team 2021) to quantify the gene expression profile for each sample and test for significant expression differences between control and the treatments. P adjusted values (*padj*) attained by the Wald test were corrected for multiple testing using the Benjamini and Hochberg method. Differential expression contrasts were performed using the Wald test and Benjamini-Hochberg p-value correction for multiple testing. Only genes with p-adjusted < 0.05 were considered as Differentially Expressed Genes (DEGs). A term enrichment analysis was performed using ClusterProfiler (Yu et al. 2012) package comparing the Gene Ontology (GO) terms of the DEGs and the terms of all the genes encoded in the genome previously annotated with InterproScan v5 (P. Jones et al. 2014).

References

Arkhipova, Irina R., Irina A. Yushenova, and Fernando Rodriguez. 2023. “Shaping Eukaryotic Epigenetic Systems by Horizontal Gene Transfer.” *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 45 (7): e2200232.

Bewick, Adam J., and Robert J. Schmitz. 2017. “Gene Body DNA Methylation in Plants.” *Current Opinion in Plant Biology* 36 (April): 103–10.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. “Trimmomatic: A Flexible Trimmer for Illumina Sequence Data.” *Bioinformatics* 30 (15): 2114–20.

Bushnell, Brian. 2014. “BBMap: A Fast, Accurate, Splice-Aware Aligner.” LBNL-7065E. Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States). https://www.osti.gov/biblio/1241166.

Cheng, Xiaodong, and Robert Blumenthal. 1999. *S-Adenosylmethionine-Dependent Methyltransferases: Structures and Functions*. World Scientific.

Chen, Hao, Lei Gu, Esteban A. Orellana, Yuanyuan Wang, Jiaojiao Guo, Qi Liu, Longfei Wang, et al. 2020. "METTL4 Is an snRNA m6Am Methyltransferase That Regulates RNA Splicing." *Cell Research* 30 (6): 544–47.

Conte, Caroline, Bernard Dastugue, and Chantal Vaury. 2002. "Promoter Competition as a Mechanism of Transcriptional Interference Mediated by Retrotransposons." *The EMBO Journal* 21 (14): 3908–16.

Deal, Roger B., and Steven Henikoff. 2010. "Gene Regulation: A Chromatin Thermostat." *Nature* 463 (7283): 887–88.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. "STAR: Ultrafast Universal RNA-Seq Aligner." *Bioinformatics* 29 (1): 15–21.

Fan, Xiao, Wentao Han, Linhong Teng, Peng Jiang, Xiaowen Zhang, Dong Xu, Chang Li, et al. 2020. "Single‐base Methylome Profiling of the Giant Kelp Saccharina Japonica Reveals Significant Differences in DNA Methylation to Microalgae and Plants." *New Phytologist*. https://doi.org/10.1111/nph.16125.

Feller, Georges. 2018. "Protein Folding at Extreme Temperatures: Current Issues." *Seminars in Cell & Developmental Biology* 84 (December): 129–37.

Filipp, Fabian V., David A. Scott, Ze'ev A. Ronai, Andrei L. Osterman, and Jeffrey W. Smith. 2012. "Reverse TCA Cycle Flux through Isocitrate Dehydrogenases 1 and 2 Is Required for Lipogenesis in Hypoxic Melanoma Cells." *Pigment Cell & Melanoma Research* 25 (3): 375–83.

Gawryluk, Ryan M. R., Ryoma Kamikawa, Courtney W. Stairs, Jeffrey D. Silberman, Matthew W. Brown, and Andrew J. Roger. 2016. "The Earliest Stages of Mitochondrial Adaptation to Low Oxygen Revealed in a Novel Rhizarian." *Current Biology: CB* 26 (20): 2729–38.

Glücksman, Edvard, Elizabeth A. Snell, and Thomas Cavalier-Smith. 2013. "Phylogeny and Evolution of Planomonadida (Sulcozoa): Eight New Species and New Genera Fabomonas and Nutomonas." *European Journal of Protistology* 49 (2): 179–200.

Grau-Bové, Xavier, Cristina Navarrete, Cristina Chiva, Thomas Pribasnig, Meritxell Antó, Guifré Torruella, Luis Javier Galindo, et al. 2022. "A Phylogenetic and Proteomic Reconstruction of Eukaryotic Chromatin Evolution." *Nature Ecology & Evolution* 6 (7):

1007–23.

Hao, Ziyang, Tong Wu, Xiaolong Cui, Pingping Zhu, Caiping Tan, Xiaoyang Dou, Kai-Wen Hsu, et al. 2020. “N6-Deoxyadenosine Methylation in Mammalian Mitochondrial DNA.” *Molecular Cell* 78 (3): 382–95.e8.

Heiss, Aaron A., Giselle Walker, and Alastair G. B. Simpson. 2010. “Clarifying the Taxonomic Identity of a Phylogenetically Important Group of Eukaryotes: Planomonas Is a Junior Synonym of Ancyromonas.” *The Journal of Eukaryotic Microbiology* 57 (3): 285–93.

Heiss, Aaron A., Giselle Walker, and Alastair G. B. Simpson. 2011. “The Ultrastructure of Ancyromonas, a Eukaryote without Supergroup Affinities.” *Protist* 162 (3): 373–93.

Hoguin, Antoine, Feng Yang, Agnès Groisillier, Chris Bowler, Auguste Genovesio, Ouardia Ait-Mohamed, Fabio Rocha Jimenez Vieira, and Leila Tirichine. 2023. “The Model Diatom Phaeodactylum Tricornutum Provides Insights into the Diversity and Function of Microeukaryotic DNA Methyltransferases.” *Communications Biology* 6 (1): 253.

Hosseini, Sara, Cécile Meunier, Diem Nguyen, Johan Reimegård, and Hanna Johannesson. 2020. “Comparative Analysis of Genome-Wide DNA Methylation in.” *Epigenetics: Official Journal of the DNA Methylation Society* 15 (9): 972–87.

Isasa, Marta, Clara Suñer, Miguel Díaz, Pilar Puig-Sàrries, Alice Zuin, Anne Bichman, Steven P. Gygi, Elena Rebollo, and Bernat Crosas. 2016. “Cold Temperature Induces the Reprogramming of Proteolytic Pathways in Yeast.” *The Journal of Biological Chemistry* 291 (4): 1664–75.

Jachowicz, Joanna W., Xinyang Bing, Julien Pontabry, Ana Bošković, Oliver J. Rando, and Maria-Elena Torres-Padilla. 2017. “LINE-1 Activation after Fertilization Regulates Global Chromatin Accessibility in the Early Mouse Embryo.” *Nature Genetics* 49 (10): 1502–10.

Johnson, L. Steven, Sean R. Eddy, and Elon Portugaly. 2010. “Hidden Markov Model Speed Heuristic and Iterative HMM Search Procedure.” *BMC Bioinformatics* 11 (August): 431.

Jones, Peter A. 2012. “Functions of DNA Methylation: Islands, Start Sites, Gene Bodies and beyond.” *Nature Reviews. Genetics* 13 (7): 484–92.

Jones, Philip, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, et al. 2014. “InterProScan 5: Genome-Scale Protein Function Classification.” *Bioinformatics* 30 (9): 1236–40.

Kashkush, Khalil, Moshe Feldman, and Avraham A. Levy. 2015. “Corrigendum: Transcriptional Activation of Retrotransposons Alters the Expression of Adjacent Genes in Wheat.” *Nature Genetics* 47 (9): 1099.

Katz, Laura A. 2006. “Genomes: Epigenomics and the Future of Genome Sciences.” *Current Biology: CB* 16 (23): R996–97.

Krueger, Felix, and Simon R. Andrews. 2011. “Bismark: A Flexible Aligner and Methylation Caller for Bisulfite-Seq Applications.” *Bioinformatics* 27 (11): 1571–72.

Li, Zi-Wen, Xing-Hui Hou, Jia-Fu Chen, Yong-Chao Xu, Qiong Wu, Josefa González, and Ya-Long Guo. 2018. “Transposable Elements Contribute to the Adaptation of Arabidopsis Thaliana.” *Genome Biology and Evolution* 10 (8): 2140–50.

Lloyd, James P. B., and Ryan Lister. 2022. “Epigenome Plasticity in Plants.” *Nature Reviews. Genetics* 23 (1): 55–68.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *Genome Biology* 15 (12): 550.

Lowdon, Rebecca F., Hyo Sik Jang, and Ting Wang. 2016. “Evolution of Epigenetic Regulation in Vertebrate Genomes.” *Trends in Genetics: TIG* 32 (5): 269–83.

Lyko, Frank. 2018. “The DNA Methyltransferase Family: A Versatile Toolkit for Epigenetic Regulation.” *Nature Reviews. Genetics* 19 (2): 81–92.

Madhani, Hiten D. 2021. “Unbelievable but True: Epigenetics and Chromatin in Fungi.” *Trends in Genetics: TIG* 37 (1): 12–20.

Mendoza, Alex de, Amandine Bonnet, Dulce B. Vargas-Landin, Nanjing Ji, Hongfei Li, Feng Yang, Ling Li, et al. 2018. “Recurrent Acquisition of Cytosine Methyltransferases into Eukaryotic Retrotransposons.” *Nature Communications* 9 (1): 1341.

Mendoza, Alex de, Ryan Lister, and Ozren Bogdanovic. 2020. “Evolution of DNA Methylome Diversity in Eukaryotes.” *Journal of Molecular Biology* 432 (6): 1687–1705.

Miousse, Isabelle R., Marie-Cecile G. Chalbot, Annie Lumen, Alesia Ferguson, Ilias G.

Kavouras, and Igor Koturbash. 2015. “Response of Transposable Elements to Environmental Stressors.” *Mutation Research-Reviews in Mutation Research* 765 (May): 19–39.

Neri, Francesco, Stefania Rapelli, Anna Krepelova, Danny Incarnato, Caterina Parlato, Giulia Basile, Mara Maldotti, Francesca Anselmi, and Salvatore Oliviero. 2017. “Intragenic DNA Methylation Prevents Spurious Transcription Initiation.” *Nature* 543 (7643): 72–77.

Osuna-Cruz, Cristina Maria, Gust Bilcke, Emmelien Vancaester, Sam De Decker, Atle M. Bones, Per Winge, Nicole Poulsen, et al. 2020. “The Seminavis Robusta Genome Provides Insights into the Evolutionary Adaptations of Benthic Diatoms.” *Nature Communications* 11 (1): 3320.

R Core Team. 2021. “R: A Language and Environment for Statistical Computing.” Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Schmitz, Robert J., Zachary A. Lewis, and Mary G. Goll. 2019. “DNA Methylation: Shared and Divergent Features across Eukaryotes.” *Trends in Genetics: TIG* 35 (11): 818–27.

Svedružić, Željko M. 2011. “Dnmt1 Structure and Function.” *Progress in Molecular Biology and Translational Science* 101: 221–54.

Tikhonenkov, Denis Victorovich, Yuri Alexandrovich Mazei, and Alexander Petrovich Mylnikov. 2006. “Species Diversity of Heterotrophic Flagellates in White Sea Littoral Sites.” *European Journal of Protistology* 42 (3): 191–200.

Vega, Enrique de la, Bernard M. Degnan, Michael R. Hall, and Kate J. Wilson. 2007. “Differential Expression of Immune-Related Genes and Transposable Elements in Black Tiger Shrimp (Penaeus Monodon) Exposed to a Range of Environmental Stressors.” *Fish & Shellfish Immunology* 23 (5): 1072–88.

Veluchamy, Alaguraj, Xin Lin, Florian Maumus, Maximo Rivarola, Jaysheel Bhavsar, Todd Creasy, Kimberly O’Brien, et al. 2013. “Insights into the Role of DNA Methylation in Diatoms by Genome-Wide Profiling in Phaeodactylum Tricornutum.” *Nature Communications* 4: 2091.

Weiner, Agnes K. M., Mario A. Cerón-Romero, Ying Yan, and Laura A. Katz. 2020. “Phylogenomics of the Epigenetic Toolkit Reveals Punctate Retention of Genes across Eukaryotes.” *Genome Biology and Evolution* 12 (12): 2196–2210.

Weiner, Agnes K. M., and Laura A. Katz. 2021. “Epigenetics as Driver of Adaptation and Diversification in Microbial Eukaryotes.” *Frontiers in Genetics* 12 (March): 642220.

Yubuki, Naoji, Guifré Torruella, Luis Javier Galindo, Aaron A. Heiss, Maria Cristina Ciobanu, Takashi Shiratori, Ken-Ichiro Ishida, et al. 2023. “Molecular and Morphological Characterization of Four New Ancyromonad Genera and Proposal for an Updated Taxonomy of the Ancyromonadida.” *The Journal of Eukaryotic Microbiology*, August, e12997.

Yu, Guangchuang, Li-Gen Wang, Yanyan Han, and Qing-Yu He. 2012. “clusterProfiler: An R Package for Comparing Biological Themes among Gene Clusters.” *Omics: A Journal of Integrative Biology* 16 (5): 284–87.

Zhou, Qiangwei, Jing-Quan Lim, Wing-Kin Sung, and Guoliang Li. 2019. “An Integrated Package for Bisulfite DNA Methylation Data Analysis with Indel-Sensitive Mapping.” *BMC Bioinformatics* 20 (1): 47.

Živaljić, Suzana, Alexandra Schoenle, Frank Nitsche, Manon Hohlfeld, Julia Piechocki, Farina Reif, Marwa Shumo, et al. 2018. “Survival of Marine Heterotrophic Flagellates Isolated from the Surface and the Deep Sea at High Hydrostatic Pressure: Literature Review and Own Experiments.” *Deep-Sea Research. Part II, Topical Studies in Oceanography* 148 (February): 251–59.

8. Discussion and perspectives

This section presents an overview of the main contributions and limitations of this thesis and the methodological approaches employed. The main findings of the results chapters and their possible implications are discussed considering these limitations and an outline of the prospects for future research in the evolution of the eukaryotic epigenome and the biology of orphan protists is proposed.

Expanding the genomic landscape of the eukaryotic tree of life

To date, in-depth functional and comparative genomics of eukaryotes has predominantly focused on a limited selection of major lineages. While this has provided critical insights into the diversity and evolutionary history of eukaryotes, our understanding remains a narrow glimpse into the broader landscape of biological diversity that exist in this domain of life (Sibbald and Archibald 2017; Blaxter et al. 2022; Richter et al. 2022). In particular, some questions of evolutionary and ecological importance can only be effectively addressed with comprehensive genome datasets that encompass divergences across the eukaryotic tree of life (eToL). The orphan branches of the eToL play a crucial role for our understanding of how eukaryotic innovations arose and evolved through their diversification, as well as our inferences of the characteristics of the last eukaryotic common ancestor.

Interestingly, many of the previously unsurveyed diversity classified in novel major lineages were first discovered and characterized trough culture methods and not by high-throughput environmental surveys (Tikhonenkov et al. 2022; Galindo, López-García, and Moreira 2022; Eglit et al. 2023; Lax et al. 2018; Brown et al. 2018), underpinning the importance of isolation and culturing that in addition allow to observe the morphology and behavior of organisms in detail.

The first objective of this thesis was thus, to generate high-quality genomic data for species that exhibit profound divergences within the eToL, while improving the sampling of species from these groups harnessing previous efforts of isolation of diverse ancyromonad and mantamonad species. The main challenges we faced in achieving this goal revolved around the impossibility of growing these organisms at a high cell density in cultures lacking the prokaryotes on which they feed. As well, we encountered some limitations during the bioinformatic scrutiny of their sequences because they lack close relative representatives in public databases. We addressed these limitations through the use of custom sequencing workflows and sensitive computational analyses respectively that allowed us to assemble, assess, decontaminate and curate the genomic data for these organisms.

For example, standard methods used to evaluate the completeness of a genome rely on the identification of a set of genes that have a wide phylogenetic distribution (Saary, Mitchell, and Finn 2020; Hanschen, Hovde, and Starkenburg 2020; Manni et al. 2021). However, these dataset are yet taxonomically biased to the species with available proteome data. We compared the prevalence of these markers in several species, which allowed us to estimate if the presence or absence of markers was exhibited by a single species or by the whole clade. We also used as a proxy the percentage of alignment of clean RNA seq data to complement this completeness estimation. In addition the comprehensive search of homologous proteins in a phylogenetically diverse custom database allowed us to effectively decontaminate these genomic sequences.

We used a combined cell sorting and genome amplification workflow to sequence six additional ancyromonad genomes. This represented an advantage when dealing with species particularly slow growing and cultures with very low ancyromonad density, however we also encountered additional challenges. Genome amplification can introduce biases into the data due to the uneven amplification of different genomic regions and an incomplete coverage of the genome (Sabina and Leamon 2015). We accounted for these biases in the short read data by subsampling the genomic reads generated through this strategy before combining it with the bulk-culture sequencing data. Moreover the amplification reduced the power of Nanopore in the generation of long reads and impeded the capacity of detecting DNA modifications from this data. Therefore, the long read data was assembled using a metagenomic-like approach to

deal with the coverage biases, this resulted in less contiguous genome assemblies than the ones generated from bulk cultures, although some of these genomes reached a high proportion of estimated proteome completeness, which in turn represented an advantage for the further comparative analyses.

The results presented in the manuscript one and two represent the first efforts that expanded our ability to study mantamonads and ancyromonads at the genomic level, making possible to address the biology, ecology and evolution of these organisms approached in the third and fourth manuscripts.

Mantamonas illuminate the ancient evolution of key eukaryotic components

Before this work, the CRuMs supergroup was represented by less than 10 described species (Adl et al. 2019), and only four partial transcriptomic datasets (Brown et al. 2018). Our results added two new species to the available proteomes for this clade and supported the position of mantamonads within this supergroup that is the sister branch of Amorphea, including opisthokonts, amoebozoans, breviates and apusomonads.

We also observed the conservation of a set of ~1800 genes conserved across the CRuMs proteomes, including an Integrin-linked kinase (TKL/DRK) involved adhesive roles in Amorphea (Kang et al. 2021) and 26 carbohydrate active enzymes (CAZy). This provides a starting point to understand the characteristics of their last common ancestor. Moreover another set of more than three thousand genes were observed to be uniquely conserved among *Mantamonas* species. These genes, representing the ~30% of *M. sphyraenae* genome, can be specific lateral transfers from outside eukaryotes or genes originated in this genus, and were not surprisingly enriched in unknown functions.

The *Mantamonas sphyraenae* nuclear genome of about 25 Mb contains TTAGGG telomeric repeats at both ends of several of the assembled contigs and based on the distribution of single nucleotide polymorphisms (SNPs) we hypothesized that this is a diploid genome. While additional analyses are required to clarify the karyotype of this

organism these results suggest the existence of 66 pairs of chromosomes. Interestingly, among the repetitive elements that make up around 12% of this genome were classified among well known families of transposable elements (TEs) such as the Bel-Pao family, which to our knowledge, was previously reported only in metazoa (de la Chaux and Wagner 2011).

This genome also revealed the retention of a very complete complement of the proteins involved in intracellular transport, and particular paralogues that are rare in other species but that were probably present in the last eukaryotic common ancestor. Altogether these results improve the resolution of the origin and evolutionary history of ancient eukaryotic machineries and represent a particular good opportunity to understand the innovations of Amorphea, the sister group of the CRuMs encompassing Opistokhonts, Amoebozoa and other protists.

Finally, during the sequencing of *M. spyraenae* we also recovered the gene-rich mitochondrial genome of this species (data not shown, Moreira, Blaz, Kim and Eme. *in preparation*). Mitochondrial gene rich genomes have been reported so far in jakobids (Burger et al. 2013), *Ancoracysta twista* (Janouškovec et al. 2017) and very recently in *Meteora sporadica* (Eglit et al. 2023), all very distantly related to the CRuMs. Depending on the position of the root of the eToL, this could imply that there were independent massive reductions of the mitochondrial genome during the early diversification of eukaryotes. Therefore, this mitochondrial genome opens a valuable opportunity to refine the understanding of the ancient evolutionary history of this important organelle.

Ancyromonas sigmoides **genome, hints of its dynamic nature and protein origins**

Ancyromonas sigmoides, was described in the nineteenth century (Kent 1882) and is considered the type species of the Ancyromonadida clade, a key lineage with a deep divergence within the eukaryotic tree of life (Paps et al. 2013; Brown et al. 2018; Atkins, McArthur, and Teske 2000; Burki et al. 2020). In addition, as phagotrophic bacterivorous

protists, ancyromonads are probably playing key roles in the trophic networks of their environments, however the biology of these organisms has just started being explored.

The genome of *A. sigmoides* of almost 40 Mb encoded more than 11K genes including anciently originated gene families such as the meiotic proteins Rec8 and SPO11, and the polycomb machinery. This species, however, has also evolved an important number of new proteins since its divergence from other eukaryotes as is suggested by the high proportion of lineage-specific proteins and genes without homologs in annotation databases. The domain analysis of these proteins indicated that the most abundant Pfam domains found in these proteins, included RNI-like and RAN GTPase activators protein families, and proteins bearing Ankyrin domains involved in diverse functions such as signal transduction, cell adhesion, and cell-cycle regulation. Altogether this suggests that ancyromonad retain very ancient features probably present in the last eukaryotic common ancestor but also has evolved new protein families with a diverse set of functions, potentially participating in cellular signaling and regulation of cellular processes that are particular in this species.

Furthermore, ~1,200 of the proteins encoded in the genome of *A. sigmoides* have prokaryotic proteins as the closest homolog. These proteins spanned 50 phyla of bacteria and archaea from the Euryarchaeota, Woesearchaeota, Heimdallarchaeota, and Lokiarchaeota phyla. Archaeal proteins are not very often reported in eukaryotic genomes (Sieber, Bromley, and Dunning Hotopp 2017). Although our capacity to track the specific origin of these proteins is limited due to the biases of the database towards particular lineages that have been more sequenced than others, the prevalence of archaeal genes in *A. sigmoides* genome suggest an unique evolutionary history of this species compared to other studied eukaryotes.

Finally, the genome architecture of this species was characterized by an important proportion of simple and interspersed repeats (~29%). Although most of these elements lacked homologous in repeat databases, we were able to trace the abundant presence of proteins with viral motifs such as the EsV-1-7 cysteine-rich motive from double strand DNA viruses and the Zinc-knuckle domain from retroviral gag proteins suggesting that diverse RNA and double strand DNA viruses populate the genome of this organism. A recent study has comprehensively revealed that many protists' genomes contain abundant and diverse viruses (Bellas et al. 2023). In addition

viral elements are proposed to participate in the lateral transfer of genetic material among eukaryotes (Sibbald et al. 2020; Irwin et al. 2022).

The further expression analysis of this species confirmed that some of these mobile elements are actually active in the genome of *A. sigmoides* and this activity changes in response to environmental stimulus. Active mobile elements are major contributors to genome diversification (Todorovska 2007; Feschotte and Pritham 2007). Some authors have previously argued that this diversification could be important for species survival under changing conditions (for example under nitrate starvation in diatoms (Maumus et al. 2009). Indeed the idea that the activity of mobile elements in response to challenges could be beneficial to the organism in some conditions was proposed by Barbara McClintock in 1984 (McClintock 1984). However, whether the complex relationship between mobile element activity and the host phenotype in response to stress is beneficial or harmful is still an open question in the field (Horváth, Merenciano, and González 2017; Capy et al. 2000).

Further analyses are required to understand how the activity of mobile elements is regulated in *A. sigmoides*, however these findings make this species an interesting model to understand the role of these elements in the genome dynamics of protists. Some of the questions that emerge from these observations are for example the tempo of the acquisition of the mobile reservoir of *A. sigmoides*, the possible intraspecific variation of viral elements between isolates of different locations and their correlation to the potential variation in the genome architecture of this species.

Reconstructing the deep evolutionary history of ancyromonad genomic repertoires using phylogenetic reconciliation

Several lines of evidence suggest that the diversification of the last eukaryotic common ancestor into the major extant lineages occurred rapidly (Knoll 2014; Cohen and Kodner 2022). How this macroevolutionary phenomenon has driven the evolution of the genome of different eukaryotic lineages is an important and open question that can only be addressed by studying the wide diversity of this domain of life.

By including several new genomes belonging to orphan protists in a large-scale comparative framework we addressed this question and provided insights into the general patterns of gene content evolution in ancyromoands since their divergence from the rest of eukaryotic supergroups using phylogenetic reconciliation. Phylogenetic reconciliation methods are powerful tools for inferring the evolutionary history of gene families across lineages (Rees et al. 2001). These approaches explicitly model evolutionary events of gene duplications, losses, transfers, and origination of a given gene family under an hypothesis of the species phylogeny. When comparing the complete genomic repertoire of different organisms this also offers a comprehensive view of the relative importance of processes in the overall gene gain and loss of the genome of different lineages. We employed the Amalgamated Likelihood Estimation (ALE) reconciliation that implements a probabilistic approach to account for the uncertainty of the gene-family phylogeny, and provides a quantitative frequency of events per gene family accounting for this uncertainty (Szöllõsi et al. 2013; Williams et al. 2023). However, it is important to consider that all reconciliation approaches, including ALE, are sensitive to the quality and accuracy of the input phylogenies. Errors or biases in gene-family tree estimation such as long branch attraction, substitution saturation or lack of signal can impact the accuracy of inferred evolutionary events especially at very ancient events. Therefore, it is important to understand these limitations when interpreting the reconciliation inferences.

ALE analysis requires a bifurcating and rooted species tree. We reconstruct a species phylogeny using the new set of 799 single copy orthologs obtained from our proteome dataset using a Maximum Likelihood method and a phylogenetic constraint. The phylogenetic constraint consisted in a multifurcating species tree based on the current consensus of the eToL (from Richter et al. 2022). In this phylogeny, several well known supergroups are monophyletic but the deep relationships among them are unresolved. Although this method limits the phylogenetic resolution within the constrained supergroups, it was a computationally efficient approach to get an overview of the relationships among eukaryotes and provide a plausible scenario for our comparisons. Our results supported an early divergence of Ancyromonads within the eukaryotic global phylogeny, as sister branch of *Gefionella okellyi* plus Metamonada, and Hemimastigophora was inferred as sister branch of that clade. The support of these

groups was 68 and 69% respectively. Moreover, these species were found to share very low specific synapomorphies, indicating that if true, these divergences occurred rapidly. Alternative positions for these lineages also remain conceivable. Indeed, we tested the impact of an alternative phylogeny into the ALE analysis with Hemimastigophora at a different position (Supplementary Material 1 Figure S15-S17). This analysis showed that the overall evolutionary trends inferred at the origin and across ancyromonads remained robust. In addition, the employed dataset, markers and method can yield further interesting comparisons with phylogenies produced by more standard maker datasets and phylogenetic workflows.

Moreover, the reconciliation analysis involved the reconstruction of large sets of gene families, their phylogenies and their ultra fast bootstrap samples, that are employed by ALE as a measure of the uncertainty of each of these phylogenies. In order to include an eukaryotic taxonomic diversity as comprehensible and balanced as possible, we used proteomes derived from all the major eukaryotic supergroups with a sequenced representative available in these years from The Comparative Set of the EukProt v3 database (Richter et al. 2022). Many relevant lineages are still only explored at the transcriptome level. Even if a transcriptome may provide an incomplete and biased representation of the genetic complement of an organism however they provide valuable insights for poorly explored eukaryotes such as malawimonads. Therefore, this data was included in our reconstruction for this reason and we used a missing portion parameter to correct the estimation of losses given the estimated completeness of the data.

Taking the above limitations into account we focused on the events that were inferred in more than the 50% of the ALE reconciliations. We also used the ratio of originations per gene family as a proxy of the uncertainty of their phylogeny. Then, gene families with a wide phylogenetic distribution and low origination fraction were considered to have an ancient uncertain origin within eukaryotes.

Furthermore, our results indicate that across the major eukaryotic lineages compared, duplication is the main evolutionary process underlying gene gain in comparison with originations and lateral gene transfers between eukaryotes (eukLGT). However, an exception to this was observed at the base of most supergroups and early splits of our species phylogeny in which eukLGT was higher than duplication. How

frequent and important it's been eukLGT into the evolution of eukaryotes it's indeed a hotly debated subject (Leger et al. 2018). Although outstanding evidence of eukLGT exists for some well studied lineages such as Ochrophyta (Dorrell et al. 2021) and Rhizaria (J. J. E. van Hooff and Eme 2023), this question has not been addressed at the scale of all eukaryotes. This is not trivial because the importance of eukLGT during the early evolution of eukaryotes has profound implications in our way to infer the gene content of the last eukaryotic common ancestor. Specifically, Can we always interpret the patchy occurrence of genes among distant eukaryotes as an ancestral LECA component? How to identify true instances of eukLGT from the latter scenario? and consequently Has the gene content of the last eukaryotic common ancestor expanded or contracted? The answer is of course not easy and scenarios are not mutually exclusive, since an anciently originated family could have been laterally transferred during its evolutionary history. Including diverse lineages of eukaryotes to improve the taxonomic balance facilitating the discrimination of true eukLGTs and implementing refined sequence evolution models could increase the resolution of these analyses. Furthermore, the study of modern eukLGT instances could also provide mechanistic understanding on how transfer occurs, evolves and if there are functional and ecological patterns tied to the genes transferred and lineages involved respectively.

Our results also suggest that a significant aspect of the evolution of ancyromonad genomic repertories that distinguished them from other eukaryotic lineages, is the high frequency of gene family originations. This was observed at the branch leading to the last common ancestor of ancyromonads and at crucial points in their evolution. Gene family origination can represent significant evolutionary events such as domain shuffling/fusion and lateral gene transfers beyond eukaryotes. These alternatives were further investigated by characterizing the domain architecture of ancyromonad genes as well as the taxonomic affiliation of these gene families outside eukaryotes as discussed in the following sections. However, almost 4,800 of these gene families were found to be taxonomically restricted to ancyromonads even when compared with the comprehensive databases *nr*, eggNOG and GTDB and could therefore represent *de novo* gene birth events across the history of these lineages. The presence of these genes raises questions about their number and prevalence compared to other orphan lineages of protists. Some aspects that can be investigated to gain

insights into their biologically significance are for example their expression patterns in different ancyromonad species and under different conditions. Analysis of selection can also be done to know if these genes are under any kind of selection or are neutrally evolving.

Evolutionary patterns across ancyromonad diversification and their possible functional implications

One of the goals of comparative genomics is to understand the evolutionary causes of extant diversity and the functional consequences of this diversity in the phenotypic and ecological characteristics of modern organisms. Collectively, our results highlight the high diversity of genome architecture and gene content between ancyromonads despite their morphological similitudes. The processes of originations and eukLGT as described above, together with genome architecture rearrangements driven by the mobilome across the further diversification of ancyromonads partially underlie these variations. In addition losses and duplications have contributed to the specific expansion and contraction of gene families through the evolution of ancyromonad lineages.

Loss is an important force in the evolution of eukaryotes, protist lineages often challenge the notion of what is an essential component of an eukaryotic cell. Metamonads are the striking example of how an eukaryote can lose something as important as mitochondria (Karnkowska et al. 2016) or canonical DNA repair processing mechanisms (Salas-Leiva et al. 2021) and yet keep thriving on this planet. However we must also be cautious when interpreting gene family loss in comparative genomic studies because differentiating true biological losses from missing data or lack of detection can be challenging. Taking this into account we opted to take a conservative approach and consider as losses when these events were shared by two or more species from the same monophyletic clade of ancyromonads. Our analyses suggest that the lineage leading to the last ancyromonad common ancestor has suffered a retraction of genes with transcription and translation initiation factors. Different lineages of eukaryotes have evolved diverse transcription factors and translation complexes in

response to specific cellular or environmental requirements contributing to the diversity in their protein translation rates and strategies (Hernández et al. 2012; Genuth and Barna 2018). Whether these genes have been lost or evolved beyond recognition in ancyromonads is still an open question that needs to be addressed using more sensitive tools of homology search.

Moreover, an important turnover (originations and losses) of gene families involved in cytoskeleton and signal transduction was also observed. Cytoskeleton changes might have impacted the particular cell shape and motion exhibited by ancyromonads (Heiss, Walker, and Simpson 2011). The changes in signal transduction-related families suggest that ancyromonads species have evolved specialized strategies to respond to cellular and environmental signals than other eukaryotic lineages.

Furthermore, diverse evolutionary patterns were observed throughout the evolution of ancyromonad species since their last common ancestor. For instance, *Fabomonas* and *Planomonas micra* displayed lower gene family duplications and more eukLGTs, potentially retaining more ancestral features. In contrast, ancyromonads within the genera *Ancyromonas, Nyramonas, Striomonas, and Nutomonas* exhibited more gene duplications across diverse functional categories. In addition, a high number of species-specific gene families was also observed across all ancyromonads. These evolutionary patterns could be associated with specific adaptations to environmental constraints and interactions. Ancyromonads have been isolated from marine and continental benthic sediments, an environment characterized by the stratification of oxygen and other conditions (Gücker and Fischer 2003). If the hypothesis that different species of ancyromonads have different adaptation strategies holds true we could expect variations in the responses to environmental signals in the transcriptomic patterns in different species. In addition, to answer this question it would be also necessary to better characterize the ecological aspects of ancyromonad distribution, interactions and limiting conditions in their natural environments.

As discussed below, some specific changes involved the possible lateral transfer of genes from bacteria and archaea. A natural question that emerges from these observations is if these gene-families can hint the long-term interactions of ancyromonads with prokaryotic species.

Prokaryotic ancestry of ancyromonad genes

In eukaryotes, the most widespread instances of LGT from bacteria were acquired from the mitochondria and chloroplast organelles. As these organelles, some prokaryotes have been shown to be intracellular symbionts of diverse eukaryotes hosts (Sieber, Bromley, and Dunning Hotopp 2017). Maybe one of the most striking examples of the lateral transfer of genetic material from symbiotic systems are illustrated by the protosexual chromosome acquired in the pillbug *Armadillidium vulgar* from its *Wolbachia* endosymbionts (Leclercq et al. 2016). Indeed, the evidence of prokaryotic gene transfer into eukaryotes is abundant in animals, plants, fungi and protists and as the genomic sampling of prokaryotes and eukaryotes improves, new intriguing LGT cases emerge and highlight the evolutionary significance of this evolutionary process into the evolution of eukaryotic genomes (Danchin 2016; Leger et al. 2018). However, robust methods are needed for the identification and verification of LGT. The discrimination between LGT and contamination is especially problematic in highly fragmented datasets in which it is not possible to study the context of the putatively transferred genes. Furthermore, insights from phylogenetic analysis on both EGT and LGT are dependent on taxon sampling, which can be uneven in our dataset in terms of the availability and quality of data from diverse lineages.

To explore the prokaryotic ancestry of the genomic novelties of ancyromonads we analyzed the taxonomic and functional affiliation of the genes inferred to be restricted to ancyromonads when compared to other eukaryotes. We identified several gene families homologous to bacterial and archaeal genes acquired into different points of ancyromonad diversification. The majority of these putative lateral gene transfers (LGTs) are protein families with diverse predicted activities and functions, underscoring the complex and ancient nature of genetic exchanges shaping the genomic repertoire of ancyromonads. A particularly interesting case of these genes were the genes involved in the denitrification pathway found in *Nutomonas limna* potentially associated to the capacity of these organisms to grow in anoxic conditions as it has been observed in other benthic organisms and as we experimentally observed in the type species *Ancyromonas sigmoides*. The further phylogenetic analysis of these putative lateral

transfers could help us to gain insight into the long-term interactions of ancyromonads in their ecosystems.

Experimental approaches to investigate the epigenetic marks and gene expression of ancyromonads

As previously discussed, a crucial aspect to unravel the evolutionary significance of ancyromonad genomic diversity is to better understand the interplay between ancyromonads and their environment. Epigenetics plays a crucial role in the adaptation and diversifications of microeukaryotes (Weiner and Katz 2021). Therefore we aimed to investigate the potential relationship between epigenetic marks, transcription profiles, and controlled environmental shifts in these organisms. At the time these results are preliminary and many downstream analyses are still missing, however we can already discuss some interesting outcomes of these experiments.

As a first glance of the diversity of ancyromonad epigenome, we generated Whole Genome Bisulfite Sequencing (WGBS) data to search 5-methylcytosine (5mC) DNA methylation marks across the genomes of our sequenced species. WGBS relies on the bisulfite conversion to distinguish methylated from unmethylated cytosines and provides single-base resolution, offering detailed insights into DNA methylation patterns across the entire genome. Our preliminary analysis showed globally a low level of methylation distributed in various sequence contexts across ancyromonad genomes. Moreover, our rough comparison between the methylation of all the TEs and protein coding genes of the model species *Ancyromonas sigmoides* revealed different patterns in the average methylation level of these genomic regions. Therefore, downstream analyses are needed in order to characterize these methylation profiles in detail and to study the distribution of methylation across genomic features of different types and origins. However, we were able to detect 5mC specific DNA methyltransferases (DNMTs) in only three of the seven species of ancyromonad encode proteins bearing 5mC MTase domain. This raises the possibility that DNMTs have been lost or evolved beyond recognition in certain ancyromonad species which is puzzling given the detected DNA methylation in all the species. Further analysis will help us to discriminate if this

DNA methylation is random or shows a particular pattern. In addition, ancyromonads encode enzymes related to other DNA modification marks such as 6mA and 4mC, opening the possibility to the existence of other layers of DNA methylation in these organisms.

Moreover, to perform a study of the molecular responses to environmental variation we used *Ancyromonas sigmoides* as a study model. This species has been isolated from coastal and deep sea sediments but also soil environments (Tikhonenkov, Mazei, and Mylnikov 2006; Yubuki et al. 2023) suggesting its ability to adapt to diverse environmental conditions. Our analysis of the gene expression under different conditions of temperature, salinity and oxygen availability provides the first insights into the molecular basis of the environmental versatility of this organism.

Under high-temperature conditions, *A. sigmoides* activate genes associated with DNA repair and ubiquitin transferase, suggesting a response to thermal stress-induced DNA damage. In contrast, genes linked to translation and protein folding were downregulated, indicating a possible trade-off between growth and DNA maintenance. In contrast, Low-temperature conditions triggered the overexpression of a diverse range of genes, with the most notable being the upregulation of genes associated with proteasomal and endocytic-vacuolar pathways. This suggests an effort to maintain protein homeostasis and adapt to colder temperatures, aligning with previous observations of temperature-driven adjustments in yeast (Isasa et al. 2016). The resilience to low temperatures could be crucial for *A. sigmoides*, particularly when considering its reported presence in deep-sea sediments with cold (3ºC) and high-pressure conditions (Živaljić et al. 2018).

Salinity shifts led to variations in the expression of genes enriched in categories related to ion transport and cytoskeleton assembly, but also several genes with unknown functions possibly involved in the strategies into this organism's ability to cope with ion imbalances.

Under low-oxygen conditions, *A. sigmoides* exhibited less expression changes than in any other tested condition contrasted to the control. Among the overexpressed genes we observed several proteins involved in the tricarboxylic acid cycle, electron transport chain. We also observed that genes encoding heme-binding proteins and

oxygen-dependent enzymes were overexpressed, pointing to a fine-tuned response to optimize oxygen utilization.

Altogether these results suggest that this species is able to thrive under low oxygen and that its responses to the reduction of oxygen involve a metabolic adjustment possibly ensuring the production of energy and the synthesis of building blocks even when oxygen is limited. Intriguingly, during this experiment *A. sigmoides* displayed differential expression of retrotransposon proteins under different environmental shifts, highlighting the dynamic nature of the genome of this organism and suggesting an important role of TEs in the genetic variation and molecular responses of this species. This finding underscores the importance of further investigating the genomic context of these elements and studying their potential influence on adjacent genes.

The further integrated analysis of these expression patterns with the generated methylation datasets will help us to understand if DNA methylation participates into the silencing of transposable elements as it has been observed in diverse eukaryotes (de Mendoza, Lister, and Bogdanovic 2019) and help us to understand is there is a role of 5mC in this species' biology. Alternatively other epigenetic marks not addressed in our study could be playing more important roles in the genome regulation of this organism. Overall our preliminary results provided the first insights into the molecular mechanisms driving *A. sigmoides* genome-environment interactions, contributing to our understanding of how these protists thrive in diverse ecological niches

Perspectives

Based on the results previously discussed, several perspectives for further studies in the diversity and evolution of the (epi)genome using orphan protists as study models are proposed below.

Rooting and solving the eukaryotic tree of life

The CRuMs and ancyromonad proteomes generated in this thesis are valuable resources for refining the reconstruction of the eukaryotic tree of life (eToL) using a different set of markers and methods such as bayesian phylogenetic reconstruction. Employing our ALE dataset could be also beneficial in order to test plausible eToL backbones and root position into the overall likelihood of the evolutionary histories of highly conserved gene families. Finally, our work has put in perspective the importance of including a balanced dataset in comparative genomics efforts aiming to study ancient evolutionary events of the eukaryotic domain. In particular the putative relationships between ancyromonads, malawimonads and Metamonada need to be revisited after improving the sampling and sequencing effort for representatives of Malawimonadida.

Investigation of unknown genes with uncharacterised functions

A more detailed analysis of the proteins restricted to mantamonas and ancyromonads respectively could be done to gain insight to explore their potential biological roles. A first question that can be addressed is for example if these protein coding genes have been evolving under purifying or diversifying selection. This kind of analysis can be done for proteins that are conserved in at least two species comparing the rate of evolution of synonym and non-synonymous sites along their coding sequences. How these genes expressed under different environmental conditions in *Ancyromonas sigmoides* and whether their expression patterns can be correlated to those of known genes are also questions that can be addressed directly by comparing the evidence generated in this work. In addition the analysis of the expression patterns of taxonomically restricted genes in different Ancyromonad species would also be very informative to answer if

these genes are being used more or less in a particular condition or if their coexpression patterns could help us to indicate their potential relationships with other genes.

Validation and further Investigation of prokaryotic gene transfer in ancyromonad genomes

Further phylogenetic analyses are needed to deepen our understanding of the evolutionary history of ancyromonad genes with a prokaryotic ancestry identified in this work. Some questions that can be addressed from these phylogenies are for example what is the identity and nature of the LGT donor and how much this protein has changed since the transfer occurred. In addition, a detailed and systematic characterization of the genomic context, gene architecture and expression patterns of these putative LGTs could shed light into the process of assimilation of these transfers into the genomes of ancyromonads. In the future, it would be also important to perform ecological studies to study ancyromonad interaction with other organisms addressing the following questions: Do ancyromonad have symbionts?, If yes what is the nature of this symbiosis?.

Investigating ancyromonad epigenetic diversity, DNA methylation and beyond

The further analysis of the domain architecture and phylogenetic affiliation of the DNA methyltransferases (DNMTs) identified in ancyromonads is still required to understand their origin and their relationship with other DNMTs in the tree of life. In addition a detailed analysis of the methylation patterns in different genomic regions is also necessary to investigate whether DNA methylation that we observed in the ancyromonad genomes follows a specific pattern or is randomly distributed across the genome. Possible comparisons could include the contrast of DNA methylation profiles across different genomic regions related to different functions, with a different age or type of origins. The comparison of the methylation levels and the expression patterns measured in our experiment in *Ancyromonas sigmoides* will further help us to

understand if there is a relationship between DNA methylation and gene silencing, particularly for transposable elements (TEs) with dynamic expression in *A. sigmoides*.

Based on the presence of other DNA modification writers found in our analysis, it would also be interesting to explore other potential epigenetic marks not addressed in this Thesis that could play important roles in the genome regulation of *A. sigmoides.* Nanopore sequencing could be employed to explore the existence of different DNA modifications (such as 6mA and 4mC) across the genome of *Acyromonas sigmoides*and other species. Finally, a fascinating prospect is to explore the prevalence of other epigenetic marks such as histone modifications and non coding RNA systems in ancyromonads a first step would consist in characterizing the presence of proteins involved in different epigenetic mechanisms in comparison to other eukaryotic lineages would shed light into the commonalities and differences of the epigenetic machinery of ancyromonads.

In depth analysis of the mobilome of Ancyromonas sigmoides

In this study, we have shown that a significant portion of the *Ancyromonas sigmoides* genome is composed of mobile genetic elements. Furthermore, we have found evidence that some of these elements remain active and exhibit dynamic expression patterns in response to various environmental conditions. These findings raise several questions. The first for example is how ancient are these elements and how they have been evolving in the genome of this organism. A sensitive comparison against specialized repeat and viral databases could also shed light into the origin of the mobile element reservoir in *A. sigmoides*. Second, based on the expression activity of many of these elements a question worthy to explore is if the expression of these elements is correlated to the expression of nearby genes and if these genes have known functions. A third question that could be approached is the potential intraspecific variation in of mobile elements between isolates of *A. sigmoides.* This could be explored in isolates from different locations or in isolates after several generations of growth in diverse controlled conditions. These studies would shed light into the specific impact of mobile elements in the genome diversity of *A. sigmoides* in a shorter evolutionary scale.

9. Concluding remarks

The primary goal of my thesis was to explore the eukaryotic (epi)genomic diversity and reduce the gaps in our understanding of deep eukaryotic evolution by integrating new and diverse model species as study models combining culture and genomic approaches for the exploration of orphan protists.

The successful generation, assembly, and curation of the first nuclear genomes for new species from the genus *Mantamonas* and the Ancyromonadida clade, marked an essential step in expanding our knowledge of these enigmatic organisms. We have provided a comprehensive view of their genome architectures characterized by diverse coding density, uncovering taxonomically restricted gene families and anciently retained protein machineries.

Our results support the placement of *Mantamonas* within the CRuMs supergroup. In contrast, the place ancyromonads into a global eukaryotic phylogeny remains enigmatic, however our analysis supports a and early divergence of this lineage within the tree, near one of the most conceivable candidate positions for the root of the eToL. The improved genomic representation of these clades adds valuable data for further phylogenomic analyses aiming to unravel the intricate relationships among major eukaryotic lineages.

By reconstructing the evolutionary history of these genomes in a wide scale using phylogenetic reconciliation analyses, we inferred that the last ancyromonad common ancestor has undergone important events of gene family turnover of proteins involved in cytoskeleton and signal transduction systems as well as the origination of an important proportion of genes. This analysis also shed light into the impact of gene duplication, transfer, loss, and origination on the diversification of ancyromonad genomic repertoires in comparison with other major eukaryotic clades, emphasizing the unique evolutionary trajectory of this protist lineage which in spite of its deep morphological conservation exhibited a wide variation of their gene content.

Notably, the identification of gene families retained in ancyromonads and distantly related species within Diphoda suggests two non mutually exclusive scenarios:

an intricate history of lateral gene transfers and an ancient origin in the last eukaryotic common ancestor of these species. Specifically, whether and how the gene content of the Last Eukaryotic Common Ancestor has been reducing or expanding through the diversification of major lineages and to which extent LGT was important in the early diversification of eukaryotes is an intriguing question.

In addition, several putative LGTs from bacteria and archaea were identified through the evolutionary history of ancyromonads, ranging from different predicted activities and biological processes, suggesting prokaryote to ancyromonad LGT has been important in the acquisition of diverse functions.

Moreover, we were not able to track canonical DNMTs across all the surveyed ancyromonad species, however our preliminary exploration of the DNA methylation patterns across ancyromonads suggested that all ancyromonads display low cytosine methylation with a mosaic distribution across their genomes. More analyses are therefore needed to conclude in which extent 5mC is a conserved epigenetic mark across these flagellates. Additionally, the conservation of other DNA modification proteins suggest that other methylation marks could coexist in the genomes of these organisms.

Through an experimental set up we have observed that *Ancyromonas sigmoides* can grow well during hypoxia, under low temperature, and under salinity changes but limitedly under high temperatures. The gene expression profiles from this experiment hint at the adaptive strategies employed by ancyromonads under a changing environment, including the differential expression of genes involved in central metabolism, structural cellular machinery and informational processes. Interestingly, we observed changes in the expression of DNA integrase genes from transposable elements during certain conditions that highlights the dynamic nature of the genome of this species. The integration of this evidence will open avenues for further functional studies in this species as well as genomic, methylome and transcriptomic comparative analyses across eukaryotes.

10. French summary

Évolution de l'(épi)génome des eucaryotes: perspectives offertes par les protistes orphelins.

Introduction et objectifs

Depuis leur diversification à partir d'un ancêtre commun (Eme et al. 2014; Dacks et al. 2016), les eucaryotes se sont divisés en plusieurs lignées majeures, également connues sous le nom de supergroupes (Fig. 12 Burki et al. 2020). La découverte de lignées profondes au sein de l'arbre des eucaryotes a considérablement amélioré notre compréhension de la diversité et l'évolution de ce domaine de la vie (Janouškovec et al. 2017 ; Brown et al. 2018 ; Schön et al. 2021 ; Lax et al. 2018 ; Galindo, López-García, et Moreira 2022 ; Tikhonenkov et al. 2022 ; Eglit et al. 2023).

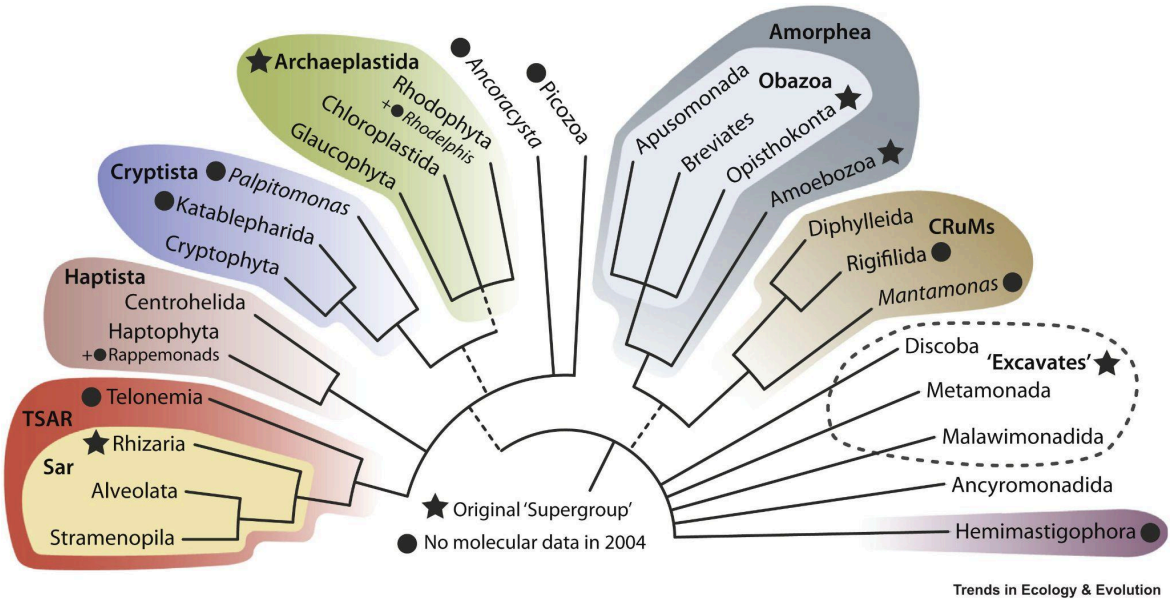

Figure 12. Modèle actuel de l'arbre de la vie eukaryota (Burki et al. 2020).

Les branches orphelines de l'arbre des eucaryotes jouent un rôle crucial pour comprendre l'émergence des caractères uniques aux eucaryotes et leur évolution tout au long de la diversification des eucaryotes en supergroupes. Les branches orphelines sont également indispensables pour reconstruire de façon fine les caractéristiques possédées par le dernier ancêtre commun des eucaryotes.

À ce jour, la recherche en génomique fonctionnelle et comparative des eucaryotes s'est principalement concentrée sur une infime fraction de lignées, limitant de façon drastique notre appréhension de la diversité et l'histoire évolutive des eucaryotes(Sibbald et Archibald 2017 ; Blaxter et al. 2022 ; Richter et al. 2022). L'objectif principal de cette thèse fut d'explorer la diversité (épi)génomique des eucaryotes et de combler les lacunes dans notre compréhension de l'évolution profonde de ce domaine de la vie. Notre stratégie a consisté en l'intégration d'espèces nouvelles et diverses de protistes orphelins comme modèles d'étude, en combinant des approches de culture et de génomique comparée.

Le projet est divisé en quatre parties; la première et la seconde se concentrent sur l'expansion des données génomiques des mantamonas et ancyromonads. Dans la troisième partie, nous étudions les principaux processus évolutifs qui ont conduit aux répertoires génétiques des différents membres du clade Ancyromonadida depuis leur origine jusqu'à nos jours. Enfin, la quatrième partie étudie pour la première fois la méthylation de l'ADN des génomes d'ancyromonades, et leur lien avec les variations d'expression en réponse aux changements environnementaux.

Manuscrit 1. De nouvelles espèces de *Mantamonas* éclairent l'évolution ancienne de composants clés des eucaryotes

Mantamonas est un genre de flagellés marins qui semblait initialement apparenté aux lignées Apusomonadida et Ancyromonadida (Glücksman et al. 2011., Cependant, les analyses phylogénétiques récentes l'ont placé dans un nouveau supergroupe nommé CRuMs, comprennant également Collodictyonidae et Rigifilidae (Brown et al. 2018). Ce groupe est important car c'est le groupe frère des Amorphea, auquel appartiennent les opisthokontes, les amoebozoa, les bréviaires et les apusomonades. Avant cette étude, les CRuMs n'étaient représentés que par trois espèces caractérisées par des données transcriptomiques partielles. Ici, nous avons isolé et décrit deux nouvelles espèces : *Mantamonas sphyraenae* et *Mantamonas vickermani*.

En combinant du séquençage PacBio et Illumina, nous avons assemblé une séquence génomique contiguë et quasi complète pour *M. sphyraenae*, dont la taille est estimée à 25 Mbp. Le génome de cette espèce contient 9,416 gènes codant pour des protéines. Le transcriptome séquencé de *M. vickermani* est également très complet.

L'analyse phylogénétique, utilisant 182 marqueurs protéiques, a confirmé la monophylie du genre *Mantamonas* au sein des CRuMs. Environ 1700 familles de gènes sont conservées dans toutes les espèces de CRuMs. En outre, les protéines conservées et uniques de *Mantamonas* comprennent environ 4000 familles de gènes, dont la plupart ont une fonction inconnue. Enfin, la présence de paralogues rares de protéines du système de trafic membranaire, telles que le complexe AP5 et la syntaxine 17 dans les espèces de *Mantamonas*, suggère la conservation de machineries protéiques d'origine ancienne.

Ce travail constitue une amélioration significative de la représentation génomique des CRuMs ainsi qu'une base importante pour des études fonctionnelles comparatives ultérieures. Ce manuscrit a été publié dans le journal *ScientificData* le 9 septembre 2023.

Manuscrit 2. La nature répétitive du génome d'*Ancyromonas sigmoides,* et les indices sur les origines de ses protéines.

Les ancyromonades sont phagotrophes et se nourrissent de procaryotes présents dans leur environnement naturel (Saville-Kent 1882 ; Heiss, Walker et Simpson 2011). Des espèces de ce groupe ont déjà été observées dans des sédiments benthiques d'environnements marins et d'eau douce ou dans des échantillons de sol à travers le monde (Yubuki et al. 2023). Depuis qu'ils ont été décrits, ils sont restés difficiles à placer dans la phylogénies eucaryotes et ne présentent aucune affinité avec un quelconque supergroupe (Atkins, McArthur et Teske 2000 ; Cavalier-Smith et Chao 2003 ; Paps et al. 2013 ; Torruella, Moreira et López-García 2017 ; Brown et al. 2018). Pour ces raisons, les ancyromonades sont considérés depuis longtemps comme une branche orpheline de l'arbre de la vie eucaryote.

Malgré leur importance évolutive, la connaissance générale sur la biologie de ces organismes est très limitée. Dans ce travail, nous présentons un assemblage contigu du génome nucléaire d'*Ancyromonas sigmoides*, l'espèce type du clade Ancyromonadida. Ce génome, généré en combinant des lectures Illumina et Pacbio, présente une taille de 39 Mbp.

Il est intéressant de noter qu'environ 29 % du génome est constitué de répétitions, principalement des familles d'éléments transposables inconnues. Sur les 11138 gènes codant pour des protéines identifiées, seuls 56 % présentaient une homologie détectable avec des gènes trouvés chez d'autres eucaryotes. De manière intrigante, 1212 gènes d'*A. sigmoides* partagent une relation évolutive étroite avec des gènes procaryotes et viraux, soulignant que les transferts latéraux de gènes ont probablement contribué de façon importante à l'acquisition de nouvelles fonctionnalités chez cette espèce.

Manuscrit 3 . Reconstruction de l'histoire évolutive profonde des répertoires génomiques des ancyromonades.

Comme mentionné ci-dessus, les ancyromonades constituent un groupe diversifié de flagellés vivant en liberté et leur divergence par rapport à tous les supergroupes eucaryotes connus leur confère une grande importance pour élucider l'évolution profonde du domaine eucaryote (Yubuki et al. 2023; Burki et al. 2020).

Dans cette étude, nous avons séquencé les génomes nucléaires de six espèces du clade Ancyromonadida, révélant leur architecture génomique et leur contenu en gènes. Dans la reconstruction phylogénétique de cet étude (Figure 13a), les ancyromonades forment un groupe monophylétique qui se ramifie comme un clade frère comprenant Metamonada et *Gefionella okellyi* (le seul représentant de Malawimonadida dans notre ensemble de données). Hemmimastoigophora se place comme branche sœur du cer dernier clade. Tous ces groupes ont des morphologies très différentes, des modes de vie variés et leurs positions phylogénétiques ont été historiquement difficiles à démêler. La reconstruction de l'histoire évolutive de ces génomes, depuis la divergence des Ancyromonadida avec le reste des supergroupes eucaryotes, a été réalisée à l'aide d'une méthode de réconciliation phylogénétique. Cette analyse suggère que le dernier ancêtre commun des ancyromonades a connu des événements majeurs de renouvellement de familles de gènes impliquées dans la traduction, le cytosquelette et les systèmes de transduction des signaux, entre autres fonctions. L'analyse suggère également que l'origine d'une grande proportion de gènes, comparée à d'autres lignées eucaryotes, est antérieure à la diversification des Ancyromonadida (Figure 13bc). Ce travail a mis en évidence l'impact de la duplication, du transfert, de la perte et de l'origine des familles de gènes sur la diversification des ancyromonades par rapport à d'autres clades eucaryotes majeurs, soulignant la trajectoire évolutive unique de cette lignée de protistes qui, en dépit de sa profonde conservation morphologique, présente une grande variation dans le contenu des gènes. En outre, l'identification de plusieurs transferts latéraux de gènes (TLG) putatifs provenant de bactéries et d'archées au cours de l'histoire évolutive des ancyromonades, avec différentes activités et processus biologiques prédits suggère que ce processus a été important pour l'acquisition de diverses fonctions.

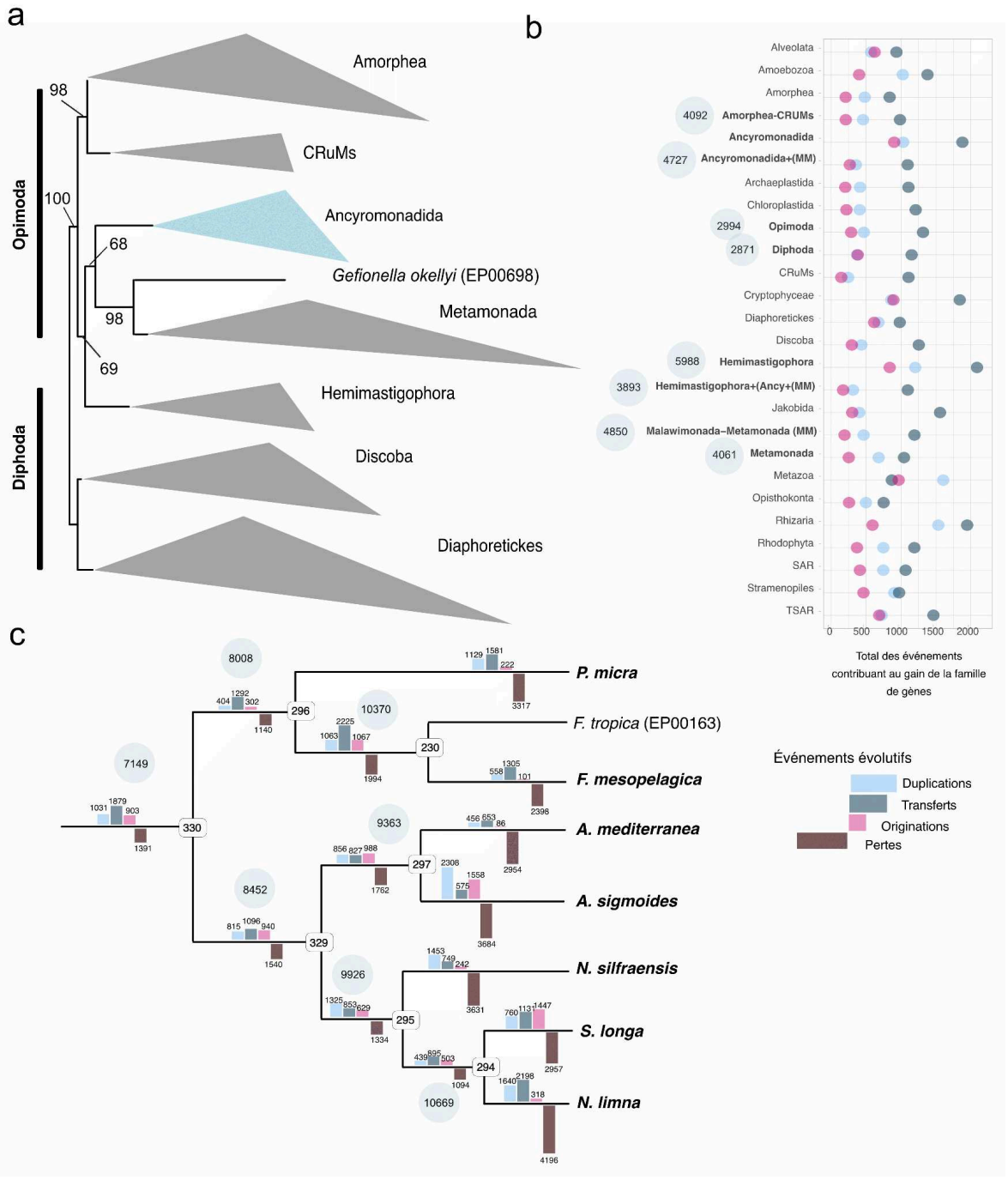

Figure 13. Position phylogénétique des nouvelles espèces d'Ancyromonadida dans l'arbre des eucaryotes et évolution de leurs répertoires génomiques. a) Phylogénie des eucaryotes basée sur l'analyse de 766 marqueurs, 205 taxons et 47611 sites et inférée à l'aide d'IQ-Tree selon le modèle LG + C60 + G. b) Nombre et types d'événements évolutifs déduits par ALE pour expliquer les gains de familles de gènes à la base des différentes lignées eucaryotes. Le soutien aux branches a été estimé à l'aide de 1 000 répliques bootstrap. b) Nombre et types d'événements évolutifs déduits par l'ALE pour expliquer les gains de familles de gènes à la base de différentes lignées eucaryotes. Les nombres représentent la somme de tous les événements évolutifs déduits par l'ALE (voir le code couleur) et le nombre de copies de gènes (dans les cercles) pour chaque nœud. c) Processus d'évolution du contenu génétique dans le clade des ancyromonades. Les nombres dans les divisions du cladogramme correspondent au nom du nœud interne.

Manuscrit 4. Exploration de la diversité épigénomique chez les ancyromonades

Les marques épigénétiques d'un organisme jouent un rôle fondamental dans sa plasticité phénotypique et sa réponse aux changements environnementaux (Weiner and Katz 2021). Actuellement, la recherche sur l'épigénétique et la régulation du génome dans la diversité eucaryote montre une grande variation entre les organismes appartenant à différents supergroupes (de Mendoza, Lister, and Bogdanovic 2019; Hoguin et al. 2023), mais reste limitée en raison de la disponibilité des ressources génomiques dans la diversité eucaryote. Les ancyromonades sont un modèle intéressant à étudier en raison de leur position profonde dans l'arbre de vie des eucaryotes, ce qui suggère qu'ils peuvent avoir conservé des caractéristiques plus proches de celles du dernier ancêtre commun eucaryote que des lignées plus dérivées telles que les animaux ou les plantes.

Dans cette étude, la diversité de la méthylation de l'ADN chez sept espèces d'ancyromonades a été explorée en utilisant le séquençage au bisulfite, qui permet de distinguer les cytosines méthylées (5mC) des cytosines non méthylées dans l'ensemble d'un génome. Les ADN méthyltransférases canoniques (DNMT) n'ont pas pu être retrouvées dans toutes les espèces étudiées. En revanche, l'exploration préliminaire des profils de méthylation de l'ADN chez les ancyromonades suggère que les génomes de ces organismes sont faiblement méthylés. La majorité des sites méthylés dans les génomes de ces espèces présentent des taux inférieurs à 20 %. D'autre part, la conservation d'autres protéines modifiant l'ADN suggère que d'autres marques de méthylation peuvent coexister dans les génomes de ces organismes.

Nous avons mené une expérience pour étudier les profils d'expression génique d'*Ancyromonas sigmoides*, l'espèce type du clade, en cas de variations de la salinité (+ et moins 3% respectivement, de la température (30º et 4º respectivement) et des basses niveaux d'oxygène.

A partir de cette expérience il a été constaté que *Ancyromonas sigmoides* peut bien se développer en hypoxie, à basse température et en cas de changements de salinité, mais difficilement à haute température. Nous avons identifié environ 6000 gènes différentiellement exprimés dans les diverses conditions, sur un total de 11138

gènes du génome de cette espèce (Figure 14). Les profils d'expression géniques de cette expérience indiquent les stratégies d'adaptation employées par les ancyromonades dans un environnement changeant, y compris l'expression différentielle des gènes impliqués dans le métabolisme central, la machinerie cellulaire structurelle et les processus d'information.

Il est à noter que les gènes avec des domaines fonctionnels de type intégrase appartenant probablement à des rétrotransposons ont montré une expression différentielle dans de multiples conditions, ce qui laisse supposer un rôle potentiel de ces éléments dans la régulation du génome chez *A. sigmoides* et souligne la nature dynamique du génome de cette espèce.

L'intégration de ces données ouvrira la voie à d'autres études fonctionnelles chez cette espèce ainsi qu'à des analyses comparatives génomiques, méthylomiques et transcriptomiques chez les eucaryotes. Cette étude est en cours et actuellement en préparation, mais les résultats préliminaires mettent en lumière la diversité épigénétique et la dynamique génomique au sein de ce groupe énigmatique de protistes.

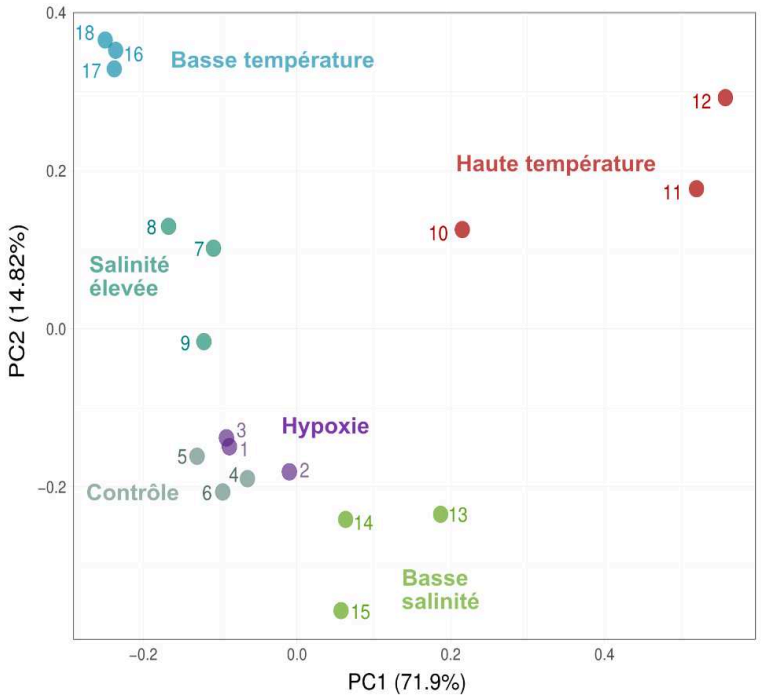

Figure 14. Regroupement et projection des échantillons par analyse des coordonnées principales en fonction du profil d'expression génétique de chaque échantillon. Les nombres d'expression ont été normalisés par la transformation de la médiane des rapports de DESeq2.

Conclusions

La génération et l'assemblage des premiers génomes nucléaires de *Mantamonas* et du clade Ancyromonadida ont constitué une étape essentielle dans l'élargissement de nos connaissances sur ces organismes énigmatiques. Ce travail fournit une vue d'ensemble de l'architecture de leurs génomes, caractérisés par une densité de codage variée, et des familles de gènes uniques à ces taxons, ainsi que des machineries protéiques conservées depuis longtemps.

La place des ancyromonades dans une phylogénie globale des eucaryotes reste énigmatique, mais notre analyse soutient une divergence précoce de cette lignée dans l'arbre, à proximité de l'une des positions candidates les plus concevables pour la racine de l'arbre des eucaryotes.

L'analyse évolutive de ces génomes à grande échelle a montré qu'un nombre important d'origines de familles de gènes est antérieur à la diversification de Ancyromonadida. Des changements majeurs dans les familles de gènes impliqués dans la transduction du signal et les protéines associées au cytosquelette ont également contribué à une grande variation du contenu en gènes parmi les espèces modernes. En outre, certaines ancyromonades ont également acquis plusieurs gènes de procaryotes qui ont pu faciliter leur adaptation aux environnements benthiques. Parmi ces acquisitions, on trouve des protéines impliquées dans le transport et le métabolisme du nitrate donnant de nouveaux indices concernant les rôles écologiques des ancyromonades dans le cycle des nutriments.

De plus, bien que nous n'ayons pas trouvé de DNMT dans les génomes de toutes les espèces d'ancyromonades, par séquençage au bisulfite nous avons observé que toutes ces espèces ont un faible niveau de méthylation. En outre, l'expérience de variation environnementale chez *Ancyromonas sigmoides* suggère que cette espèce est capable de prospérer dans des conditions à faible teneur en oxygène et que ses réponses à la raréfaction de l'oxygène impliquent un ajustement métabolique pour assurer la production d'énergie et la synthèse des lipides dans ces conditions.

Il est intéressant de noter que dans cette expérience, *A. sigmoides* a montré une expression différentielle des protéines de rétrotransposons sous différents changements environnementaux, soulignant la nature dynamique du génome de cet organisme et suggérant un rôle important de ces éléments dans la variation génétique et les réponses moléculaires de cette espèce. La comparaison de ces profils d'expression avec les ensembles de données de méthylation générés nous aidera à comprendre si la méthylation de l'ADN est impliquée dans la répression des éléments transposables, comme cela a été observé chez plusieurs eucaryotes (de Mendoza et al. 2020), et nous aidera à comprendre si le 5mC joue un rôle dans la biologie de cette espèce. Alternativement, d'autres marques épigénétiques non abordées dans notre étude pourraient jouer des rôles plus importants dans la régulation du génome de cet organisme.

Dans l'ensemble, les résultats préliminaires de ces études ont fourni les premières informations sur les mécanismes moléculaires à l'origine des interactions entre le génome et l'environnement d'*A. sigmoides*, contribuant ainsi à une meilleure compréhension de la manière dont ces protistes prospèrent dans diverses niches écologiques.

Cette thèse offre un aperçu nouveau de l'évolution ancienne des eucaryotes, a fourni des données génomiques, transcriptomiques et de méthylome fondamentales pour un groupe clé des eucaryotes encore sous-étudié, et pose également de nouvelles questions pour la recherche future sur l'épigénome eucaryote.

11. Literature cited

Adl, Sina M., David Bass, Christopher E. Lane, Julius Lukeš, Conrad L. Schoch, Alexey Smirnov, Sabine Agatha, et al. 2019. “Revisions to the Classification, Nomenclature, and Diversity of Eukaryotes.” *The Journal of Eukaryotic Microbiology* 66 (1): 4–119.

Adl, Sina M., Brian S. Leander, Alastair G. B. Simpson, John M. Archibald, O. Roger Anderson, David Bass, Samuel S. Bowser, et al. 2007. “Diversity, Nomenclature, and Taxonomy of Protists.” *Systematic Biology* 56 (4): 684–89.

Adl, Sina M., Alastair G. B. Simpson, Christopher E. Lane, Julius Lukeš, David Bass, Samuel S. Bowser, Matthew W. Brown, et al. 2012. “The Revised Classification of Eukaryotes.” *The Journal of Eukaryotic Microbiology* 59 (5): 429–93.

Akıl, Caner, and Robert C. Robinson. 2018. “Genomes of Asgard Archaea Encode Profilins That Regulate Actin.” *Nature* 562 (7727): 439–43.

Al Jewari, Caesar, and Sandra L. Baldauf. 2023. “An Excavate Root for the Eukaryote Tree of Life.” *Science Advances* 9 (17): eade4973.

Alonge, Michael, Sebastian Soyk, Srividya Ramakrishnan, Xingang Wang, Sara Goodwin, Fritz J. Sedlazeck, Zachary B. Lippman, and Michael C. Schatz. 2019. “RaGOO: Fast and Accurate Reference-Guided Scaffolding of Draft Genomes.” *Genome Biology* 20 (1): 224.

Ammar, Ron, Dax Torti, Kyle Tsui, Marinella Gebbia, Tanja Durbic, Gary D. Bader, Guri Giaever, and Corey Nislow. 2012. “Chromatin Is an Ancient Innovation Conserved between Archaea and Eukarya.” *eLife* 1 (December): e00078.

Archibald, John M., Matthew B. Rogers, Michael Toop, Ken-Ichiro Ishida, and Patrick J. Keeling. 2003. “Lateral Gene Transfer and the Evolution of Plastid-Targeted Proteins in the Secondary Plastid-Containing Alga Bigelowiella Natans.” *Proceedings of the National Academy of Sciences of the United States of America* 100 (13): 7678–83.

Atkins, M. S., A. G. McArthur, and A. P. Teske. 2000. “Ancyromonadida: A New Phylogenetic Lineage among the Protozoa Closely Related to the Common Ancestor of Metazoans, Fungi, and Choanoflagellates (Opisthokonta).” *Journal of Molecular*

Evolution 51 (3): 278–85.

Barreat, Jose Gabriel Nino, and Aris Katzourakis. 2022. “Paleovirology of the DNA Viruses of Eukaryotes.” *Trends in Microbiology* 30 (3): 281–92.

Bellas, Christopher, Thomas Hackl, Marie-Sophie Plakolb, Anna Koslová, Matthias G. Fischer, and Ruben Sommaruga. 2023. “Large-Scale Invasion of Unicellular Eukaryotic Genomes by Integrating DNA Viruses.” *Proceedings of the National Academy of Sciences of the United States of America* 120 (16): e2300465120.

Bestor, T. H., V. L. Chandler, and A. P. Feinberg. 1994. “Epigenetic Effects in Eukaryotic Gene Expression.” *Developmental Genetics* 15 (6): 458–62.

Betts, Holly C., Mark N. Puttick, James W. Clark, Tom A. Williams, Philip C. J. Donoghue, and Davide Pisani. 2018. “Integrated Genomic and Fossil Evidence Illuminates Life’s Early Evolution and Eukaryote Origin.” *Nature Ecology & Evolution* 2 (10): 1556–62.

Bewick, Adam J., Brigitte T. Hofmeister, Rob A. Powers, Stephen J. Mondo, Igor V. Grigoriev, Timothy Y. James, Jason E. Stajich, and Robert J. Schmitz. 2019. “Diversity of Cytosine Methylation across the Fungal Tree of Life.” *Nature Ecology & Evolution* 3 (3): 479–90.

Blaxter, Mark, John M. Archibald, Anna K. Childers, Jonathan A. Coddington, Keith A. Crandall, Federica Di Palma, Richard Durbin, et al. 2022. “Why Sequence All Eukaryotes?” *Proceedings of the National Academy of Sciences of the United States of America* 119 (4). https://doi.org/10.1073/pnas.2115636118.

Bochtler, Matthias, and Humberto Fernandes. 2021. “DNA Adenine Methylation in Eukaryotes: Enzymatic Mark or a Form of DNA Damage?” *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 43 (3): e2000243.

Boscaro, Vittorio, and Patrick J. Keeling. 2023. “How Ciliates Got Their Nuclei.” *Proceedings of the National Academy of Sciences of the United States of America* 120 (7): e2221818120.

Boulias, Konstantinos, and Eric Lieberman Greer. 2022. “Means, Mechanisms and Consequences of Adenine Methylation in DNA.” *Nature Reviews. Genetics* 23 (7): 411–28.

Bourque, Guillaume, Kathleen H. Burns, Mary Gehring, Vera Gorbunova, Andrei

Seluanov, Molly Hammell, Michaël Imbeault, et al. 2018. “Ten Things You Should Know about Transposable Elements.” *Genome Biology* 19 (1): 199.

Brown, Matthew W., Aaron A. Heiss, Ryoma Kamikawa, Yuji Inagaki, Akinori Yabuki, Alexander K. Tice, Takashi Shiratori, et al. 2018. “Phylogenomics Places Orphan Protistan Lineages in a Novel Eukaryotic Super-Group.” *Genome Biology and Evolution* 10 (2): 427–33.

Burger, Gertraud, Michael W. Gray, Lise Forget, and B. Franz Lang. 2013. “Strikingly Bacteria-like and Gene-Rich Mitochondrial Genomes throughout Jakobid Protists.” *Genome Biology and Evolution* 5 (2): 418–38.

Burki, Fabien. 2014. “The Eukaryotic Tree of Life from a Global Phylogenomic Perspective.” *Cold Spring Harbor Perspectives in Biology* 6 (5): a016147.

Burki, Fabien, Andrew J. Roger, Matthew W. Brown, and Alastair G. B. Simpson. 2020. “The New Tree of Eukaryotes.” *Trends in Ecology & Evolution* 35 (1): 43–55.

Capy, P., G. Gasperi, C. Biémont, and C. Bazin. 2000. “Stress and Transposable Elements: Co-Evolution or Useful Parasites?” *Heredity* 85 (Pt 2) (August): 101–6.

Caron, David A., Alexandra Z. Worden, Peter D. Countway, Elif Demir, and Karla B. Heidelberg. 2009. “Protists Are Microbes Too: A Perspective.” *The ISME Journal* 3 (1): 4–12.

Cavalier-Smith, Thomas, Ema E. Chao, Elizabeth A. Snell, Cédric Berney, Anna Maria Fiore-Donno, and Rhodri Lewis. 2014. “Multigene Eukaryote Phylogeny Reveals the Likely Protozoan Ancestors of Opisthokonts (animals, Fungi, Choanozoans) and Amoebozoa.” *Molecular Phylogenetics and Evolution* 81 (December): 71–85.

Cavalier-Smith, Thomas, and Ema E-Y Chao. 2003. “Phylogeny of Choanozoa, Apusozoa, and Other Protozoa and Early Eukaryote Megaevolution.” *Journal of Molecular Evolution* 56 (5): 540–63.

Cerón-Romero, Mario A., Miguel M. Fonseca, Leonardo de Oliveira Martins, David Posada, and Laura A. Katz. 2022. “Phylogenomic Analyses of 2,786 Genes in 158 Lineages Support a Root of the Eukaryotic Tree of Life between Opisthokonts and All Other Lineages.” *Genome Biology and Evolution* 14 (8). https://doi.org/10.1093/gbe/evac119.

Chaux, Nicole de la, and Andreas Wagner. 2011. “BEL/Pao Retrotransposons in Metazoan Genomes.” *BMC Evolutionary Biology* 11 (June): 154.

Clark, James W., and Philip C. J. Donoghue. 2018. “Whole-Genome Duplication and Plant Macroevolution.” *Trends in Plant Science* 23 (10): 933–45.

Cohen, Phoebe A., and Robin B. Kodner. 2022. “The Earliest History of Eukaryotic Life: Uncovering an Evolutionary Story through the Integration of Biological and Geological Data.” *Trends in Ecology & Evolution* 37 (3): 246–56.

Csurös, Miklós. 2010. “Count: Evolutionary Analysis of Phylogenetic Profiles with Parsimony and Likelihood.” *Bioinformatics* 26 (15): 1910–12.

Danchin, Etienne G. J. 2016. “Lateral Gene Transfer in Eukaryotes: Tip of the Iceberg or of the Ice Cube?” *BMC Biology*.

Delsuc, Frédéric, Henner Brinkmann, and Hervé Philippe. 2005. “Phylogenomics and the Reconstruction of the Tree of Life.” *Nature Reviews. Genetics* 6 (5): 361–75.

Derelle, Evelyne, Conchita Ferraz, Stephane Rombauts, Pierre Rouzé, Alexandra Z. Worden, Steven Robbens, Frédéric Partensky, et al. 2006. “Genome Analysis of the Smallest Free-Living Eukaryote Ostreococcus Tauri Unveils Many Unique Features.” *Proceedings of the National Academy of Sciences of the United States of America* 103 (31): 11647–52.

Derelle, Romain, Guifré Torruella, Vladimír Klimeš, Henner Brinkmann, Eunsoo Kim, Čestmír Vlček, B. Franz Lang, and Marek Eliáš. 2015. “Bacterial Proteins Pinpoint a Single Eukaryotic Root.” *Proceedings of the National Academy of Sciences of the United States of America* 112 (7): E693–99.

Donoghue, Philip C. J., Chris Kay, Anja Spang, Gergely Szöllősi, Anna Nenarokova, Edmund R. R. Moody, Davide Pisani, and Tom A. Williams. 2023. “Defining Eukaryotes to Dissect Eukaryogenesis.” *Current Biology: CB* 33 (17): R919–29.

Doolittle, W. F. 1998. “You Are What You Eat: A Gene Transfer Ratchet Could Account for Bacterial Genes in Eukaryotic Nuclear Genomes.” *Trends in Genetics: TIG* 14 (8): 307–11.

Dorrell, Richard G., Adrien Villain, Benoît Perez-Lamarque, Guillemette Audren de Kerdrel, Giselle McCallum, Andrew K. Watson, Ouardia Ait-Mohamed, et al. 2021.

"Phylogenomic Fingerprinting of Tempo and Functions of Horizontal Gene Transfer within Ochrophytes." *Proceedings of the National Academy of Sciences of the United States of America* 118 (4). https://doi.org/10.1073/pnas.2009974118.

Eglit, Yana, Takashi Shiratori, Jon Jerlström-Hultqvist, Kelsey Williamson, Andrew J. Roger, Ken-Ichiro Ishida, and Alastair G. B. Simpson. 2023. "Meteora Sporadica, a Protist with Incredible Cell Architecture, Is Related to Hemimastigophora." *bioRxiv*. https://doi.org/10.1101/2023.08.13.553137.

Eme, Laura, Susan C. Sharpe, Matthew W. Brown, and Andrew J. Roger. 2014. "On the Age of Eukaryotes: Evaluating Evidence from Fossils and Molecular Clocks." *Cold Spring Harbor Perspectives in Biology* 6 (8). https://doi.org/10.1101/cshperspect.a016139.

Eme, Laura, Anja Spang, Jonathan Lombard, Courtney W. Stairs, and Thijs J. G. Ettema. 2018. "Archaea and the Origin of Eukaryotes." *Nature Reviews. Microbiology* 16 (2): 120.

Eme, Laura, Daniel Tamarit, Eva F. Caceres, Courtney W. Stairs, Valerie De Anda, Max E. Schön, Kiley W. Seitz, et al. 2023. "Inference and Reconstruction of the Heimdallarchaeial Ancestry of Eukaryotes." *Nature* 618 (7967): 992–99.

Emms, David M., and Steven Kelly. 2019. "OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics." *Genome Biology* 20 (1): 238.

Fedoroff, N. V. 2012. "Transposable Elements, Epigenetics, and Genome Evolution." *Science*. https://doi.org/10.1126/science.338.6108.758.

Fernández, Rosa, and Toni Gabaldón. 2020. "Gene Gain and Loss across the Metazoan Tree of Life." *Nature Ecology & Evolution* 4 (4): 524–33.

Feschotte, Cédric. 2008. "Transposable Elements and the Evolution of Regulatory Networks." *Nature Reviews. Genetics* 9 (5): 397–405.

Feschotte, Cédric. 2023. "Transposable Elements: McClintock's Legacy Revisited." *Nature Reviews. Genetics*, September. https://doi.org/10.1038/s41576-023-00652-3.

Feschotte, Cédric, and Ellen J. Pritham. 2007. "DNA Transposons and the Evolution of Eukaryotic Genomes." *Annual Review of Genetics* 41: 331–68.

Finnegan, D. J. 1989. "Eukaryotic Transposable Elements and Genome Evolution." *Trends*

in Genetics: TIG 5 (4): 103–7.

Fischer, Matthias G., and Thomas Hackl. 2016. “Host Genome Integration and Giant Virus-Induced Reactivation of the Virophage Mavirus.” *Nature* 540 (7632): 288–91.

Fischer, Matthias G., and Curtis A. Suttle. 2011. “A Virophage at the Origin of Large DNA Transposons.” *Science* 332 (6026): 231–34.

Frank, John A., and Cédric Feschotte. 2017. “Co-Option of Endogenous Viral Sequences for Host Cell Function.” *Current Opinion in Virology* 25 (August): 81–89.

Freire, Borja, Susana Ladra, and Jose R. Parama. 2021. “Memory-Efficient Assembly Using Flye.” *IEEE/ACM Transactions on Computational Biology and Bioinformatics / IEEE, ACM* PP (September). https://doi.org/10.1109/TCBB.2021.3108843.

Fukuda, Yasuhiro, and Toshinobu Suzaki. 2015. “Unusual Features of Dinokaryon, the Enigmatic Nucleus of Dinoflagellates.” In *Marine Protists: Diversity and Dynamics*, edited by Susumu Ohtsuka, Toshinobu Suzaki, Takeo Horiguchi, Noritoshi Suzuki, and Fabrice Not, 23–45. Tokyo: Springer Japan.

Galindo, Luis Javier, Purificación López-García, and David Moreira. 2022. “First Molecular Characterization of the Elusive Marine Protist Meteora Sporadica.” *Protist* 173 (4): 125896.

Genuth, Naomi R., and Maria Barna. 2018. “Heterogeneity and Specialized Functions of Translation Machinery: From Genes to Organisms.” *Nature Reviews. Genetics* 19 (7): 431–52.

Gibney, E. R., and C. M. Nolan. 2010. “Epigenetics and Gene Expression.” *Heredity* 105 (1): 4–13.

Gilbert, W. 1978. “Why Genes in Pieces?” *Nature* 271 (5645): 501.

Glöckner, Gernot, Norbert Hülsmann, Michael Schleicher, Angelika A. Noegel, Ludwig Eichinger, Christoph Gallinger, Jan Pawlowski, et al. 2014. “The Genome of the Foraminiferan Reticulomyxa Filosa.” *Current Biology: CB* 24 (1): 11–18.

Glücksman, Edvard, Elizabeth A. Snell, Cédric Berney, Ema E. Chao, David Bass, and Thomas Cavalier-Smith. 2011. “The Novel Marine Gliding Zooflagellate Genus Mantamonas (Mantamonadida Ord. N.: Apusozoa).” *Protist* 162 (2): 207–21.

Gornik, Sebastian G., Kristina L. Ford, Terrence D. Mulhern, Antony Bacic, Geoffrey I. McFadden, and Ross F. Waller. 2012. “Loss of Nucleosomal DNA Condensation Coincides with Appearance of a Novel Nuclear Protein in Dinoflagellates.” *Current Biology: CB* 22 (24): 2303–12.

Grattepanche, Jean-David, Laura M. Walker, Brittany M. Ott, Daniela L. Paim Pinto, Charles F. Delwiche, Christopher E. Lane, and Laura A. Katz. 2018. “Microbial Diversity in the Eukaryotic SAR Clade: Illuminating the Darkness Between Morphology and Molecular Data.” *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 40 (4): e1700198.

Grau-Bové, Xavier, Cristina Navarrete, Cristina Chiva, Thomas Pribasnig, Meritxell Antó, Guifré Torruella, Luis Javier Galindo, et al. 2022. “A Phylogenetic and Proteomic Reconstruction of Eukaryotic Chromatin Evolution.” *Nature Ecology & Evolution* 6 (7): 1007–23.

Gücker, B., and H. Fischer. 2003. “Flagellate and Ciliate Distribution in Sediments of a Lowland River: Relationships with Environmental Gradients and Bacteria.” *Aquatic Microbial Ecology: International Journal* 31: 67–76.

Haeckel, Ernst. 1866. *Generelle Morphologie der Organismen: Bd. Allgemeine Anatomie der Organismen*. G. Reimer.

Haeckel, Ernst. 2004. *Haeckel’s Art Forms from Nature*. Courier Corporation.

Haghani, Amin, Caesar Z. Li, Todd R. Robeck, Joshua Zhang, Ake T. Lu, Julia Ablaeva, Victoria A. Acosta-Rodríguez, et al. 2023. “DNA Methylation Networks Underlying Mammalian Traits.” *Science* 381 (6658): eabq5693.

Hampton, Hannah G., Bridget N. J. Watson, and Peter C. Fineran. 2020. “The Arms Race between Bacteria and Their Phage Foes.” *Nature* 577 (7790): 327–36.

Hanschen, Erik R., Blake T. Hovde, and Shawn R. Starkenburg. 2020. “An Evaluation of Methodology to Determine Algal Genome Completeness.” *Algal Research* 51 (October): 102019.

Heiss, Aaron A., Martin Kolisko, Fleming Ekelund, Matthew W. Brown, Andrew J. Roger, and Alastair G. B. Simpson. 2018. “Combined Morphological and Phylogenomic Re-Examination of Malawimonads, a Critical Taxon for Inferring the Evolutionary

History of Eukaryotes." *Royal Society Open Science* 5 (4): 171707.

Heiss, Aaron A., Giselle Walker, and Alastair G. B. Simpson. 2010. "Clarifying the Taxonomic Identity of a Phylogenetically Important Group of Eukaryotes: Planomonas Is a Junior Synonym of Ancyromonas." *The Journal of Eukaryotic Microbiology* 57 (3): 285–93.

Heiss, Aaron A., Giselle Walker, and Alastair G. B. Simpson. 2011. "The Ultrastructure of Ancyromonas, a Eukaryote without Supergroup Affinities." *Protist* 162 (3): 373–93.

Hernández, Greco, Christopher G. Proud, Thomas Preiss, and Armen Parsyan. 2012. "On the Diversification of the Translation Apparatus across Eukaryotes." *Comparative and Functional Genomics* 2012 (May): 256848.

Hoguin, Antoine, Feng Yang, Agnès Groisillier, Chris Bowler, Auguste Genovesio, Ouardia Ait-Mohamed, Fabio Rocha Jimenez Vieira, and Leila Tirichine. 2023. "The Model Diatom Phaeodactylum Tricornutum Provides Insights into the Diversity and Function of Microeukaryotic DNA Methyltransferases." *Communications Biology* 6 (1): 253.

Holliday, Robin. 2006. "Epigenetics: A Historical Overview." *Epigenetics: Official Journal of the DNA Methylation Society* 1 (2): 76–80.

Hooff, Jolien J. E. van, and Laura Eme. 2023. "Lateral Gene Transfer Leaves Lasting Traces in Rhizaria." *bioRxiv*. https://doi.org/10.1101/2023.01.27.525846.

Hooff, Jolien Je van, Eelco Tromer, Leny M. van Wijk, Berend Snel, and Geert Jpl Kops. 2017. "Evolutionary Dynamics of the Kinetochore Network in Eukaryotes as Revealed by Comparative Genomics." *EMBO Reports* 18 (9): 1559–71.

Horváth, Vivien, Miriam Merenciano, and Josefa González. 2017. "Revisiting the Relationship between Transposable Elements and the Eukaryotic Stress Response." *Trends in Genetics: TIG* 33 (11): 832–41.

Hug, Laura A., Brett J. Baker, Karthik Anantharaman, Christopher T. Brown, Alexander J. Probst, Cindy J. Castelle, Cristina N. Butterfield, et al. 2016. "A New View of the Tree of Life." *Nature Microbiology* 1 (April): 16048.

Hunter, Richard G., Khatuna Gagnidze, Bruce S. McEwen, and Donald W. Pfaff. 2015. "Stress and the Dynamic Genome: Steroids, Epigenetics, and the Transposome."

Proceedings of the National Academy of Sciences of the United States of America 112 (22): 6828–33.

Imachi, Hiroyuki, Masaru K. Nobu, Nozomi Nakahara, Yuki Morono, Miyuki Ogawara, Yoshihiro Takaki, Yoshinori Takano, et al. 2020. “Isolation of an Archaeon at the Prokaryote-Eukaryote Interface.” *Nature* 577 (7791): 519–25.

Innan, Hideki, and Fyodor Kondrashov. 2010. “The Evolution of Gene Duplications: Classifying and Distinguishing between Models.” *Nature Reviews. Genetics* 11 (2): 97–108.

Irwin, Nicholas A. T., Alexandros A. Pittis, Thomas A. Richards, and Patrick J. Keeling. 2022. “Systematic Evaluation of Horizontal Gene Transfer between Eukaryotes and Viruses.” *Nature Microbiology* 7 (2): 327–36.

Irwin, Nicholas A. T., and Thomas A. Richards. 2023. “Self-Assembling Viral Histones Unravel Early Nucleosome Evolution.” *bioRxiv*. https://doi.org/10.1101/2023.09.20.558576.

Isasa, Marta, Clara Suñer, Miguel Díaz, Pilar Puig-Sàrries, Alice Zuin, Anne Bichman, Steven P. Gygi, Elena Rebollo, and Bernat Crosas. 2016. “Cold Temperature Induces the Reprogramming of Proteolytic Pathways in Yeast.” *The Journal of Biological Chemistry* 291 (4): 1664–75.

Iyer, Lakshminarayan M., Saraswathi Abhiman, and L. Aravind. 2011. “Chapter 2 - Natural History of Eukaryotic DNA Methylation Systems.” In *Progress in Molecular Biology and Translational Science*, edited by Xiaodong Cheng and Robert M. Blumenthal, 101:25–104. Academic Press.

Iyer, Lakshminarayan M., Dapeng Zhang, and L. Aravind. 2016. “Adenine Methylation in Eukaryotes: Apprehending the Complex Evolutionary History and Functional Potential of an Epigenetic Modification.” *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 38 (1): 27–40.

Janouškovec, Jan, Denis V. Tikhonenkov, Fabien Burki, Alexis T. Howe, Forest L. Rohwer, Alexander P. Mylnikov, and Patrick J. Keeling. 2017. “A New Lineage of Eukaryotes Illuminates Early Mitochondrial Genome Reduction.” *Current Biology: CB* 27 (23): 3717–24.e5.

Jeudy, Sandra, Sofia Rigou, Jean-Marie Alempic, Jean-Michel Claverie, Chantal Abergel, and Matthieu Legendre. 2020. “The DNA Methylation Landscape of Giant Viruses.” *Nature Communications* 11 (1): 2657.

Kang, Seungho, Alexander K. Tice, Courtney W. Stairs, Robert E. Jones, Daniel J. G. Lahr, and Matthew W. Brown. 2021. “The Integrin-Mediated Adhesive Complex in the Ancestor of Animals, Fungi, and Amoebae.” *Current Biology: CB* 31 (14): 3073–85.e3.

Karnkowska, Anna, Vojtěch Vacek, Zuzana Zubáčová, Sebastian C. Treitli, Romana Petrželková, Laura Eme, Lukáš Novák, et al. 2016. “A Eukaryote without a Mitochondrial Organelle.” *Current Biology: CB* 26 (10): 1274–84.

Katz, Laura A. 2012. “Origin and Diversification of Eukaryotes.” *Annual Review of Microbiology* 66 (July): 411–27.

Kent, S. W. 1882. “A Manual of the Infusoria (D. Bogue, London).” Vols.

Knoll, Andrew H. 2014. “Paleobiological Perspectives on Early Eukaryotic Evolution.” *Cold Spring Harbor Perspectives in Biology* 6 (1). https://doi.org/10.1101/cshperspect.a016121.

Koonin, Eugene V. 2007. “The Biological Big Bang Model for the Major Transitions in Evolution.” *Biology Direct* 2 (August): 21.

Koonin, Eugene V. 2010. “Taming of the Shrewd: Novel Eukaryotic Genes from RNA Viruses.” *BMC Biology* 8 (January): 2.

Krueger, Felix, and Simon R. Andrews. 2011. “Bismark: A Flexible Aligner and Methylation Caller for Bisulfite-Seq Applications.” *Bioinformatics* 27 (11): 1571–72.

Lamka, Gina F., Avril M. Harder, Mekala Sundaram, Tonia S. Schwartz, Mark R. Christie, J. Andrew DeWoody, and Janna R. Willoughby. 2022. “Epigenetics in Ecology, Evolution, and Conservation.” *Frontiers in Ecology and Evolution* 10. https://doi.org/10.3389/fevo.2022.871791.

Lax, Gordon, Yana Eglit, Laura Eme, Erin M. Bertrand, Andrew J. Roger, and Alastair G. B. Simpson. 2018. “Hemimastigophora Is a Novel Supra-Kingdom-Level Lineage of Eukaryotes.” *Nature* 564 (7736): 410–14.

Leclercq, Sébastien, Julien Thézé, Mohamed Amine Chebbi, Isabelle Giraud, Bouziane Moumen, Lise Ernenwein, Pierre Grève, Clément Gilbert, and Richard Cordaux.

2016. “Birth of a W Sex Chromosome by Horizontal Transfer of Wolbachia Bacterial Symbiont Genome.” *Proceedings of the National Academy of Sciences of the United States of America* 113 (52): 15036–41.

Leger, Michelle M., Laura Eme, Courtney W. Stairs, and Andrew J. Roger. 2018. “Demystifying Eukaryote Lateral Gene Transfer (Response to Martin 2017 DOI: 10.1002/bies.201700115).” *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 40 (5): e1700242.

Li, En, and Yi Zhang. 2014. “DNA Methylation in Mammals.” *Cold Spring Harbor Perspectives in Biology* 6 (5): a019133.

Lizarraga, Ayelen, Zach Klapholz O’Brown, Konstantinos Boulias, Lara Roach, Eric Lieberman Greer, Patricia J. Johnson, Pablo H. Strobl-Mazzulla, and Natalia de Miguel. 2020. “Adenine DNA Methylation, 3D Genome Organization, and Gene Expression in the Parasite *Trichomonas Vaginalis*.” *Proceedings of the National Academy of Sciences of the United States of America* 117 (23): 13033–43.

Løbner-Olesen, Anders, Ole Skovgaard, and Martin G. Marinus. 2005. “Dam Methylation: Coordinating Cellular Processes.” *Current Opinion in Microbiology* 8 (2): 154–60.

López-García, Purificación, Laura Eme, and David Moreira. 2017. “Symbiosis in Eukaryotic Evolution.” *Journal of Theoretical Biology* 434 (December): 20–33.

López-García, Purificación, and David Moreira. 2015. “Open Questions on the Origin of Eukaryotes.” *Trends in Ecology & Evolution* 30 (11): 697–708.

López-García, Purificación, and David Moreira. 2023. “The Symbiotic Origin of the Eukaryotic Cell.” *Comptes Rendus Biologies* 346 (May): 55–73.

Manni, Mosè, Matthew R. Berkeley, Mathieu Seppey, Felipe A. Simão, and Evgeny M. Zdobnov. 2021. “BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes.” *Molecular Biology and Evolution* 38 (10): 4647–54.

Margulis, L. 1971. “Symbiosis and Evolution.” *Scientific American* 225 (2): 48–57.

Margulis, L., and D. Bermudes. 1985. “Symbiosis as a Mechanism of Evolution: Status of Cell Symbiosis Theory.” *Symbiosis* 1: 101–24.

Marsh, Joseph A., and Sarah A. Teichmann. 2010. “How Do Proteins Gain New Domains?”

Genome Biology 11 (7): 126.

Maumus, Florian, Andrew E. Allen, Corinne Mhiri, Hanhua Hu, Kamel Jabbari, Assaf Vardi, Marie-Angèle Grandbastien, and Chris Bowler. 2009. "Potential Impact of Stress Activated Retrotransposons on Genome Evolution in a Marine Diatom." *BMC Genomics* 10 (December): 624.

Maurer-Alcalá, Xyrus X., and Laura A. Katz. 2015. "An Epigenetic Toolkit Allows for Diverse Genome Architectures in Eukaryotes." *Current Opinion in Genetics & Development*. https://doi.org/10.1016/j.gde.2015.10.005.

Mayr, Christine. 2016. "Evolution and Biological Roles of Alternative 3'UTRs." *Trends in Cell Biology* 26 (3): 227–37.

McCLINTOCK, B. 1950. "The Origin and Behavior of Mutable Loci in Maize." *Proceedings of the National Academy of Sciences of the United States of America* 36 (6): 344–55.

McClintock, Barbara. 1984. "The Significance of Responses of the Genome to Challenge." *Science* 226 (4676): 792–801.

Mendoza, Alex de, Ryan Lister, and Ozren Bogdanovic. 2019. "Evolution of DNA Methylome Diversity in Eukaryotes." *Journal of Molecular Biology*, November. https://doi.org/10.1016/j.jmb.2019.11.003.

Mendoza, Alex de, Daniel Poppe, Sam Buckberry, Jahnvi Pflueger, Caroline B. Albertin, Tasman Daish, Stephanie Bertrand, et al. 2021. "The Emergence of the Brain Non-CpG Methylation System in Vertebrates." *Nature Ecology & Evolution* 5 (3): 369–78.

Mendoza, Alex de, and Arnau Sebé-Pedrós. 2019. "Origin and Evolution of Eukaryotic Transcription Factors." *Current Opinion in Genetics & Development* 58-59 (October): 25–32.

Mereschkowsky, C. 1905. "Über Natur Und Ursprung Der Chromatophoren Im Pflanzenreiche." *Biol Centralbl* 25: 593.

Mereschkowsky, C. 1910. "Theorie Der Zwei Plasmaarten Als Grundlage Der Symbiogenesis, Einer Neuen Lehre von Der Entstehung Der Organismen." *Biologisches Centralblatt* 30: 278–88,289–303,322–74,353–67.

Miller, Wolfgang J., John F. McDonald, Danielle Nouaud, and Dominique Anxolabéhère.

2000. “Molecular Domestication — More than a Sporadic Episode in Evolution.” In *Transposable Elements and Genome Evolution*, edited by John F. McDonald, 197–207. Dordrecht: Springer Netherlands.

Minh, Bui Quang, Heiko A. Schmidt, Olga Chernomor, Dominik Schrempf, Michael D. Woodhams, Arndt von Haeseler, and Robert Lanfear. 2020. “IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era.” *Molecular Biology and Evolution* 37 (5): 1530–34.

Misteli, Tom. 2020. “The Self-Organizing Genome: Principles of Genome Architecture and Function.” *Cell* 183 (1): 28–45.

More, Kiran, Christen M. Klinger, Lael D. Barlow, and Joel B. Dacks. 2020. “Evolution and Natural History of Membrane Trafficking in Eukaryotes.” *Current Biology: CB* 30 (10): R553–64.

Muñoz-Gómez, Sergio A., Edward Susko, Kelsey Williamson, Laura Eme, Claudio H. Slamovits, David Moreira, Purificación López-García, and Andrew J. Roger. 2022. “Site-and-Branch-Heterogeneous Analyses of an Expanded Dataset Favour Mitochondria as Sister to Known Alphaproteobacteria.” *Nature Ecology & Evolution* 6 (3): 253–62.

Nilsen, Timothy W. 2003. “The Spliceosome: The Most Complex Macromolecular Machine in the Cell?” *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 25 (12): 1147–49.

Ocaña-Pallarès, Eduard, Tom A. Williams, David López-Escardó, Alicia S. Arroyo, Jananan S. Pathmanathan, Eric Bapteste, Denis V. Tikhonenkov, Patrick J. Keeling, Gergely J. Szöllősi, and Iñaki Ruiz-Trillo. 2022. “Divergent Genomic Trajectories Predate the Origin of Animals and Fungi.” *Nature* 609 (7928): 747–53.

Oggenfuss, Ursula, and Daniel Croll. 2023. “Recent Transposable Element Bursts Are Associated with the Proximity to Genes in a Fungal Plant Pathogen.” *PLoS Pathogens* 19 (2): e1011130.

O’Malley, Maureen A., Jeremy G. Wideman, and Iñaki Ruiz-Trillo. 2016. “Losing Complexity: The Role of Simplification in Macroevolution.” *Trends in Ecology & Evolution* 31 (8): 608–21.

Paps, Jordi, Luis A. Medina-Chacón, Wyth Marshall, Hiroshi Suga, and Iñaki Ruiz-Trillo. 2013. “Molecular Phylogeny of Unikonts: New Insights into the Position of Apusomonads and Ancyromonads and the Internal Relationships of Opisthokonts.” *Protist* 164 (1): 2–12.

Parks, Donovan H., Maria Chuvochina, Christian Rinke, Aaron J. Mussig, Pierre-Alain Chaumeil, and Philip Hugenholtz. 2022. “GTDB: An Ongoing Census of Bacterial and Archaeal Diversity through a Phylogenetically Consistent, Rank Normalized and Complete Genome-Based Taxonomy.” *Nucleic Acids Research* 50 (D1): D785–94.

Pauling, Linus, Emile Zuckerkandl, Thormod Henriksen, and Rolf Lövstad. 1963. “Chemical Paleogenetics. Molecular ‘Restoration Studies’ of Extinct Forms of Life.” *Acta Chemica Scandinavica* 17 supl.: 9–16.

Philippe, Hervé, Philippe Lopez, Henner Brinkmann, Karine Budin, Agnés Germot, Jacqueline Laurent, David Moreira, Miklós Müller, and Hervé Le Guyader. 2000. “Early–branching or Fast–evolving Eukaryotes? An Answer Based on Slowly Evolving Positions.” *Proceedings of the Royal Society of London. Series B: Biological Sciences* 267 (1449): 1213–21.

Pittis, Alexandros A., and Toni Gabaldón. 2016. “Late Acquisition of Mitochondria by a Host with Chimaeric Prokaryotic Ancestry.” *Nature* 531 (7592): 101–4.

Ponce-Toledo, Rafael I., Philippe Deschamps, Purificación López-García, Yvan Zivanovic, Karim Benzerara, and David Moreira. 2017. “An Early-Branching Freshwater Cyanobacterium at the Origin of Plastids.” *Current Biology: CB* 27 (3): 386–91.

Prjibelski, Andrey, Dmitry Antipov, Dmitry Meleshko, Alla Lapidus, and Anton Korobeynikov. 2020. “Using SPAdes De Novo Assembler.” *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis … [et Al.]* 70 (1): e102.

Raymond, Jason, and Robert E. Blankenship. 2003. “Horizontal Gene Transfer in Eukaryotic Algal Evolution.” *Proceedings of the National Academy of Sciences of the United States of America*.

Rees, D. J., B. C. Emerson, P. Oromí, and G. M. Hewitt. 2001. “Reconciling Gene Trees with Organism History: The mtDNA Phylogeography of Three Nesotes Species (Coleoptera: Tenebrionidae) on the Western Canary Islands.” *Journal of Evolutionary*

Biology 14 (1): 139–47.

Richter, Daniel J., Cédric Berney, Jürgen F. H. Strassert, Yu-Ping Poh, Emily K. Herman, Sergio A. Muñoz-Gómez, Jeremy G. Wideman, Fabien Burki, and Colomban de Vargas. 2022. "EukProt: A Database of Genome-Scale Predicted Proteins across the Diversity of Eukaryotes." *Peer Community Journal* 2 (e56). https://doi.org/10.24072/pcjournal.173.

Ritter, Eleanore J., and Chad E. Niederhuth. 2021. "Intertwined Evolution of Plant Epigenomes and Genomes." *Current Opinion in Plant Biology* 61 (January): 101990.

Rodriguez, Fernando, Irina A. Yushenova, Daniel DiCorpo, and Irina R. Arkhipova. 2022. "Bacterial N4-Methylcytosine as an Epigenetic Mark in Eukaryotic DNA." *Nature Communications* 13 (1): 1072.

Roger, Andrew J., Edward Susko, and Michelle M. Leger. 2021. "Evolution: Reconstructing the Timeline of Eukaryogenesis." *Current Biology: CB* 31 (4): R193–96.

Rogozin, Igor B., Liran Carmel, Miklos Csuros, and Eugene V. Koonin. 2012. "Origin and Evolution of Spliceosomal Introns." *Biology Direct* 7 (April): 11.

Roitman, Sheila, Andrey Rozenberg, Tali Lavy, Corina P. D. Brussaard, Oded Kleifeld, and Oded Béjà. 2023. "Isolation and Infection Cycle of a Polinton-like Virus Virophage in an Abundant Marine Alga." *Nature Microbiology* 8 (2): 332–46.

Rothschild, L. J. 1989. "Protozoa, Protista, Protoctista: What's in a Name?" *Journal of the History of Biology* 22 (2): 277–305.

Saary, Paul, Alex L. Mitchell, and Robert D. Finn. 2020. "Estimating the Quality of Eukaryotic Genomes Recovered from Metagenomic Analysis with EukCC." *Genome Biology* 21 (1): 244.

Sabina, Jeffrey, and John H. Leamon. 2015. "Bias in Whole Genome Amplification: Causes and Considerations." *Methods in Molecular Biology* 1347: 15–41.

Salas-Leiva, Dayana E., Eelco C. Tromer, Bruce A. Curtis, Jon Jerlström-Hultqvist, Martin Kolisko, Zhenzhen Yi, Joan S. Salas-Leiva, et al. 2021. "Genomic Analysis Finds No Evidence of Canonical Eukaryotic DNA Processing Complexes in a Free-Living Protist." *Nature Communications* 12 (1): 6003.

Sánchez-Romero, María A., and Josep Casadesús. 2020. "The Bacterial Epigenome."

Nature Reviews. Microbiology 18 (1): 7–20.

Saville-Kent, William. 1882. *A Manual of the Infusoria: Plates*.

Schmitz, Robert J., Zachary A. Lewis, and Mary G. Goll. 2019. "DNA Methylation: Shared and Divergent Features across Eukaryotes." *Trends in Genetics: TIG* 35 (11): 818–27.

Sémon, Marie, and Kenneth H. Wolfe. 2007. "Consequences of Genome Duplication." *Current Opinion in Genetics & Development* 17 (6): 505–12.

Seppey, Mathieu, Mosè Manni, and Evgeny M. Zdobnov. 2019. "BUSCO: Assessing Genome Assembly and Annotation Completeness." *Methods in Molecular Biology* 1962: 227–45.

Shalchian-Tabrizi, K., W. Eikrem, D. Klaveness, D. Vaulot, M. A. Minge, F. Le Gall, K. Romari, et al. 2006. "Telonemia, a New Protist Phylum with Affinity to Chromist Lineages." *Proceedings. Biological Sciences / The Royal Society* 273 (1595): 1833–42.

Sibbald, Shannon J., and John M. Archibald. 2017. "More Protist Genomes Needed." *Nature Ecology & Evolution* 1 (5): 145.

Sibbald, Shannon J., Laura Eme, John M. Archibald, and Andrew J. Roger. 2020. "Lateral Gene Transfer Mechanisms and Pan-Genomes in Eukaryotes." *Trends in Parasitology* 36 (11): 927–41.

Sieber, Karsten B., Robin E. Bromley, and Julie C. Dunning Hotopp. 2017. "Lateral Gene Transfer between Prokaryotes and Eukaryotes." *Experimental Cell Research* 358 (2): 421–26.

Simpson, Alastair G. B. 2003. "Cytoskeletal Organization, Phylogenetic Affinities and Systematics in the Contentious Taxon Excavata (Eukaryota)." *International Journal of Systematic and Evolutionary Microbiology* 53 (Pt 6): 1759–77.

Simpson, Alastair G. B., and Andrew J. Roger. 2004. "The Real 'Kingdoms' of Eukaryotes." *Current Biology: CB* 14 (17): R693–96.

Skala, Anna Marie. 2014. "Retroviral DNA Transposition: Themes and Variations." *Microbiology Spectrum* 2 (5). https://doi.org/10.1128/microbiolspec.MDNA3-0005-2014.

Smith, David Roy, and Patrick J. Keeling. 2016. "Protists and the Wild, Wild West of Gene

Expression: New Frontiers, Lawlessness, and Misfits." *Annual Review of Microbiology* 70 (September): 161–78.

Spang, Anja. 2023. "Is an Archaeon the Ancestor of Eukaryotes?" *Environmental Microbiology* 25 (4): 775–79.

Stairs, Courtney W., Michelle M. Leger, and Andrew J. Roger. 2015. "Diversity and Origins of Anaerobic Metabolism in Mitochondria and Related Organelles." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 370 (1678): 20140326.

Stevens, Kathryn M., Jacob B. Swadling, Antoine Hocher, Corinna Bang, Simonetta Gribaldo, Ruth A. Schmitz, and Tobias Warnecke. 2020. "Histone Variants in Archaea and the Evolution of Combinatorial Chromatin Complexity." *Proceedings of the National Academy of Sciences of the United States of America* 117 (52): 33384–95.

Szitenberg, Amir, Soyeon Cha, Charles H. Opperman, David M. Bird, Mark L. Blaxter, and David H. Lunt. 2016. "Genetic Drift, Not Life History or RNAi, Determine Long-Term Evolution of Transposable Elements." *Genome Biology and Evolution* 8 (9): 2964–78.

Szöllõsi, Gergely J., Wojciech Rosikiewicz, Bastien Boussau, Eric Tannier, and Vincent Daubin. 2013. "Efficient Exploration of the Space of Reconciled Gene Trees." *Systematic Biology* 62 (6): 901–12.

Talbert, Paul B., and Steven Henikoff. 2010. "Histone Variants--Ancient Wrap Artists of the Epigenome." *Nature Reviews. Molecular Cell Biology* 11 (4): 264–75.

Tautz, Diethard, and Tomislav Domazet-Lošo. 2011. "The Evolutionary Origin of Orphan Genes." *Nature Reviews. Genetics* 12 (10): 692–702.

Tikhonenkov, Denis Victorovich, Yuri Alexandrovich Mazei, and Alexander Petrovich Mylnikov. 2006. "Species Diversity of Heterotrophic Flagellates in White Sea Littoral Sites." *European Journal of Protistology* 42 (3): 191–200.

Tikhonenkov, Denis V., Kirill V. Mikhailov, Ryan M. R. Gawryluk, Artem O. Belyaev, Varsha Mathur, Sergey A. Karpov, Dmitry G. Zagumyonnyi, et al. 2022. "Microbial Predators Form a New Supergroup of Eukaryotes." *Nature* 612 (7941): 714–19.

Todorovska, E. 2007. "Retrotransposons and Their Role in Plant—Genome Evolution." *Biotechnology, Biotechnological Equipment* 21 (3): 294–305.

Vosseberg, Julian, Jolien J. E. van Hooff, Marina Marcet-Houben, Anne van Vlimmeren, Leny M. van Wijk, Toni Gabaldón, and Berend Snel. 2021. “Timing the Origin of Eukaryotic Cellular Complexity with Ancient Duplications.” *Nature Ecology & Evolution* 5 (1): 92–100.

Vosseberg, Julian, Michelle Schinkel, Sjoerd Gremmen, and Berend Snel. 2022. “The Spread of the First Introns in Proto-Eukaryotic Paralogs.” *Communications Biology* 5 (1): 476.

Wallau, Gabriel Luz, Pierre Capy, Elgion Loreto, and Aurélie Hua-Van. 2014. “Genomic Landscape and Evolutionary Dynamics of Mariner Transposable Elements within the Drosophila Genus.” *BMC Genomics* 15 (1): 727.

Weiner, Agnes K. M., Mario A. Cerón-Romero, Ying Yan, and Laura A. Katz. 2020. “Phylogenomics of the Epigenetic Toolkit Reveals Punctate Retention of Genes across Eukaryotes.” *Genome Biology and Evolution* 12 (12): 2196–2210.

Weiner, Agnes K. M., and Laura A. Katz. 2021. “Epigenetics as Driver of Adaptation and Diversification in Microbial Eukaryotes.” *Frontiers in Genetics* 12 (March): 642220.

Whittaker, R. H., and L. Margulis. 1978. “Protist Classification and the Kingdoms of Organisms.” *Bio Systems* 10 (1-2): 3–18.

Williams, Tom A., Adrián A. Davín, Benoit Morel, Lénárd L. Szánthó, Anja Spang, Alexandros Stamatakis, Philip Hugenholtz, and Gergely J. Szöllősi. 2023. “Parameter Estimation and Species Tree Rooting Using ALE and GeneRax.” *Genome Biology and Evolution* 15 (7). https://doi.org/10.1093/gbe/evad134.

Woese, C. R. 1994. “There Must Be a Prokaryote Somewhere: Microbiology’s Search for Itself.” *Microbiological Reviews* 58 (1): 1–9.

Woese, C. R., and G. E. Fox. 1977. “Phylogenetic Structure of the Prokaryotic Domain: The Primary Kingdoms.” *Proceedings of the National Academy of Sciences of the United States of America* 74 (11): 5088–90.

Yabuki, Akinori, Ken-Ichiro Ishida, and Thomas Cavalier-Smith. 2013. “Rigifila Ramosa N. Gen., N. Sp., a Filose Apusozoan with a Distinctive Pellicle, Is Related to Micronuclearia.” *Protist* 164 (1): 75–88.

Yi, Soojin V., and Michael A. D. Goodisman. 2021. “The Impact of Epigenetic Information

on Genome Evolution." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 376 (1826): 20200114.

Yubuki, Naoji, Guifré Torruella, Luis Javier Galindo, Aaron A. Heiss, Maria Cristina Ciobanu, Takashi Shiratori, Ken-Ichiro Ishida, et al. 2023. "Molecular and Morphological Characterization of Four New Ancyromonad Genera and Proposal for an Updated Taxonomy of the Ancyromonadida." *The Journal of Eukaryotic Microbiology*, August, e12997.

Zemach, Assaf, Ivy E. McDaniel, Pedro Silva, and Daniel Zilberman. 2010. "Genome-Wide Evolutionary Analysis of Eukaryotic DNA Methylation." *Science* 328 (5980): 916–19.

Zhao, Sen, Fabien Burki, Jon Bråte, Patrick J. Keeling, Dag Klaveness, and Kamran Shalchian-Tabrizi. 2012. "Collodictyon--an Ancient Lineage in the Tree of Eukaryotes." *Molecular Biology and Evolution* 29 (6): 1557–68.

Zhou, Qiangwei, Jing-Quan Lim, Wing-Kin Sung, and Guoliang Li. 2019. "An Integrated Package for Bisulfite DNA Methylation Data Analysis with Indel-Sensitive Mapping." *BMC Bioinformatics* 20 (1): 47.

Zhou, Wanding, Gangning Liang, Peter L. Molloy, and Peter A. Jones. 2020. "DNA Methylation Enables Transposable Element-Driven Genome Expansion." *Proceedings of the National Academy of Sciences of the United States of America* 117 (32): 19359–66.

Živaljić, Suzana, Alexandra Schoenle, Frank Nitsche, Manon Hohlfeld, Julia Piechocki, Farina Reif, Marwa Shumo, et al. 2018. "Survival of Marine Heterotrophic Flagellates Isolated from the Surface and the Deep Sea at High Hydrostatic Pressure: Literature Review and Own Experiments." *Deep-Sea Research. Part II, Topical Studies in Oceanography* 148 (February): 251–59.

Zuckerkandl, E., and L. Pauling. 1965. "Molecules as Documents of Evolutionary History." *Journal of Theoretical Biology* 8 (2): 357–66.

12. Supplementary material of manuscript 3

Genome sequencing and assembly

We selected seven ancyromonad species of the DEEM team culture collection representing the main branches within the clade: *Fabomonas mesopelagica (Fmes), Ancyromonas sigmoides (Asig), Striomonas longa (Nlon), Planomonas micra (Pmic), Nutomonas limna (Nlim), Ancyromonas mediterranea (Amed) and Nyramonas silfraensis (Nsil).* In order to obtain high quality genomic data of these strains, we used three sequencing workflows.

1) Sequencing workflow 1 (SW1): nucleic acids were purified from the non axenic cultures of each species and used for DNA, and RNA Illumina HiSeq sequencing.

2) SW2: with the aim to reduce the prokaryotic cells (i.e. contaminants) present in the non-axenic cultures and to increase the quantity of DNA required for single molecule sequencing, we assessed two parallel methods consisting of a combination of cell sorting and Whole Genome Amplification (WGA) approaches. Approximately ten samples of 200 cells were obtained from the culture of each species using Fluorescent Activated Cell sorting (FACs). These samples were further used to test the yield of two WGA protocols based on the True Prime and Repli G kits, respectively, and sequenced using Illumina HiSeq (2x150 bp).

3) SW3: based on the results of the FACs + WGA experiments and genome assembly qualities, we generated genomic data for the seven ancyromonad species using the Repli G WGA protocol for DNA amplification, followed by fragment size selection and long-read sequencing (minIT Oxford Nanopore Technologies (ONT) platform).

To evaluate the yield of each sequencing workflow, an independent draft genome assembly was first generated for each dataset obtained from SW1 and SW2 using SPAdes(Prjibelski et al. 2020). The draft assemblies were further assessed considering their contiguity and completeness using the eukaryota_obd10 database of BUSCO (Seppey, Manni, and Zdobnov 2019). Evident contaminant sequences from the SW1 and SW2 only short read assemblies were identified based using Blobtools with taxonomic assignment based on diamond blastx searches with (> 80 % identity), query coverage (> 50%) and e-value (< $1xe^{-15}$ of BLAST hits against viral, bacterial and archaeal sequences from the non-redundant nucleotide database and excluded from the assembly.

Furthermore we generated a hybrid assembly for each species combining the data generated in all the sequencing workflows (Table S1 and Table S2). To handle the coverage artifacts generated by the WGA in the SW2 datasets we normalized the coverage by down-sampling the reads over the high-depth areas of the genome with BBNorm, using as “target read abundance” the precalculated coverage obtained from the SW1 non-amplified genomic library (100-200x). We then pooled short read data coming from the SW1 and SW2 downsampled and generated assembled contigs using SPAdes. Paralerly, the SW3 datasets were basecalled using Guppy5 with the *super accuracy* model and the long reads were then assembled using Flye (Freire, Ladra, and Parama 2021) (with the uneven coverage aware mode --meta). The long-read assemblies were corrected using racon and medaka by subsequent rounds of mapping the long reads. Finally the long-read assembly was used to scaffold and patch the accurate contigs coming from the short reads coassembly workflow using RagTag (Alonge et al. 2019).

Table S1. Data from sequencing workflows and genome assemblies comparison.

Species ID	sequencing strategy ID	sequenced bases after QC (Gb)	Assembly size (Mb)	number of contigs	N50	BUSCO score eukaryota (odb10)
Fmes	*SW1*	2.04	40.05	3,696	58,353	87.3
Amed	*SW1*	5.1	61.38	7,204	18,017	42.5
Slon	*SW1*	5.05	22.84	9,651	9,990	58.4
Nlim	*SW1*	11.95	0.31	3,820	4,024	10.3
Pmic	*SW1*	4.78	26.98	8,767	24,448	89.2
Nsil	*SW1*	16.5	20.71	7204	26,179	84.7
Fmes	*SW2_1 RG*	1.566	43.64	41610	5,842	71.4
Fmes	*SW2_1 TP*	1.989	5.33	4581	6,455	2.7
Amed	*SW2_1 RG*	2.488	10.10	2906	26,988	3.5
Amed	*SW2_1 TP*	1.439	4.21	2336	8,488	2.0
Slon	*SW2_1 RG*	1.758	25.65	71503	1,234	2.7
Slon	*SW2_1 TP*	1.72	13.90	110649	143	30.2
Pmic	*SW2_1 RG*	1.619	5.58	8123	1,044	2.4
Pmic	*SW2_1 TP*	1.817	11.84	6773	18,400	5.5
Nsil	*SW2_1 RG*	1.634	2.27	3307	2221	0.8
Nsil	*SW2_1 TP*	1.585	2.49	2532	8698	1.6
Fmes	*SW2_2 RG*	7.2	58.24	29704	20814	89.4
Amed	*SW2_2 RG*	6.6	14.50	7789	27795	6.7
Slon	*SW2_2 RG*	7.3	40.72	27328	4501	43.5
Nlim	*SW2_2 RG*	6.3	20.65	25814	2670	40.4
Pmic	*SW2_2 RG*	6.6	11.38	11157	3761	4.3
Nsil	*SW2_2 RG*	6.9	19.33	9830	26779	10.2

Table S2. Nanopore datasets (SW3), Values after filtering on quality score (>9) (value for non quality filtered *fail+pass* files).

Species ID	sequencing strategy ID	basecalled bases (Gb)	number of reads	longest read	average read length	Reads N50/N90
Fmes	*SW3*	3.45 (4.3)	749,759	82,435	4,601	5,523 / 2639
Pmic	*SW3*	9.57 (12.8)	4,024,017	29,239	2,379	2,784 / 1,448
Amed	*SW3*	10.47 (12.87)	4,239,613	33,566	2,470	2,871 / 1,504
Nsil	*SW3*	31.4 (32.9)	14,692,424	565,029	2,244	2,782 / 1,312
Slon	*SW3*	4.16 (5.1)	759,226	55,244	3,279	4,325 / 1,903
Nlim	*SW3*	2.74 (3.07)	789,377	62,894	3,475	5,198 / 1,727

Genomes quality assessment

Table S3. Genome statistics after the two steps of contamination screening.

Species	Assembly length (Mbp)	Number of contigs	Longest contig (Kbp)	N count	Number of gaps	N50 (Kbp)	N50n
Fmes	38.02	893	522.894	658	27	107344	110
Pmic	24.99	1561	206.215	1912	94	32831	221
Asig	39.77	202	891.693	0	0	399306	35
Amed	25.17	8338	147.178	7435	648	3590	2023
Nsil	24.87	1745	145.932	6601	149	30355	250
Slon	24.76	4426	62.089	2941	279	9075	775
Nlim	38.07	6873	1710.564	822	78	8382	1089

The completeness of the final genomic assemblies (Table S3) was evaluated based on the identification of BUSCO markers on the genome and the comparison against the transcriptome of the same species. Briefly, the transcriptome was assembled de novo and decontaminated using BlobToolkit2. Clean transcripts were used to retrieve clean transcriptomic reads that were mapped to the genomic sequences, the proportion of clean reads from each species transcriptome was considered as a proxy of the representation of the gene space of the genome (Table S4).

Most of the sequenced genomes (Figure S1) harbor up to 75% of BUSCOs (including, complete single, complete duplicated and fragmented), except for *Amed*. In this species the transcriptome also displays a higher proportion of BUSCOs than the genomic sequence, indicating an incomplete representation of the genome for this species.

Similarly, the transcriptome of *Nsil* has a higher BUSCO score than the genome and 91% of the transcriptomic reads could be aligned back to the genome, indicating

incompleteness of the gene complement of this species. Nlim, had the highest proportion of fragmented BUSCOs. This could be explained in part by the presence of repeated elements, abundant in this species also observed by the fact that an important proportion of the transcritomic reads align to multiple regions of the genome. Finally several of the missing BUSCO markers are absent in several species, suggesting these markers could be really absent within those clades.

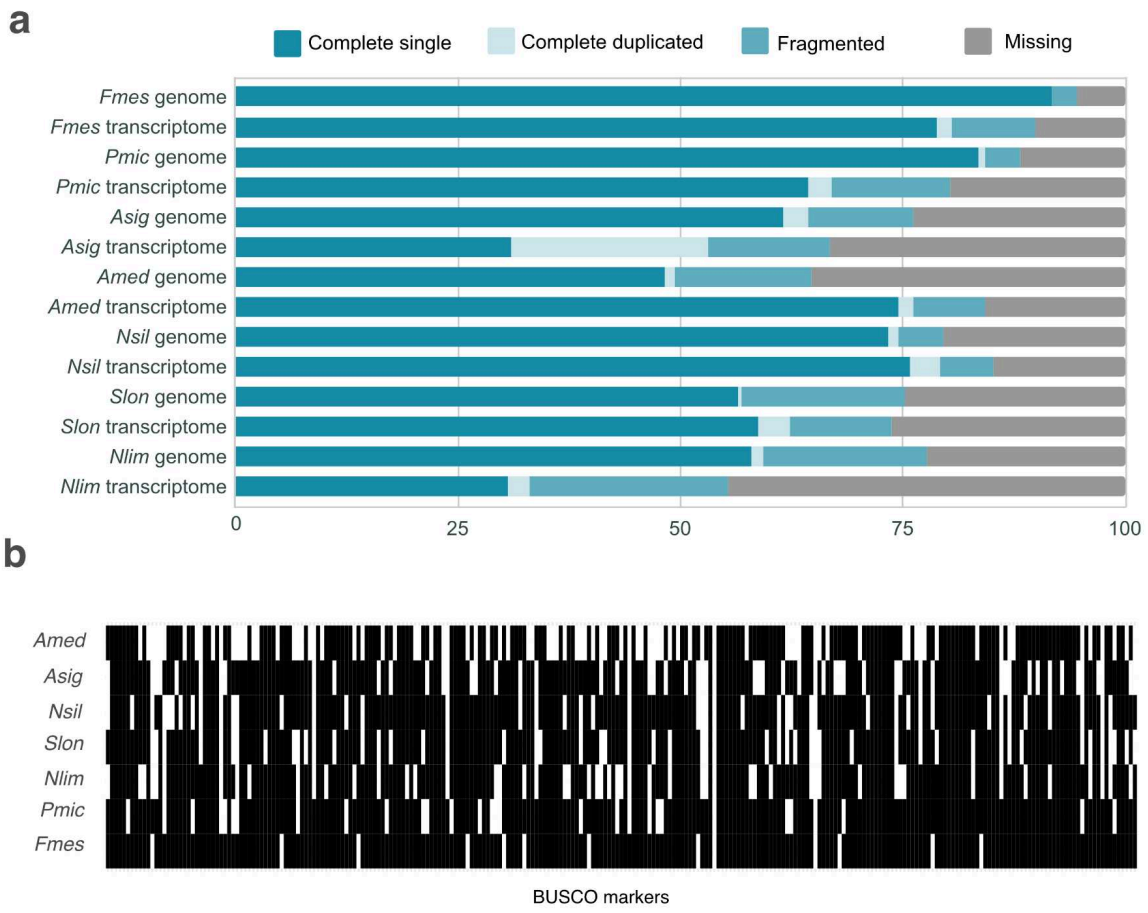

Figure S1. eukaryota_odb10 BUSCO markers tracked in the ancyromonad datasets. a) Percentage of BUSCO makers in the genomic and transcriptomic datasets of seven species of ancyromonads. b) Patterns of presence and absence of individual BUSCO markers across the genomic datasets.

Table S4. Mapping statistics of the RNA-seq clean reads to the ancyromonad genomic sequences.

Species ID	% of reads uniquely mapped	% of reads mapped to multiple loci	Total % of mapped reads	Number of splices:	Number of splices: GT/AG	Number of splices: GC/AG	Number of splices: AT/AC	Number of splices: Non canonical
Fmes	94.6	3.51	98.11	36,1289	35,7076	2416	0	1797
Pmic	85.3	10.3	95.6	133,263	12,9694	2308	17	3111
Asig *	70.31	23.7	94.01	401,439	39,1789	2981	59	6610
Amed	813.7	8.22	89.92	189,973	17,1542	8911	169	14554
Nsil	86.43	6.12	92.55	1,245,920	1,230,722	13,860	109	13622
Slon	88.21	6.45	94.66	318,034	311,532	3606	30	8260
Nlim	57.73	39.39	97.12	187,526	185,717	313	4	1492

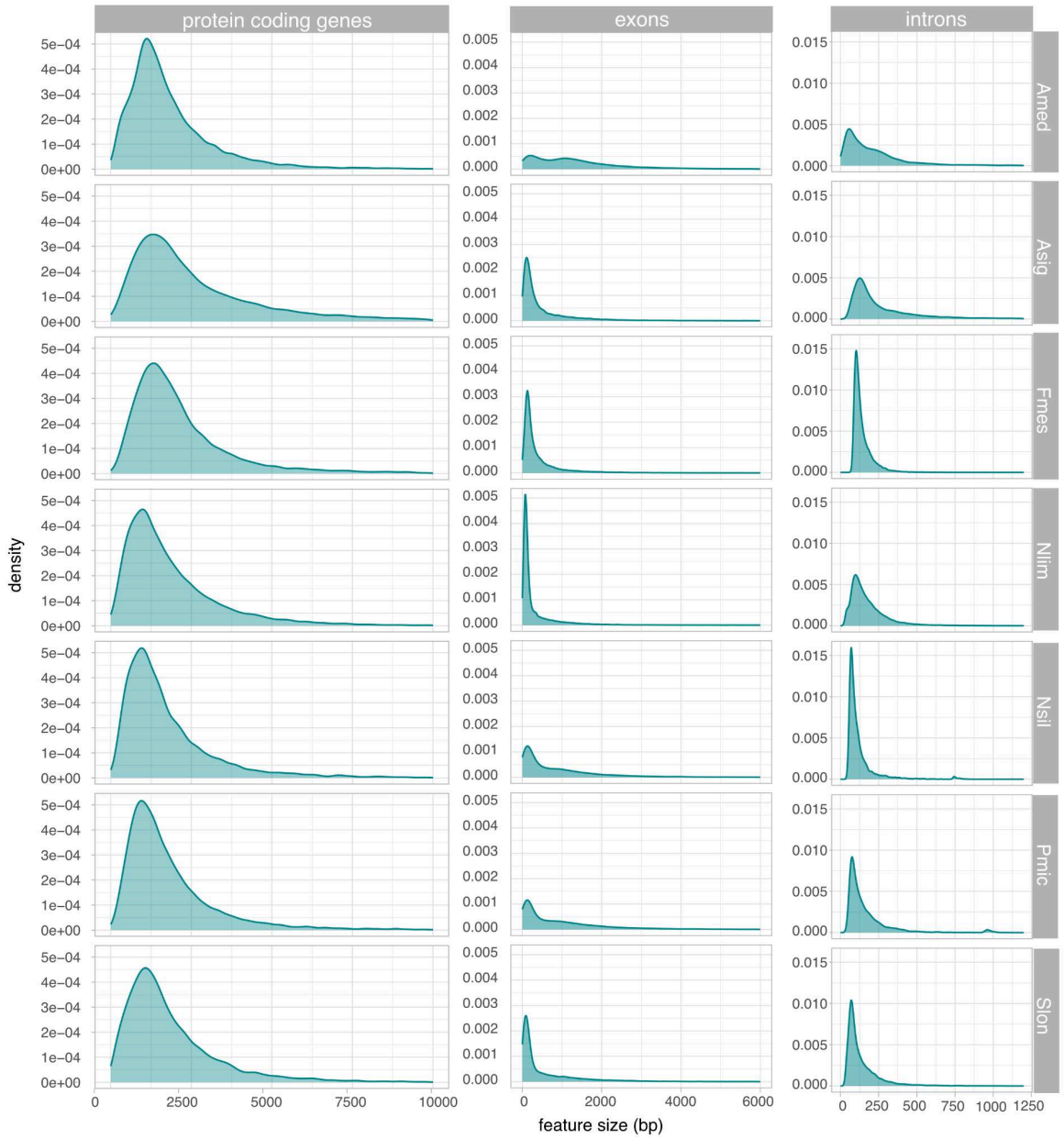

Figure S2. Distribution of gene, intron and exon sizes

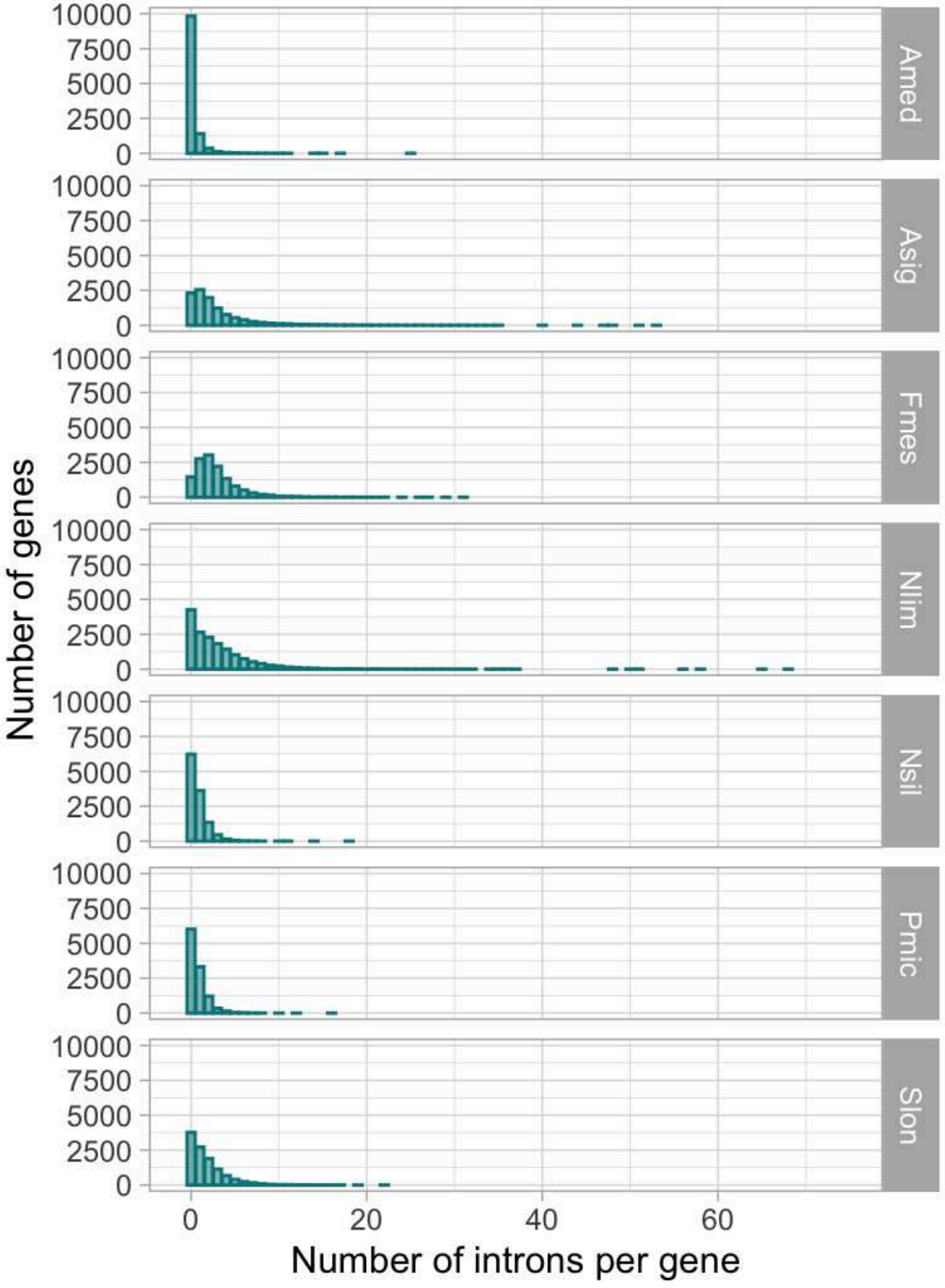

Figure S3. Intron density per gene.

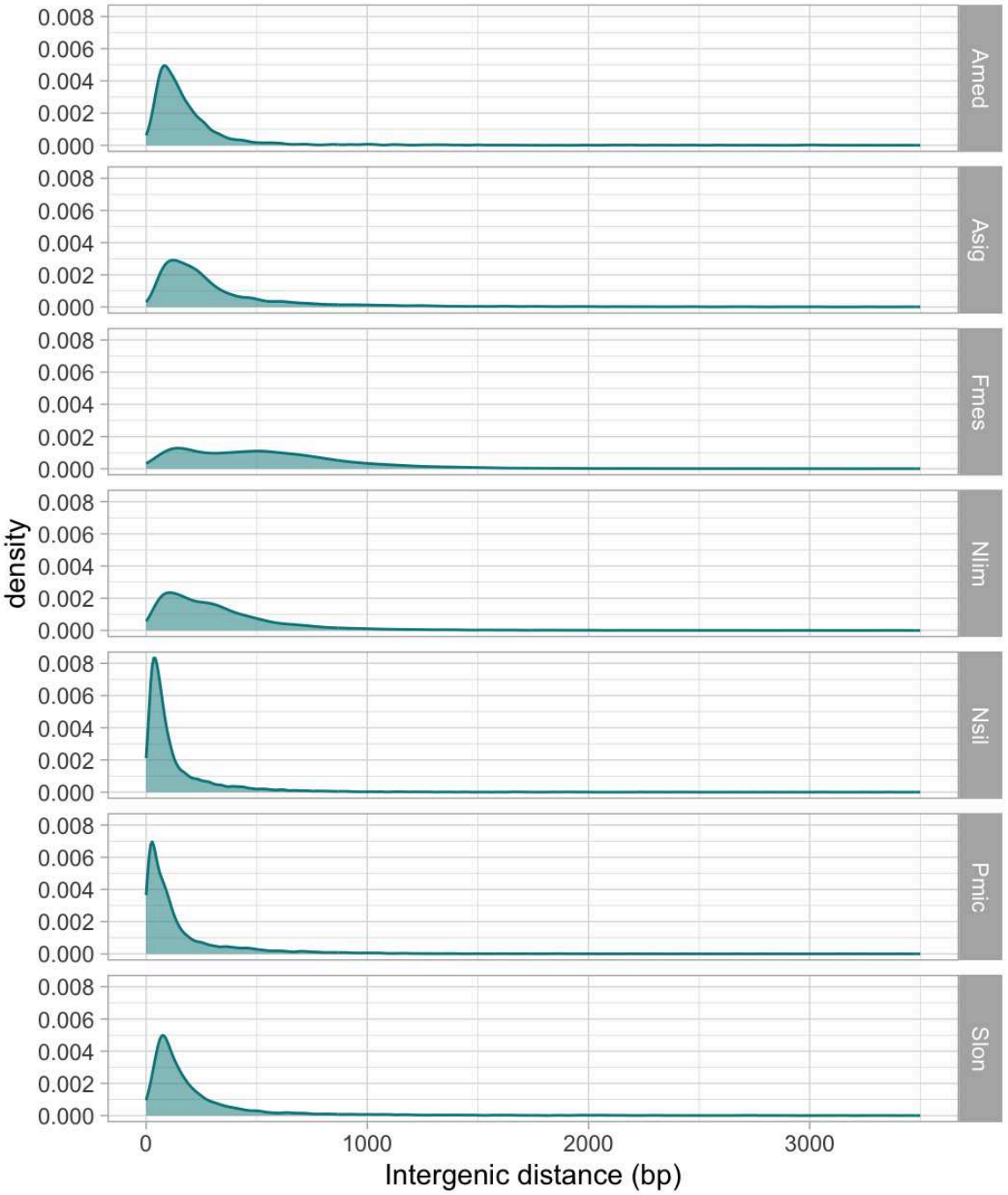

Figure S4. Distribution of intergenic distance.

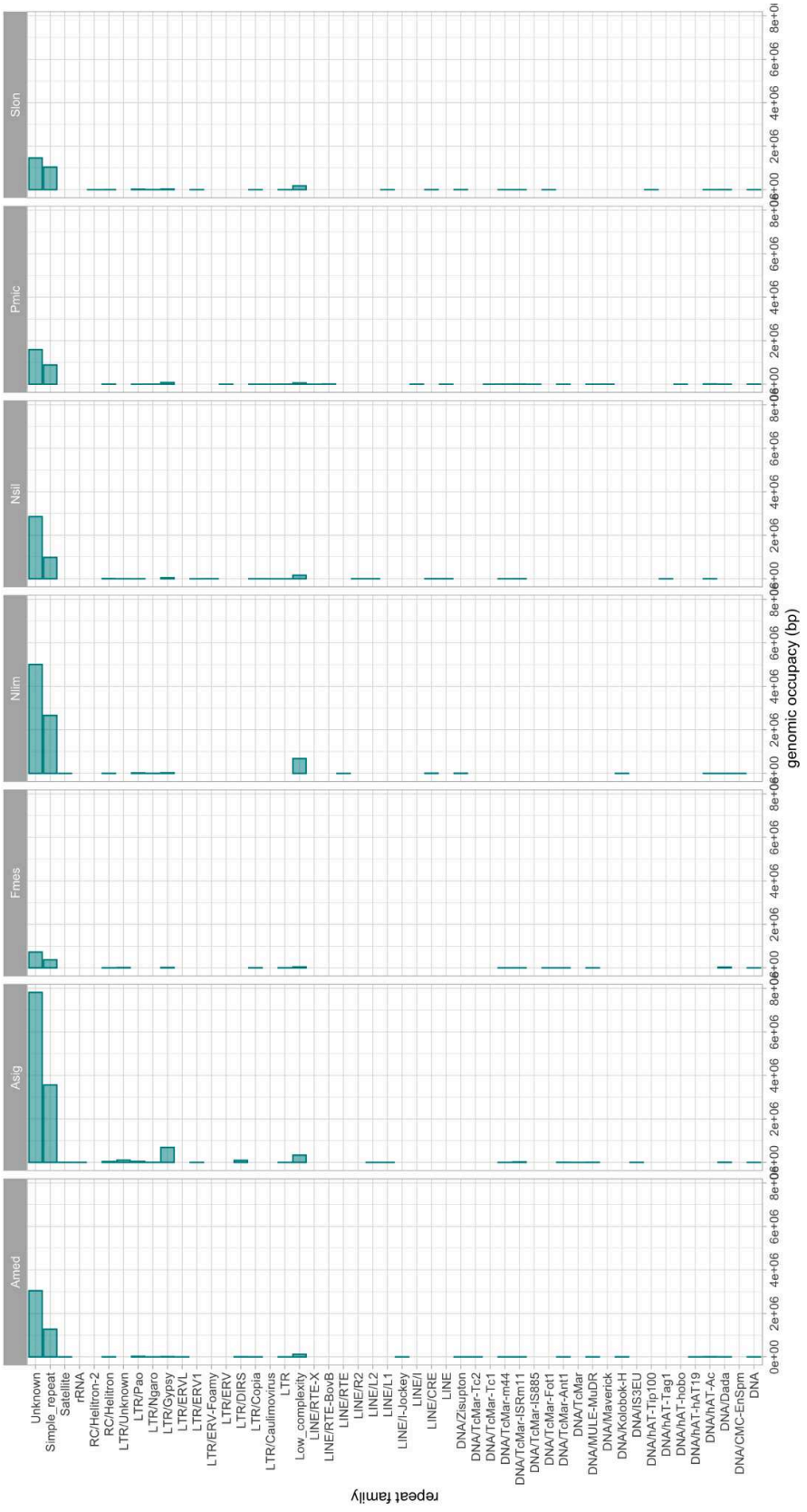

Figure S5. Comparison of the repeated element families identified in the genomic sequences.

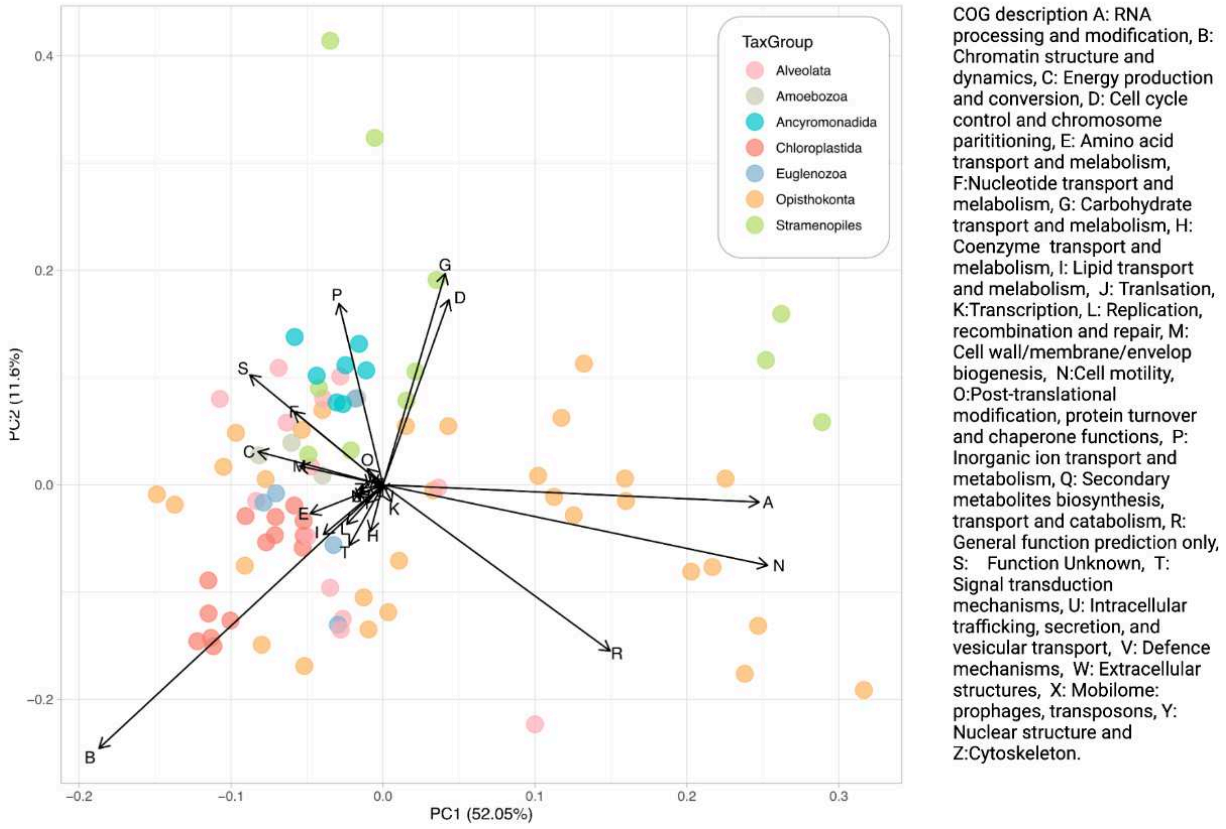

Figure S6. a) Proportion of COG categories in the ancyromonad predicted using egg-NOG mapper v2. b) Principal Coordinates Analysis of the distribution of COG categories in ancyromonads compared with a selection of representative eukaryotic proteomes from the EukProt v3 database.

a

	Fmes	*Pmic*	*Asig*	*Amed*	*Nsil*	*Slon*	*Nlim*
Fmes	6159	4536	3808	3452	4187	3677	3757
Pmic	4536	5480	3406	3118	3765	3256	3351
Asig	3808	3406	5967	4254	3937	3446	3524
Amed	3452	3118	4254	5436	3586	3252	3257
Nsil	4187	3765	3937	3586	6501	4482	4508
Slon	3677	3256	3446	3252	4482	5590	4021
Nlim	3757	3351	3524	3257	4508	4021	6163

b

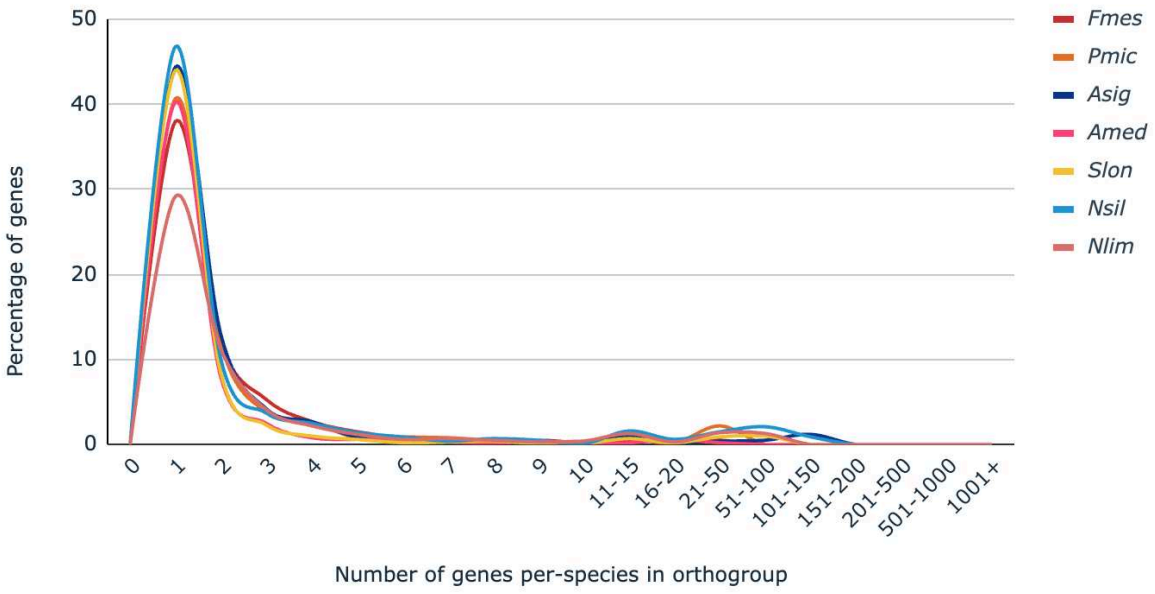

Figure S7. a) Gene families shared across ancyromonadida species. b) Distribution of gene family sizes.

Figure S8. Constraint tree used in the phylogenomic reconstruction.

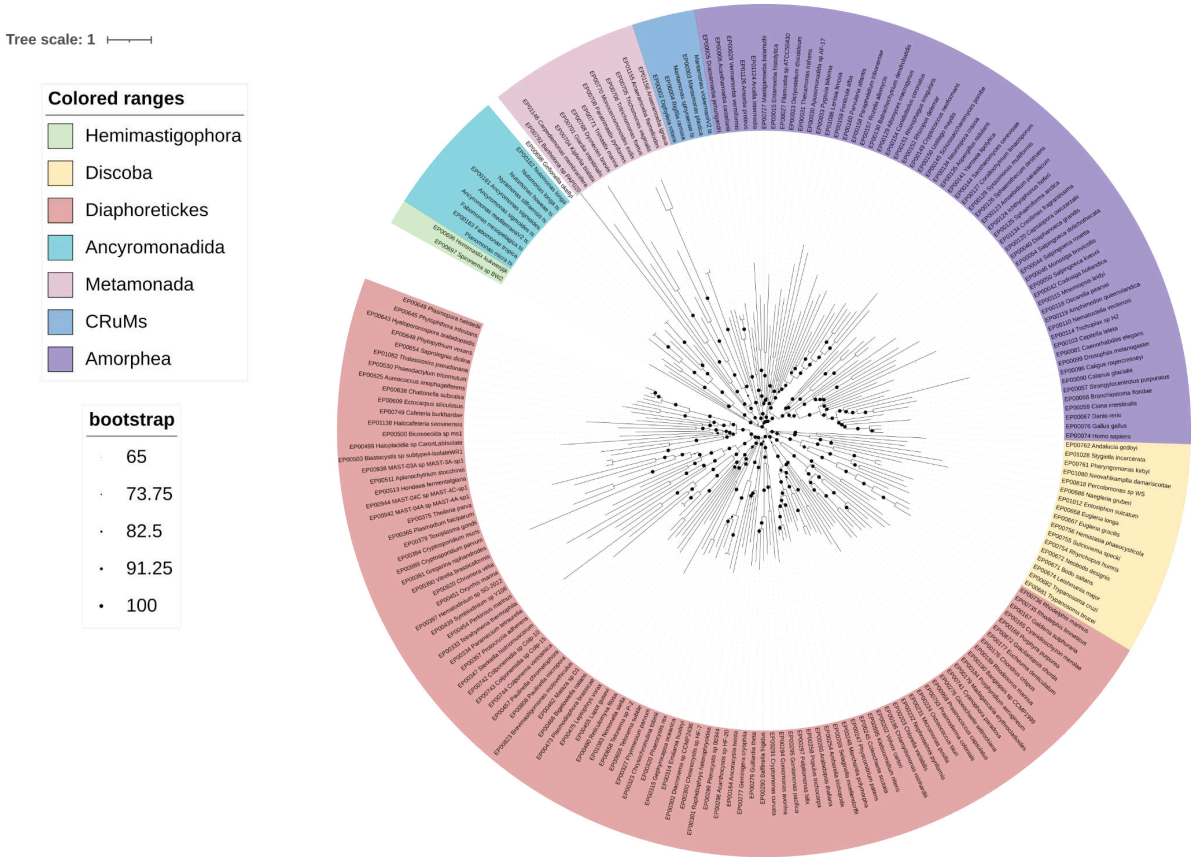

Figure S9. Phylogenomic analysis. Maximum Likelihood phylogeny based on 766 conserved proteins from OrthoFinder. The tree was obtained using 62,088 amino acid positions with the LG+C60+G model and 1000 bootstrap replicates. Values > 65% are indicated by black dots. The tree was rooted between Diphoda (Discoba+Diaphoretickes) and everything else.

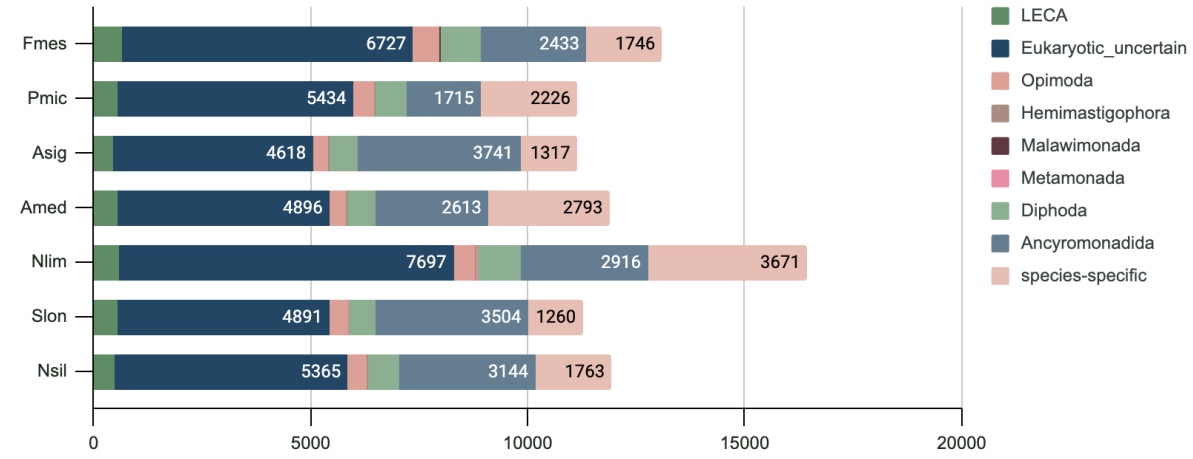

Figure S10. Origin of ancyromonad genomic repertoires (In proportions) according to the maximum origination value inferred by ALE reconciliation.

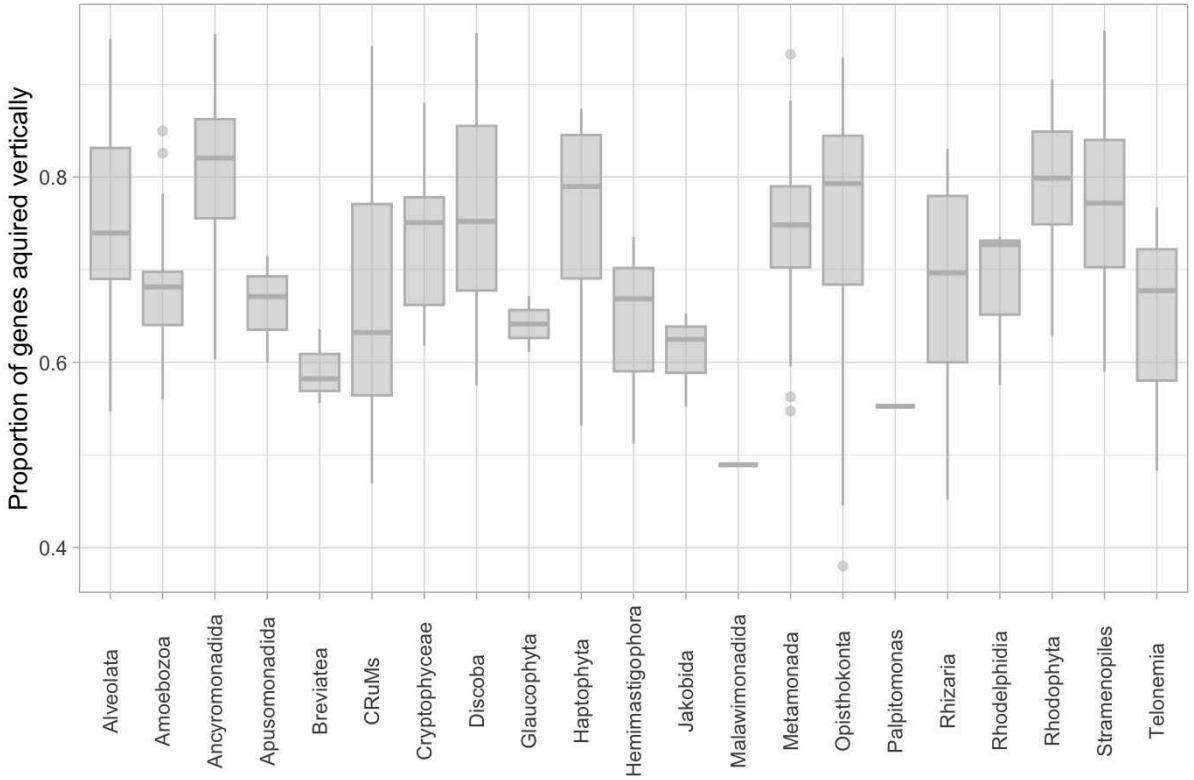

Figure S11. Verticality across all the nodes belonging to different clades of our species phylogeny.

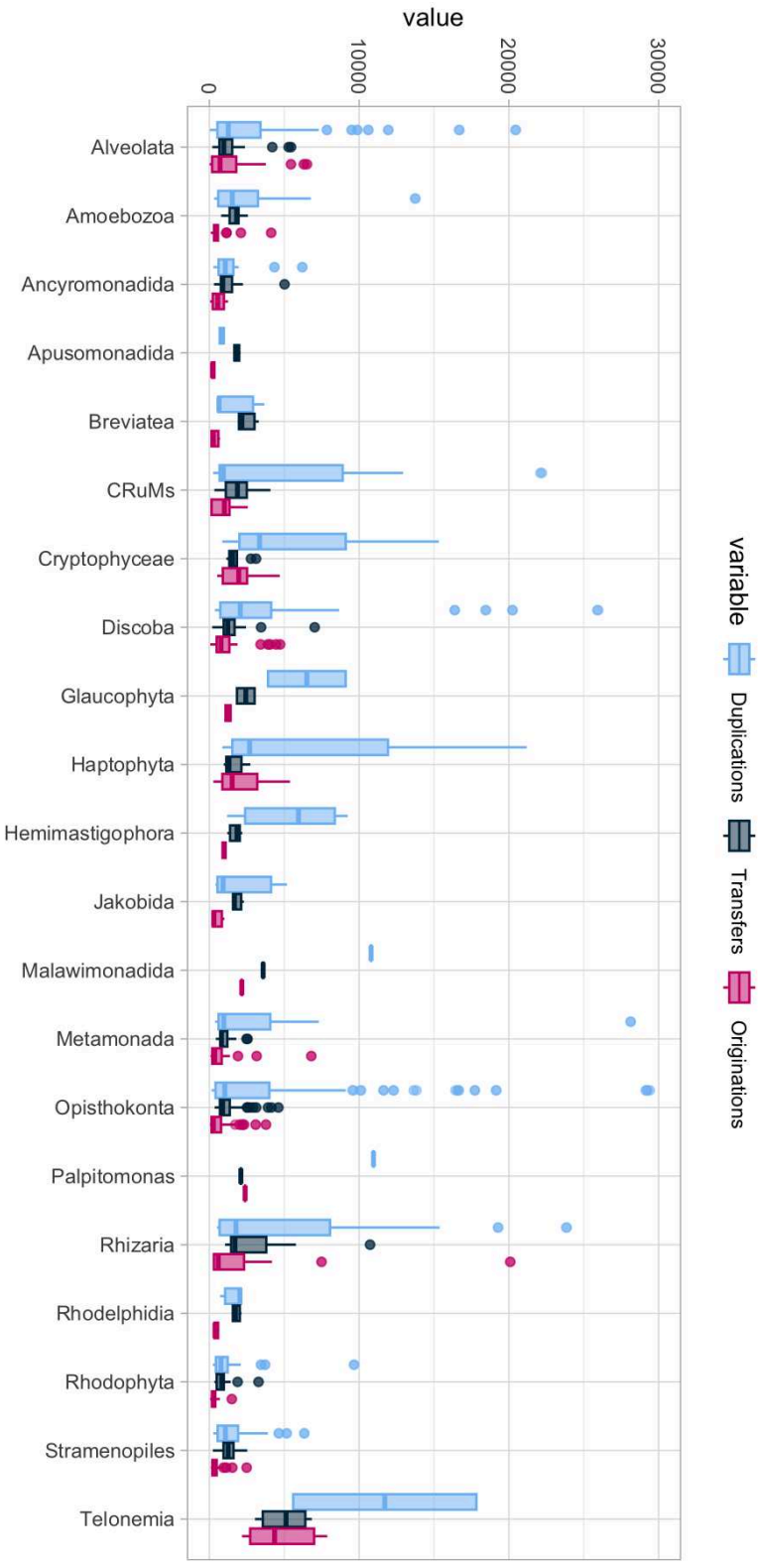

Figure S12. Mechanisms of gene gain across all the nodes belonging to different clades of our species phylogeny.

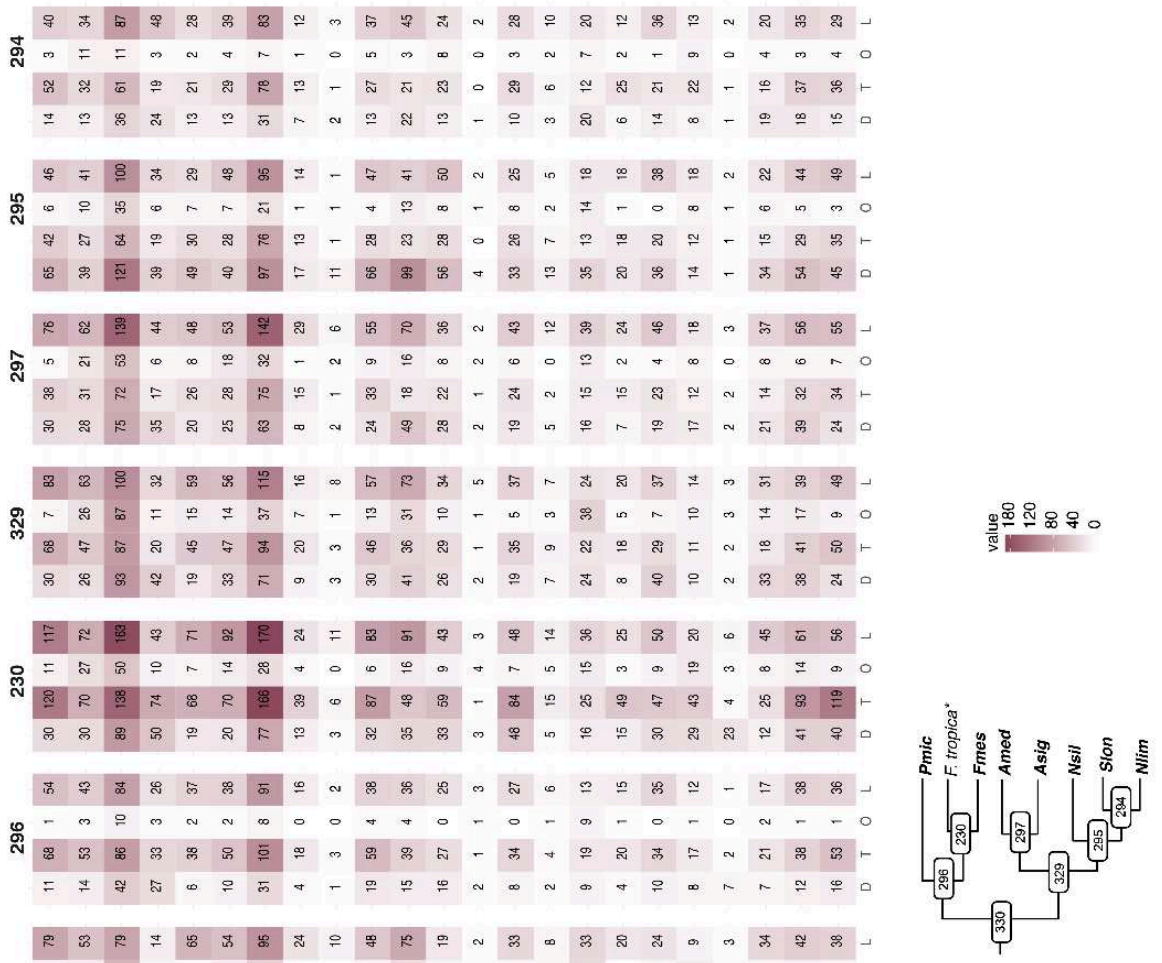

Figure S13. Evolutionary changes shaping the proteomes of ancyromonad ancestors inferred by ALE (D: duplications, T: transfers, O: originations, L: losses.). Gene families were splitted by the Cluster of Orthologous Groups (COG) category according to their classification with eggNOG-mapper. **F. tropica* only has transcriptomic data, numbers for this branch should be interpreted cautiously.

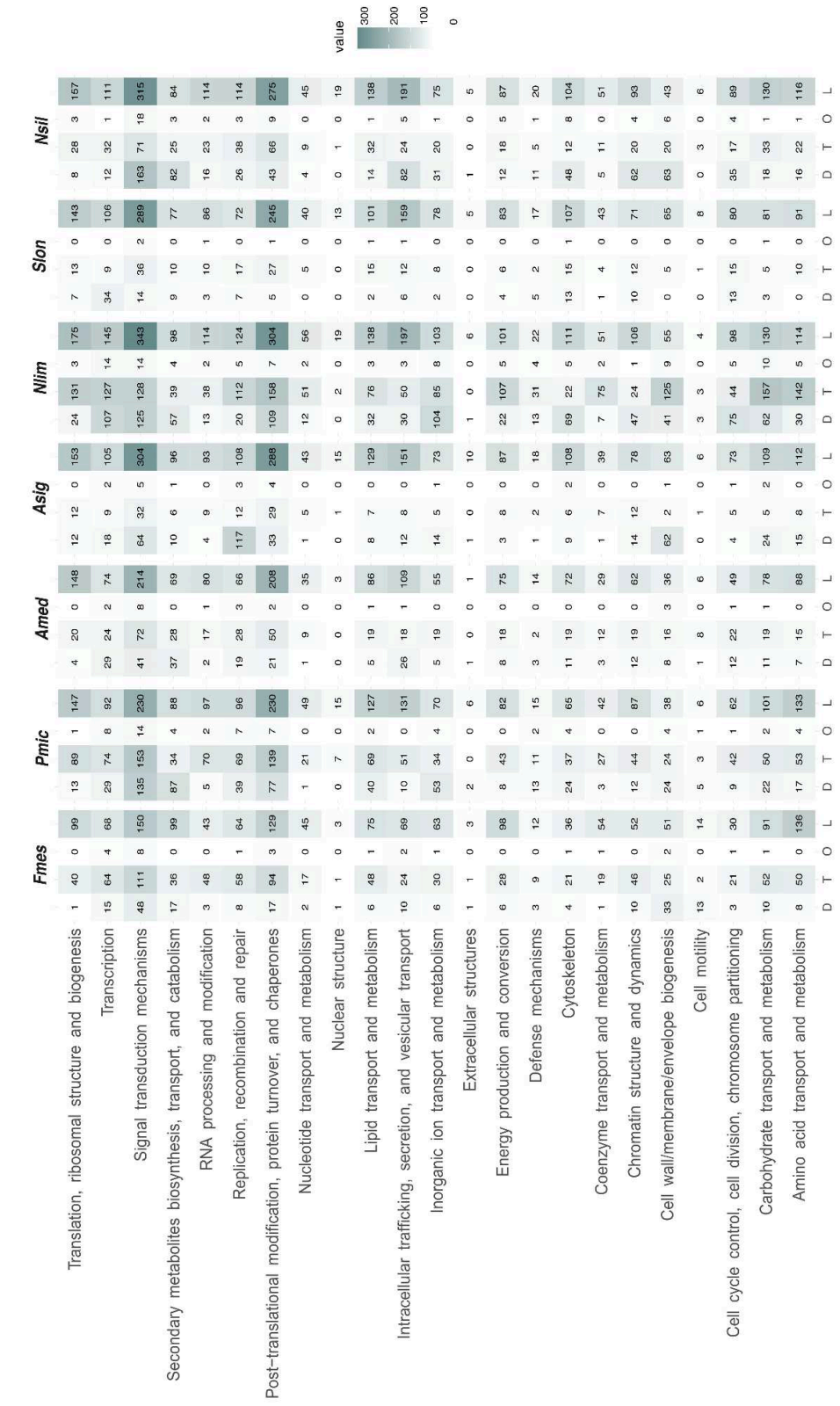

Figure S14. Evolutionary changes shaping the proteomes of modern ancyromonads inferred by ALE (D: duplications, T: transfers, O: originations, L: losses.). Gene families were splitted by the Cluster of Orthologous Groups (COG) category according to their classification with eggNOG-mapper.

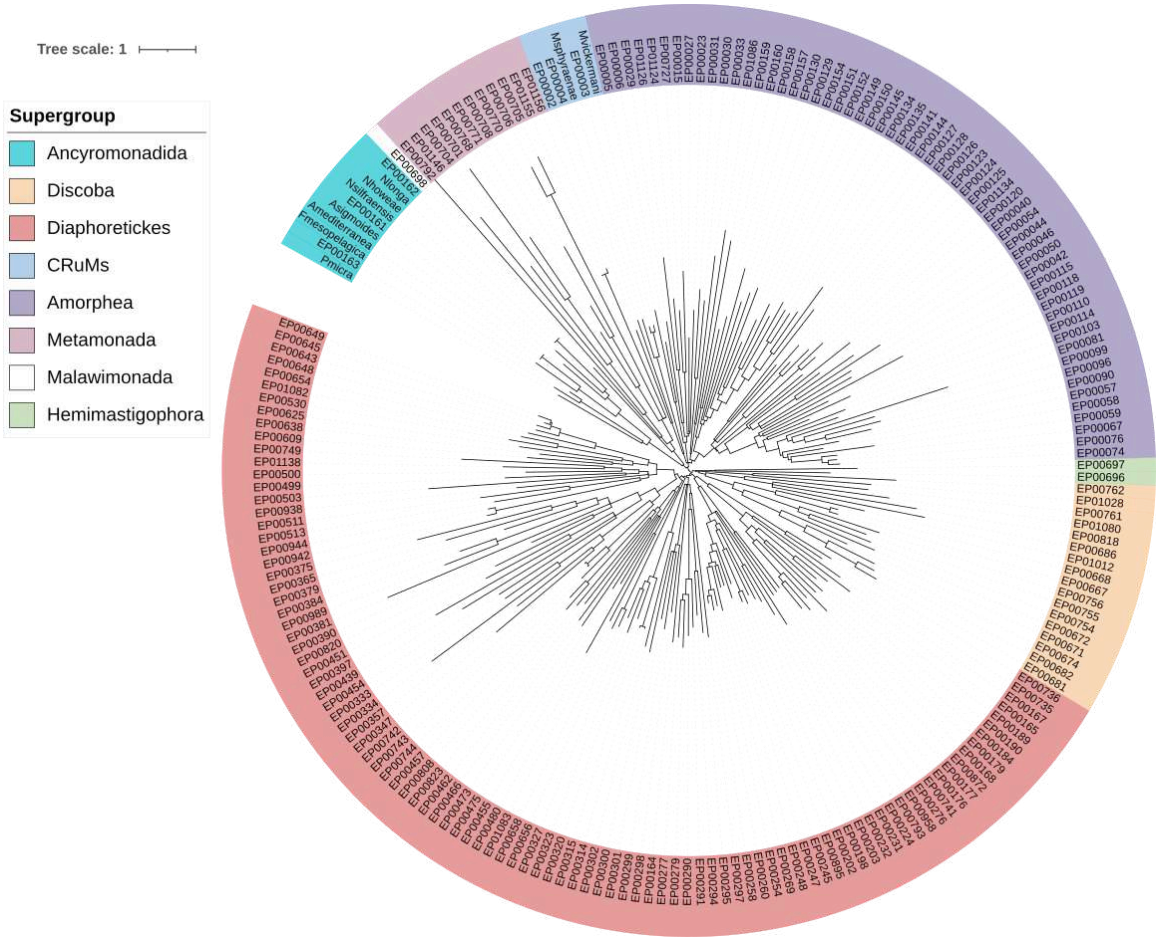

Figure S15. Alternative tree topology tested with ALE.

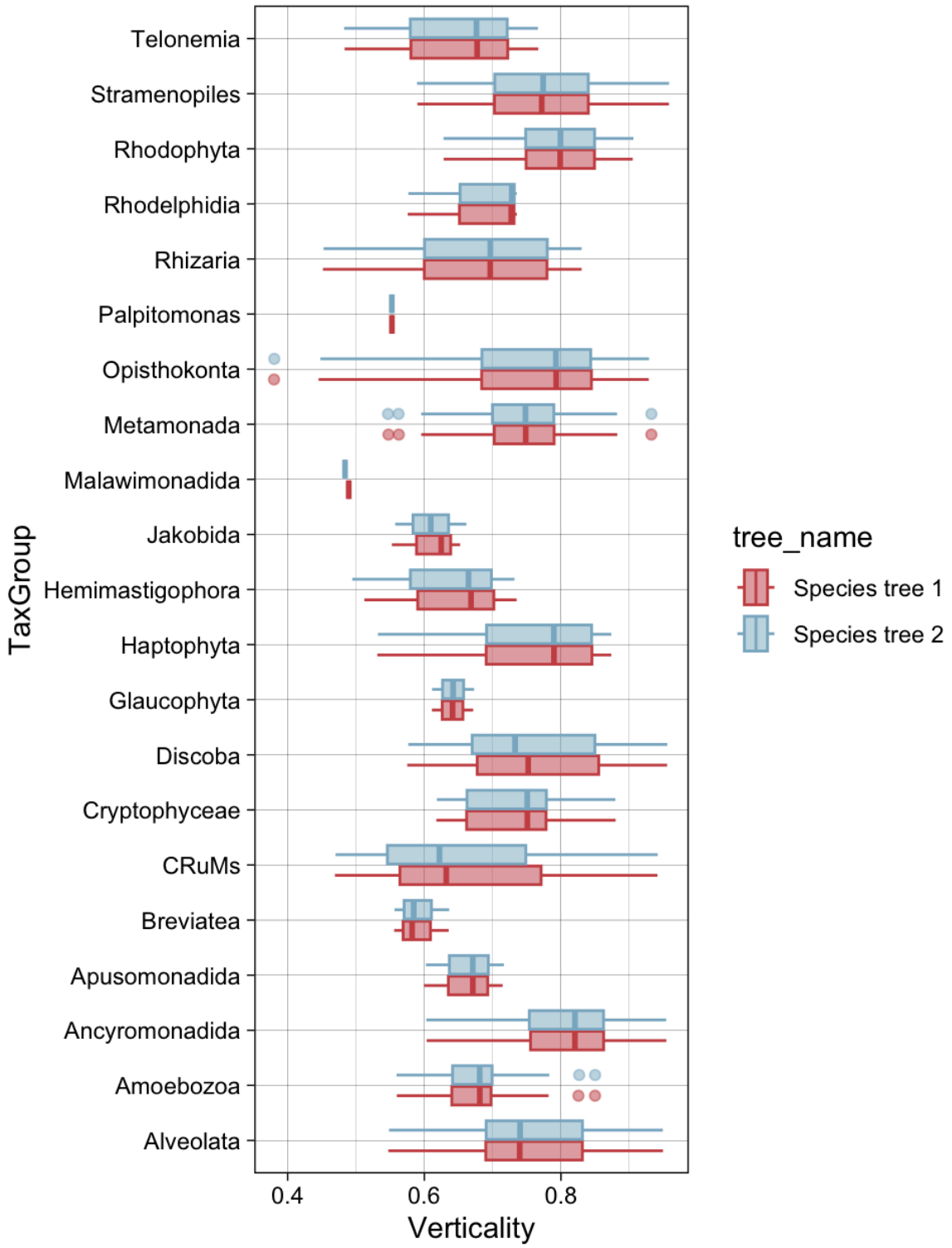

Figure S16. Verticality (singletons/singletons+transfers) estimated using the two alternative species tree topologies.

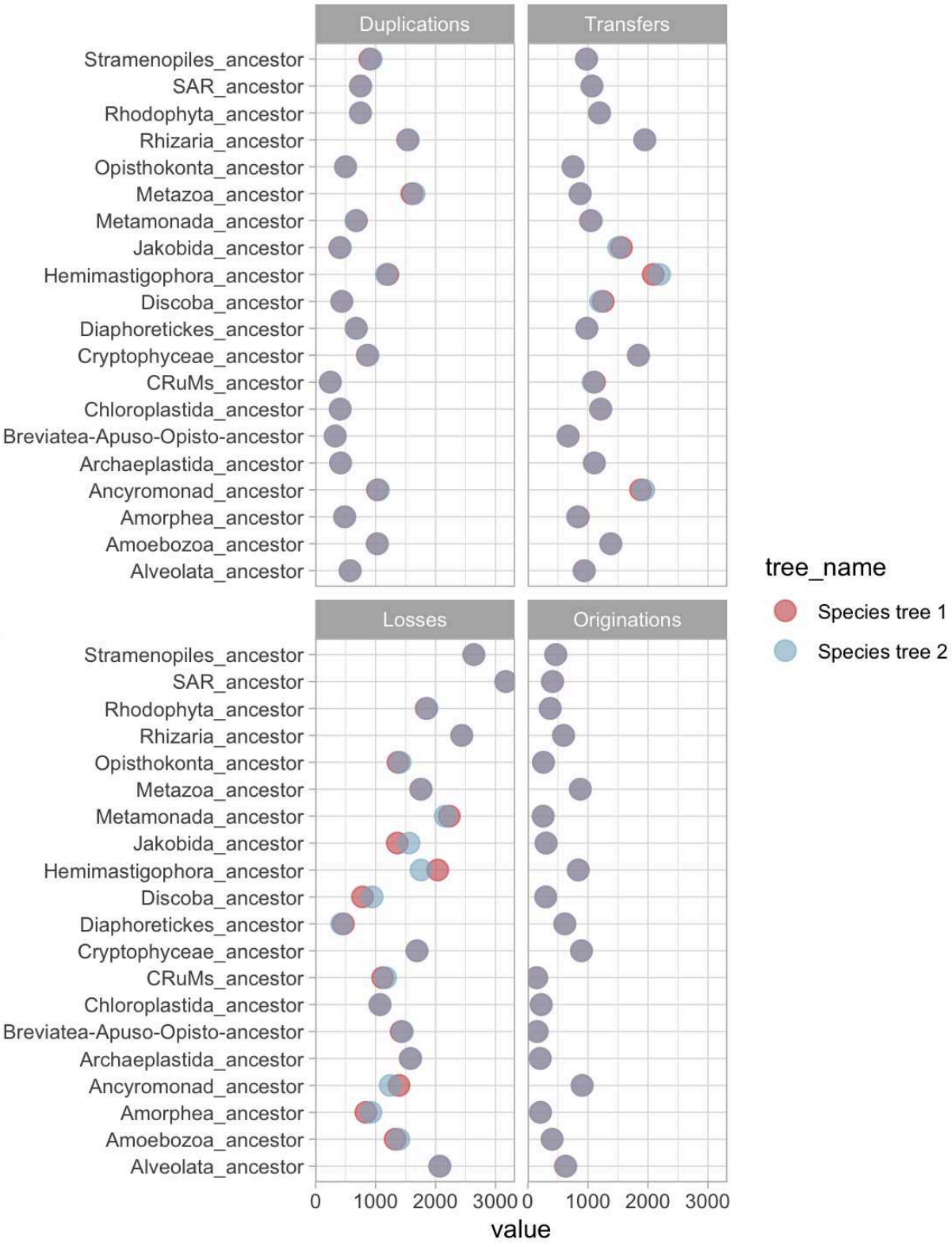

Figure S17. Evolutionary events inferred at key ancestral nodes using the different species tree topologies

Additional references

Alonge, Michael, Sebastian Soyk, Srividya Ramakrishnan, Xingang Wang, Sara Goodwin, Fritz J. Sedlazeck, Zachary B. Lippman, and Michael C. Schatz. 2019. “RaGOO: Fast and Accurate Reference-Guided Scaffolding of Draft Genomes.” Genome Biology 20 (1): 224.

Freire, Borja, Susana Ladra, and Jose R. Parama. 2021. “Memory-Efficient Assembly Using Flye.” IEEE/ACM Transactions on Computational Biology and Bioinformatics / IEEE, ACM PP (September). https://doi.org/10.1109/TCBB.2021.3108843.

Prjibelski, Andrey, Dmitry Antipov, Dmitry Meleshko, Alla Lapidus, and Anton Korobeynikov. 2020. “Using SPAdes De Novo Assembler.” Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.] 70 (1): e102.

Seppey, Mathieu, Mosè Manni, and Evgeny M. Zdobnov. 2019. “BUSCO: Assessing Genome Assembly and Annotation Completeness.” Methods in Molecular Biology 1962: 227–45.