View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Explore Bristol Research

This electronic thesis or dissertation has been downloaded from Explore Bristol Research, http://research-information.bristol.ac.uk

Author:
Scambler, Ross D

Title:
Exploring the evolutionary relationships amongst eukaryote groups using comparative genomics, with a particular focus on the excavate taxa

General rights
Access to the thesis is subject to the Creative Commons Attribution - NonCommercial-No Derivatives 4.0 International Public License. A copy of this may be found at https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode This license sets out your rights and the restrictions that apply to your access to the thesis so it is important you read this before proceeding.

Take down policy
Some pages of this thesis may have been removed for copyright restrictions prior to having it been deposited in Explore Bristol Research. However, if you have discovered material within the thesis that you consider to be unlawful e.g. breaches of copyright (either yours or that of a third party) or any other law, including but not limited to those relating to patent, trademark, confidentiality, data protection, obscenity, defamation, libel, then please contact collections-metadata@bristol.ac.uk and include the following information in your message:

• Your contact details
• Bibliographic details for the item, including a URL
• An outline nature of the complaint

Your claim will be investigated and, where appropriate, the item in question will be removed from public view as soon as possible.

Exploring the evolutionary relationships amongst eukaryote groups using comparative genomics, with a particular focus on the excavate taxa

Ross Daniel Scambler

Supervisor: Dr. Tom A. Williams

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Master of Science (by research) in the Faculty of Life Sciences, November 2020.

Word count: 11747

Abstract

Large-scale genomic data have provided new insights into the evolutionary relationships of various eukaryotes. Of particular interest were the excavates, a group of morphologically similar protists whose placement within the eukaryote Tree of Life (eToL) has long been problematic for researchers. The protein sequences of different taxa, representing all major lineages within Eukaryota, were here compared to identify the number of orthologous sequences that are shared amongst pairs of lineages. A high number of proteins shared uniquely between two eukaryote groups was proposed as evidence that the gene families encoding these proteins represented synapomorphies. This approach is an alternative to conventional phylogenetic analyses which do not always provide consistent results when inferring deep relationships amongst eukaryotes.

Analysis of the three excavate lineages Metamonada, Discoba and Malawimonadidae did not return a significant number of uniquely shared orthogroups in any pairwise comparison, therefore lending no support to the idea of a monophyletic Excavata. Other groups shared a considerably greater number of orthogroups, however. The orphan lineage Telonemia, comprising the sole genus *Telonema*, was inferred to have a specific relationship with another recently discovered orphan species, *Ancoracysta twista*. Similarly, Ancyromonadida was found to be related to members of the CRuMs (Collodictyonidae, Rigifilida, and *Mantamonas*) lineage, specifically *Diphylleia rotans* and *Rigifila ramosa*. Phylogenetic constraint analyses were performed to test these relationships. Neither the topology constraining Telonemia + *A. twista* nor that constraining Ancyromonadida + CRuMs could be rejected by the approximately unbiased (AU) test at a significance of 5%.

Based on these results, and the presence of a collection of 'excavate-like' morphological characters found with punctate distribution throughout the eToL, it is suggested here that the excavates may be a paraphyletic lineage and that the last eukaryotic common ancestor (LECA) may have been an excavate-like organism.

Dedication

To Mum and Dad

Acknowledgements

I would like to thank my supervisor, Dr. Tom Williams, for his valuable insight during our discussions of the research, as well as for his Python programming advice, for providing feedback on the dissertation, and for his good general advice over the past year. I would also like to thank Dr. Celine Petitjean, who performed the clustering analysis of sequences, providing me with the resulting orthogroup data used in this study, and has also provided much-appreciated advice over the course of my research.

Author's Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: Ross Scambler DATE: 19/11/2020

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

0.1 List of Abbreviations

AU = approximately unbiased

BLAST = Basic Local Alignment Search Tool

CRuMs = Collodictyonidae, Rigifilida and *Mantamonas*

eggNOG = evolutionary genealogy of genes: Non-supervised Orthologous Groups.

eToL = eukaryote Tree of Life

LECA = Last Eukaryotic Common Ancestor

LGT = lateral gene transfer

LBA = long-branch attraction

MRO = mitochondrion-related organelle

OG = orthogroup

SAR = Stramenopiles, Alveolates and Rhizaria

TSAR = Telonemia, Stramenopiles, Alveolates and Rhizaria.

Morphology

B1 = basal body 1

B2 = basal body 2

R1, R2, R3, R4 = microtubular roots.

Chapter 1

Introduction

1.1 Background

In recent years there has been a massive increase in the amount of genomic data made available in public repositories. New sequencing techniques, such as single-cell transcriptomics as well as environmental metagenomics projects, are providing huge amounts of information for an ever-broader range of taxa (Keeling et al., 2014; Escobar-Zepeda et al., 2015). This has enabled comparisons to be made amongst species spanning the tree of life, including those that have hitherto been largely ignored, or were until recently completely unknown. By carefully analysing the abundance of new sequence data, researchers are working to resolve some of the key unanswered questions in evolutionary biology.

Despite advances in our understanding of eukaryote evolution, many of the deep relationships amongst major groups remain unresolved. Understanding the shape of the tree of life is crucial – it provides the framework for deducing patterns of change and adaptation seen in different taxa, and has use as a predictive tool when comparing related species (Keeling et al., 2005).

The general appearance of the eukaryote Tree of Life (eToL) has shifted over the past few decades to one increasingly resembling a 'supergroups' model, wherein all eukaryote diversity is divided into a handful of major lineages (Baldauf, 2003; Parfrey et al., 2006; Burki et al., 2019). One possible configuration of the eToL is shown here as an example (Fig. 1). As both taxon and gene sampling have improved, many of these groups have been fleshed out, becoming better defined, in some cases coalescing with one another or alternatively splitting and shuffling to other locations on the tree. For instance, the chromalveolates were originally thought to have arisen from a common ancestor that underwent a single endosymbiosis

event with a red algal symbiont (Cavalier-Smith, 1999). This grouping has since been disrupted as its constituent members, the stramenopiles (formerly heterokonts) and alveolates were shown to form a robust clade with rhizarians, forming the supergroup known as SAR (Burki et al., 2007). Meanwhile, Cryptista (formerly cryptomonads) and Haptista (comprising the haptophytes and centrohelids) have emerged as distinct clades in their own right (Yabuki et al., 2014; Cavalier-Smith et al., 2015) with affinities to Archaeplastida and SAR, respectively (Burki et al., 2016). Together, Cryptista, Haptista, Archaeplastida and SAR have all been assigned to the megagroup Diaphoretickes (Adl et al., 2012), along with some orphan lineages. One of these groups, the ecologically abundant but seemingly less diverse telonemids (sole genus *Telonema*) (Klaveness et al., 2005) frequently branches with SAR in molecular analyses (Burki et al., 2012), leading to an extension of the original moniker: TSAR (Strassert et al., 2019).

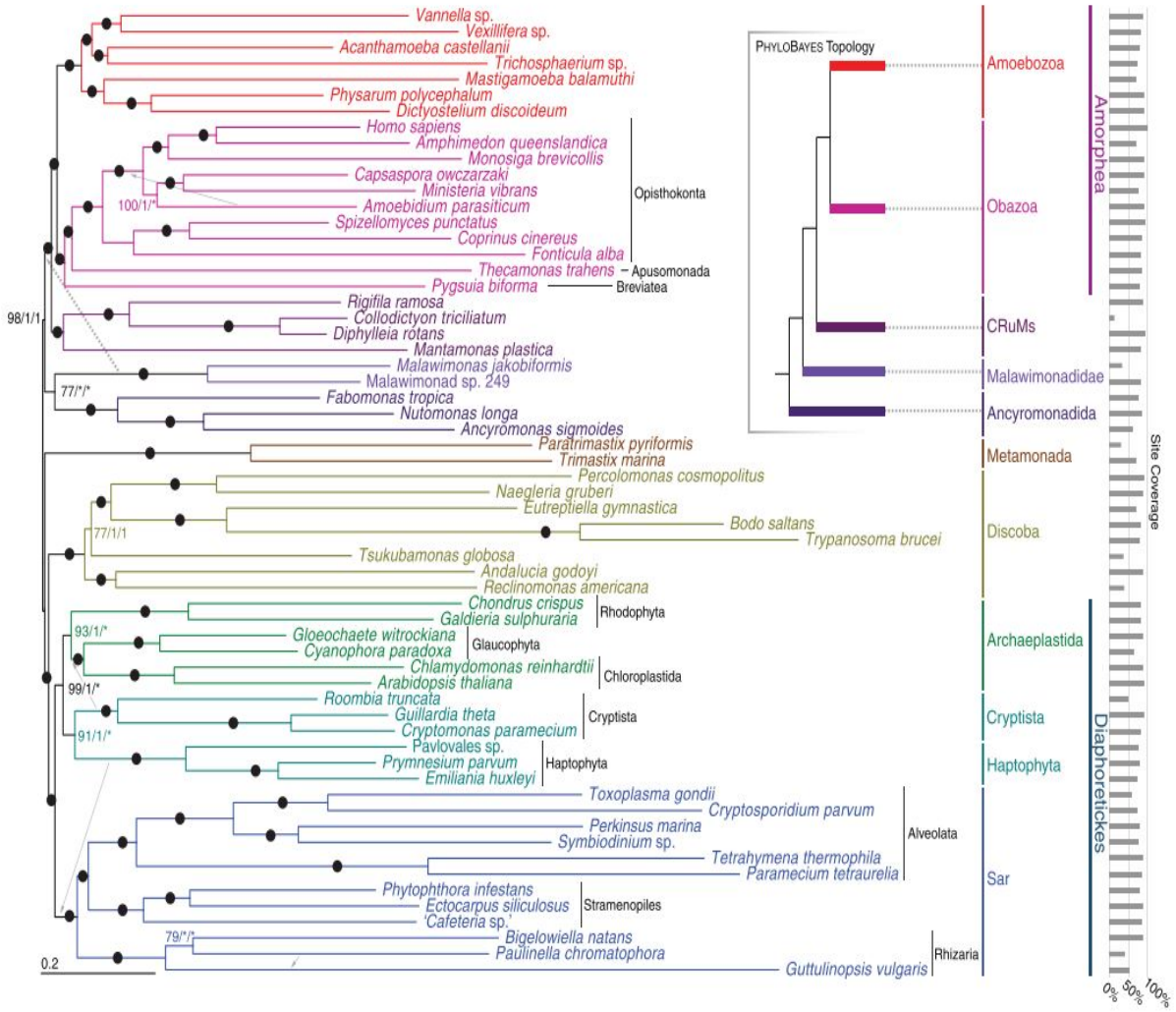

Figure 1 – **Unrooted eukaryote phylogeny.** Maximum Likelihood tree of major eukaryote groups recovered under the LG+C60+F+Γ-PMSF model. Taken from Brown et al. (2018).

Elsewhere in the tree, a trio of orphan lineages have recently been recognised as monophyletic. The group comprising Collodictyonidae, Rigifilida and *Mantamonas* has been informally termed the CRuMS lineage (Brown et al., 2018). Earlier studies were inconsistent in the placement of these taxa, with analyses missing some of the representatives (Cavalier-Smith et al., 2014) or otherwise using only a very small number of genes to infer phylogenies (Yabuki et al., 2013b). Obazoa, comprised of Opisthokonta, Breviatea and Apusomonadida (Brown et al., 2013), and Amoebozoa, which together form an even larger assemblage known as Amorphea (Burki, 2014), have shown affinity to the CRuMs lineage in recent phylogenomic analyses (Brown et al., 2018; Lax et al., 2018). Another lineage that may be related to Amorphea are the ancyromonads (Paps et al., 2013), although this group has so far evaded consistent placement in the eToL. It may yet emerge that this taxon occupies its own deeply diverged position, as appears to be the case for the newly established Hemimastigophora (Lax et al., 2018). Alternatively, it may be related to the excavate taxa, as the marked similarities in cellular ultrastructure would suggest (Heiss et al., 2011). More rigorous analyses are needed to determine the ancyromonads' true evolutionary position. Given that it may be a lineage diverging close to the root of the tree, and that it has the potential to shed light on another controversial group, the excavates, a better understanding of this group's relationships may answer key questions in the field of eukaryote evolution.

1.2 The excavates

The excavate taxa are some of the best-studied of all protists, yet remain perhaps the most puzzling pieces in the eToL. Early research found morphological links between the jakobids and retortamonads (O'Kelly, 1993), which, it soon emerged, were a suite of characters common (for the most part) to a wider group also including the diplomonads, heteroloboseids, *Carpediemonas*, *Trimastix* and *Malawimonas* (O'Kelly and Nerad, 1999; O'Kelly et al., 1999; Simpson and Patterson, 1999, 2001). This group was subsequently proposed as a formal taxon, Excavata, by Cavalier-Smith (2002).

Three excavate lineages are currently recognised: Metamonada (Cavalier-Smith, 2003), Discoba (Hampl et al., 2009) and the orphan lineage Malawimonadidae (O'Kelly and Nerad, 1999). Most notable amongst their unifying morphological features are the ventral feeding groove and cytoskeletal features of the flagellar apparatus (Simpson, 2003). Authors put forward the idea that, because of these typical 'excavate' characteristics, the group may be monophyletic, or possibly paraphyletic, in what has been termed the excavate hypothesis (Simpson and Patter-

son, 1999; Simpson et al., 2000). The distinction between monophyly and paraphyly is significant, as the latter hypothesis suggests that the last eukaryotic common ancestor (LECA) may have been an excavate-like organism, and that over the course of evolution many eukaryote lineages have lost the characteristic excavate morphology. This idea was originally proposed by (O'Kelly, 1993), and more recently an analysis of cytoskeletal architecture has similarly suggested that the excavate morphology may represent an ancestral condition (Yubuki and Leander, 2013). This was reiterated by Keeling and Burki (2019), who suggest that the excavates may be paraphyletic and interpret the presence of a ventral feeding groove in a number of different protist lineages as evidence that the excavate condition may be ancestral.

1.2.1 Ecology

The excavates exhibit a variety of lifestyles: many are free-living heterotrophic flagellates, some of which occupy oxygen-poor environments (Bernard et al., 2000). Others are known to be medically significant parasites of humans and other animals, responsible for diseases including African Sleeping Sickness (*Trypanosoma*) (Barrett et al., 2003), trichomoniasis (*Trichomonas*) (Carlton et al., 2007) and giardiasis (*Giardia*) (Ortega and Adam, 1997). Tendency towards a parasitic lifestyle has primarily occurred in metamonads, leading some authors to suggest that parasitism was an ancestral trait in this lineage, arising after the appearance of the animal gut (Cavalier-Smith, 2003). Adaptation to the anoxic environments found within host species has involved the reductive evolution of mitochondria into the hydrogenosomes and other mitochondrion-related organelles (MROs) of parasitic metamonads (Roger et al., 2017). In the extreme case of *Monocercomonoides*, this has led to the complete loss of the organelle (Karnkowska et al., 2016).

A small number of euglenids (Discoba) are photosynthetic, a nutritional mode that is combined with osmotrophy in most plastid-bearing species (Vesteg et al., 2019). The existence of photosynthetic excavates has led to the suggestion that phototrophy may have been an ancestral trait for the group (assuming excavate monophyly) and was subsequently lost in the majority of species (Cavalier-Smith, 2003). However, there is now compelling evidence that a far more recent endosymbiosis event occurred between a euglenid ancestor and a prasinophyte green alga (Turmel et al., 2009).

1.2.2 Morphology in detail

Protists can usually be compared by examining the various cytoskeletal structures associated with their flagella. These and their associated basal bodies may have arisen just once over the course of eukaryote evolution, i.e. they are homologous amongst protists (Moestrup, 2000).

The ventral feeding groove is found in a number of protist groups, and this feature has been proposed as being homologous between the excavate taxa and at least some other groove-bearing organisms (Heiss et al., 2011), although whether or not the groove had a single origin amongst all eukaryotes remains unclear. In any case, it appears to have been lost in a number of excavates – most euglenozoans, parabasalids and oxymonads (Simpson, 2003). Aside from the ventral feeding groove, the excavate taxa can be characterised by the number and orientation of their basal bodies (the intracellular components of flagella that give rise to axonemes), and the various microtubular structures that are found supporting them. Most species have either two or four basal bodies; three is uncommon, although this appears to be the case for *Carpediemonas membranifera* (Simpson and Patterson, 1999), and in most diplomonads a total of eight are recognised (McInally and Dawson, 2016). In typical excavates, one of these is a posteriorly-directed basal body associated with a flagellum that beats in order to generate a current, thus directing food particles into the ventral feeding groove (Simpson, 2003). This basal body (known as B1) is most often associated on either side by a right and a left microtubular root, which each extend to support the feeding groove. In the case of the Preaxostyla (i.e. *Trimastix* and the oxymonads), these microtubular structures occur as a sheet, known as the preaxostyle, and this is inferred to be homologous to the roots found in typical excavates (Simpson et al., 2002a). Similarly, the euglenozoans, which in most species lack a ventral feeding groove, have nevertheless retained a structural link between the right microtubular root and the feeding apparatus (Simpson, 2003).

A second basal body, the anterior basal body (known as B2), is found in some excavates to associate with an anterior root, which itself is associated, to a greater or lesser degree, with an expanding dorsal fan of microtubules (Simpson and Patterson, 1999). In other species the microtubular root of B2 is instead found on its ventral side, or simply not present, for example in the case of most jakobids (O'Kelly and Nerad, 1999).

The right microtubular root of excavates is usually split into an outer root and an inner root. This split occurs close to the root's origin in some taxa, for example *Malawimonas* (O'Kelly and Nerad, 1999). In others, such as *Reclinomonas*, the split occurs much further from the origin (O'Kelly, 1997). The right root of some heteroloboseids also has the characteristic of being

C-shaped (viewed in cross-section) at its origin. This has been inferred as being homologous to the hook-band microtubular feature found in retortamonads (Bernard et al., 1997).

Another interesting comparison is that of the costae of trichomonad parabasalids and the C fibre seen in other excavates. Costae are striated roots found in association with the basal bodies (Viscogliosi and Brugerolle, 1994). The layered sheet substructure of costae appears to be highly similar to that seen in the C fibre (although costae are composed of many more layers) and as such it has been suggested that the two structures could be homologous (Simpson and Patterson, 2001)

In addition to the C fibre, excavates also possess a number of other fibres that support the cytoskeleton: the B fibre, first described as 'root B' (Patterson, 1990), originates on the right side of B1 and runs alongside the right microtubular root. Also connected to the right root is the I fibre (O'Kelly, 1997). These fibres are not present in all excavates. In diplomonads, for instance, I fibres appear to be absent, although I fibre-like structures have been reported in some species, e.g. *Brugerolleia algonquinensis* (Desser et al., 1993).

1.2.3 Molecular analyses

Though a suite of morphological characters unite the excavate taxa, the supporting evidence provided by molecular analyses has been somewhat mixed. Early molecular analyses recovered the diplomonads as one of the most deeply diverged eukaryote groups (Sogin et al., 1989; Leipe et al., 1993), which, being apparently amitochondriate organisms, was congruent with the idea that they had split before the acquisition of the first mitochondrion in eukaryotes (Cavalier-Smith, 1993). Phylogenies of alpha- and beta-tubulins found the diplomonads to be closely related to *Carpediemonas*, and those of small subunit ribosomal RNA (SSU rRNA) further recovered the retortamonads in this clade (Simpson et al., 2002b). The inclusion of these other excavate taxa led authors to question the legitimacy of the diplomonads' deep position within the eToL, since molecular analyses excluding this group found the retortamonads and *Carpediemonas* to be nested much further from the base. The possibility that the topologies of SSU rRNA phylogenies arose due to long-branch attraction (LBA) artifacts also fuelled uncertainty (Embley and Hirt, 1998).

Meanwhile, other phylogenetic studies bolstered support for intra-excavate relationships: Silberman et al. (2002) recovered a retortamonad + diplomonad clade; Dacks et al. (2001) recovered a *Trimastix* + *Pyrsonympha* (oxymonads) clade. The above studies highlighted the possibility that the amitochondriate taxa under examination may have lost their organelles secon-

darily, given the presence of certain membrane-bounded organelles in their inferred sister taxa (Roger, 1999), which have since been confirmed as hydrogenosomes or other mitochondrion-related organelles (MROs) (Hampl et al., 2008; Roger et al., 2017). Aside from shedding light on the evolution of mitochondria, these initial molecular studies showed little support for an overall Excavata clade. Phylogenies based on just one or two loci found the excavate taxa to be polyphyletic (Edgcomb et al., 2001; Simpson et al., 2002b), but with the addition of more taxa analysed over more loci, a more consistent picture of excavate evolution began to emerge.

A clade resembling the discobid excavates (then named Discicristata) was first recovered by Baldauf et al. (2000) using a combined protein phylogeny of elongation factor 1 (EF-1), actin, alpha- and beta- tubulin. Notably, this clade was not recovered with strong support in most molecular analyses of individual loci. A similar clade, with the addition of jakobids, was recovered in a six-gene phylogeny with a greater focus on the excavate taxa, along with a clade resembling Metamonada (when alpha-tubulin was excluded from the analysis), albeit weakly supported (Simpson et al., 2006a). Metamonada was recovered by Hampl et al. (2005), who also opted to exclude tubulins from part of the analysis due to conflicting phylogenetic signal.

The increase from a handful of loci to dozens or hundreds of proteins in molecular analyses further cemented support for Discoba and Metamonada, yet the malawimonads' position remained unstable (Rodríguez-Ezpeleta et al., 2007; Burki et al., 2007; Katz and Grant, 2015). The excavate taxa, including Malawimonadidae, were recovered as a monophyletic group by Hampl et al. (2009) after certain long-branching gene sequences were removed from the analysis, and then only with moderate bootstrap support. Parfrey et al. (2010) also found the excavates to be monophyletic in a 451-taxon analysis using 16 loci, and Heiss et al. (2018) recovered a Metamonada + Malawimonadidae clade after the removal of fast-evolving sites and long-branch taxa from their initial dataset. However, these studies have been the exceptions rather than the rule. The monophyletic concept of Excavata has proved elusive in the majority of phylogenomic analyses to date; more often the group has been recovered as polyphyletic (Burki et al., 2009; Zhao et al., 2012; Brown et al., 2013; Cavalier-Smith et al., 2014; Yabuki et al., 2014; Brown et al., 2018; Lax et al., 2018).

To complicate matters, metamonads – especially parabasalids and diplomonads – are often the most problematic taxa in phylogenetic analyses, and due to their fast-evolving nature are known to be some of the longest-branching of all eukaryotes (Simpson et al., 2006a). As a result these taxa are sometimes omitted from pan-eukaryotic phylogenies (Burki et al., 2016; Strassert et al., 2019) or removed following initial analysis because they are fast-evolving (Hampl et al.,

2009).

1.3 Rooting the EtoL

One of the enduring challenges faced by researchers has been to pinpoint the root of the eukaryotic tree. The vast evolutionary timescales separating extant Eukaryotes from their prokaryotic ancestors, i.e. the Asgard archaea and Alphaproteobacteria, respectively, makes this an especially difficult task. One approach has been to propose the presence/absence of genomic markers as synapomorphies. These are chosen due to their highly conserved yet complex nature, the idea being that such markers are unlikely to arise multiple times in different lineages. For instance, the dihydrofolate reductase-thymidylate synthase (DHFR-TS) gene fusion was put forward as a possible derived trait in those eukaryotes that possess it (Chromalveolata, Excavata, Plantae and Rhizaria), whilst those in which the genes are split (Opisthokonta) were inferred to be ancestral since the two genes are also split in bacteria, therefore this ought to be the primitive condition, and the root ought to lie between these two eukaryote clades (Stechmann and Cavalier-Smith, 2002). This was soon revised to include Amoebozoa (whose position was previously uncertain) on the same side of the root as opisthokonts when it was discovered that this group also lacked the DHFR-TS gene fusion, and furthermore it was found that members of this Opisthokonta/Amoebozoa clade also possess a triple-fused multienzyme protein involved in pyrimidine synthesis – a fusion not observed in any bikont organisms (Stechmann and Cavalier-Smith, 2003).

The unikont-bikont root, positioned between opisthokonts and all other eukaryotes (Cavalier-Smith, 2002), was based on the inferred cytoskeletal condition of the ancestral eukaryote, although with the phylogenomic placement of non-unikont taxa into the unikont clade, for example breviates (Minge et al., 2009), this morphology-based scheme appeared somewhat uncertain (Roger and Simpson, 2009). However, with the unikont-bikont root recovered in molecular analyses (Derelle and Lang, 2012) and the unikont clade frequently found in unrooted analyses (Burki et al., 2007; Hampl et al., 2009; Burki et al., 2016), the unikont-bikont divide has remained a popular notion.

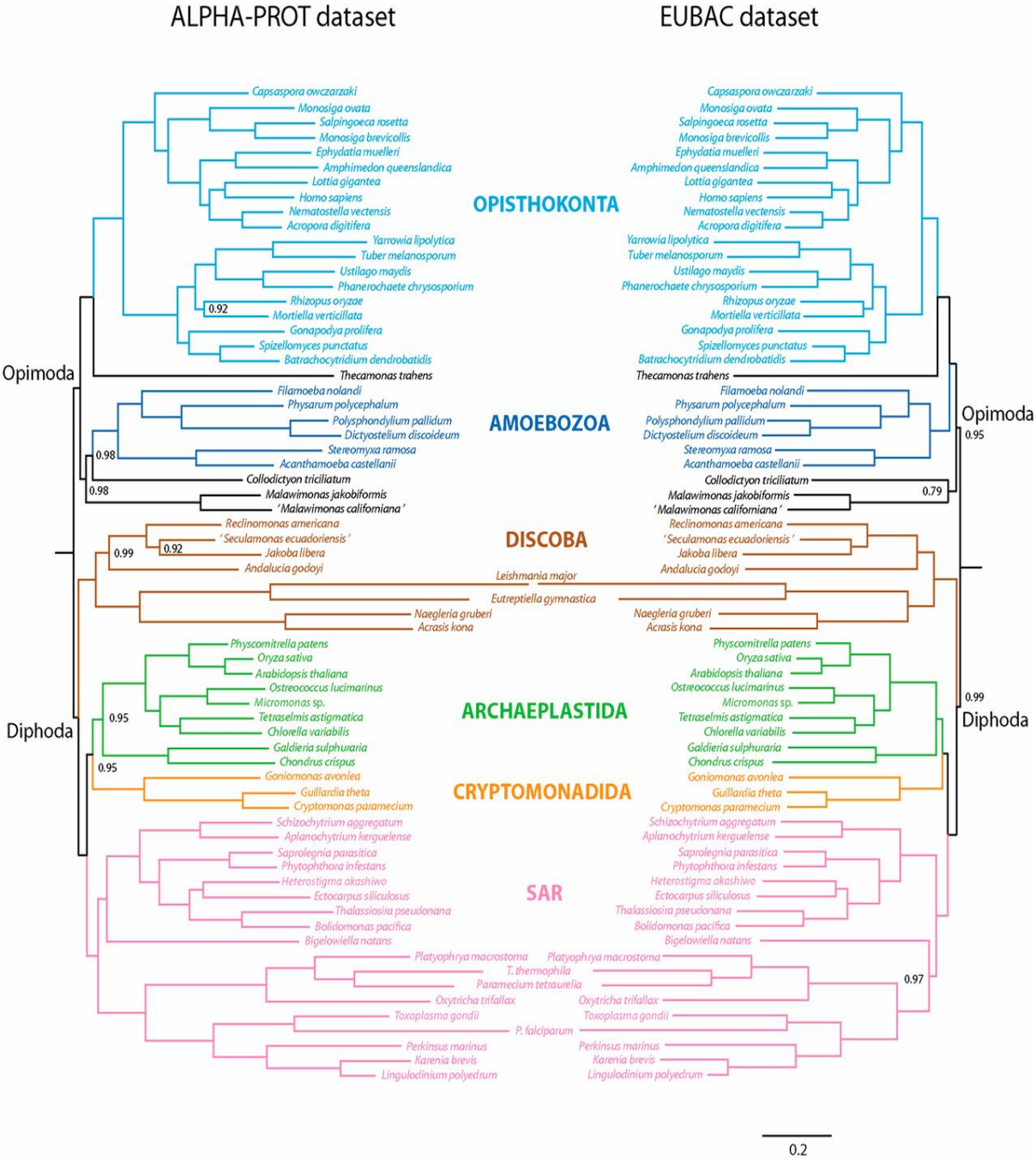

Figure 2 – **Rooted eukaryote phylogeny.** Two bayesian consensus trees constructed with different datasets display possible root positions for the eukaryote tree under the CAT-GTR+Γ4 model. Taken from Derelle et al. (2015).

An alternative position placed the root between Euglenozoa (comprising euglenids, trypanosomatids, kinetoplastids and relatives) and all other eukaryotes on the basis of a number of genomic and cellular characters (Cavalier-Smith, 2010), an arrangement that put the excavate taxa as a paraphyletic group with respect to all other eukaryotes. A phylogeny somewhat resembling this hypothesis was recovered by He et al. (2014), who used a set of mostly mi-

tochondrial genes, and included a number of bacteria – those with sequences inferred to be most closely related to the 37 eukaryote genes in the dataset – so as to obtain as close an outgroup as possible. Here the root was located between Discoba and all other eukaryotes in a well-supported tree, however metamonad representatives were conspicuously absent from the analysis (excluded on account of their lack of aerobic mitochondria).

Derelle et al. (2015) then reanalysed the dataset used in He et al. (2014) under a different model, and recovered a tree rooted between what closely resembled a unikont-bikont divide (the main differences being the position of *Malawimonas* and *Collodictyon* at the base of the unikonts). In the same study a highly similar tree was recovered using a modified version of an earlier dataset from Derelle and Lang (2012). Both trees can be seen in Fig. 2. The position of Discoba and Malawimonadidae on either side of the root alluded to the ancestral excavate hypothesis, though the authors alternatively suggested that typical excavate characters may have evolved multiple times; again the lack of metamonad representatives (due to the genes selected for this analysis being poorly represented in a group with unusual mitochondrial states) meant that part of the picture was still missing (Derelle et al., 2015).

These studies highlight the lack of consensus in attempts to locate the root of the eToL.

1.4 Excavate affinities to other groups

Beyond the recognised excavate taxa there are a number of other eukaryotes with morphological characters that bear resemblance to the typical excavate morphology. The question of whether these evolved independently or are retained plesiomorphies (plesiomorphy = ancestral character state) has not been answered convincingly. Most obvious of these putative homologs is the ventral groove. This character has been identified in CRuMs members *Collodictyon triciliatum*, *Diphylleia rotans* (Brugerolle et al., 2002), and *Sulcomonas lacustris* (Brugerolle, 2006), apusomonads (Karpov, 2007), ancyromonads (Heiss et al., 2011), colponemids (Tikhonenkov et al., 2014), and *Ancoracysta twista* (Janouškovec et al., 2017).

Another member of the CRuMs lineage, *Rigifila ramosa*, has a ventral aperture with a different structure. Nevertheless, it has been suggested that this character is still derived from the same ventral groove-bearing ancestor as the diphylleids (Yabuki et al., 2013b). Although ventral groove-bearing cells are present in the CRuMs lineage (*Collodictyon* and *Diphylleia* spp.), their relatedness to the excavates on the basis of other ultrastructural features of the flagellar apparatus has been doubted (Brugerolle et al., 2002). In contrast to this, the diphylleid

cytoskeleton has been inferred to be derived from an ancestral excavate (undergoing a reconfiguration of the microtubular roots and basal bodies), as has the apusomonad cytoskeleton, wherein both left and right roots are found to be highly similar in supporting the ventral groove, and the bicosoecid stramenopiles on the basis of the shared split right root character (Cavalier-Smith and Chao, 2010). Similarly it has been proposed, based on the scattered distribution of various components of the flagellar apparatus in the major eukaryote groups, retained in entirety only by the excavates, that these components were present ancestrally in an excavate-like organism, and that Excavata is a paraphyletic group (Yubuki and Leander, 2013).

To summarise, *A. twista*, ancyromonads, bicosoecids (Strampenopiles), colponemids (Alveolata), and CRuMs members all have cytoskeletal components that are putatively homologous to the excavate cytoskeleton.

1.5 Evolutionary relationships amongst non-excavate orphans

Aside from the excavates, a number of "orphan" eukaryotes – *Ancoracysta twista*, ancyromonads, CRuMs, telonemids – have also proved difficult to place in the eToL. On the basis of morphology it is possible to make tentative links, for example the sectorised extrusomes of *Telonema subtilis* may be related to the ancoracyst of *A. twista* (Janouškovec et al., 2017), which may also be related to the extrusomes of *Metromonas simplex* (Rhizaria) (Yabuki et al., 2013a). *A. twista* has been linked to *Colponema marisrubri* (Alveolata), based primarily on similarities of the cortical alveoli (Cavalier-Smith et al., 2018). Though these morphological comparisons may appear tenuous, associations with the SAR supergroup are nevertheless interesting given the increasing phylogenomic evidence for a Telonemia + SAR relationship (Strassert et al., 2019).

Elsewhere in the eToL the CRuMs member *Mantamonas* has been associated with the ancyromonads on the basis of morphology, as has the orphan *Micronuclearia* (Glücksman et al., 2011). Molecular studies have supported the former relationship, albeit only weakly (Cavalier-Smith et al., 2014). Conversely CRuMs form a well-supported clade that does not branch with the ancyromonads in the study of Brown et al. (2018). In the case of the latter, *Micronuclearia* has been mostly absent from phylogenetic studies, yet it has been recovered in a concatenated 18S and 28S rDNA tree as the sister to *Rigifila ramosa* in a strongly supported clade, which was itself the sister clade to *Collodictyon triciliatum* (Yabuki et al., 2013b).

By comparing the number of OGs shared uniquely by different pairs of lineages, this study

aims to identify potential evolutionary relationships amongst deep-branching protists. In particular, it aims to clarify relationships amongst the three excavate taxa Metamonada, Discoba and Malawimonadidae, which exhibit striking morphological similarities, yet tend not to be recovered as a natural group in most molecular analyses.

Chapter 2

Data Chapter

2.1 Materials and methods

2.1.1 Proteome analysis

The protein sequence data of 97 taxa were studied, representing all known major eukaryote lineages. 15 distinct groups of eukaryotes were initially analysed (Fig. 3A). The original 15 were then further divided and reanalysed in order to obtain more fine-grained results for certain lineages: SAR was divided into Stramenopiles, Alveolata and Rhizaria; Haptista was divided into Haptophyta and Centrohelida, for a total of 18 eukaryote groups (Fig. 3B). Initially, 2,355,252 sequences were analysed using OrthoFinder 2.1 (Emms and Kelly, 2015), using DIAMOND (Buchfink et al., 2015) to align sequences and IQ-TREE 2.0 (Minh et al., 2020) as the tree inference program. In this way, sequences were clustered into orthogroups (OGs) – groups in which all sequences present share an evolutionary history. To be considered an OG, a minimum of four sequences were required to cluster together. During the OrthoFinder clustering, a total of 843,109 sequences were not assigned to any OG. The remaining 1,512,143 sequences clustered into 45,821 OGs. The OG data used in this analysis can be viewed and downloaded using the following DOI: 10.6084/m9.figshare.14465367.

Custom Python scripts were used to identify which OGs were unique to each pair of eukaryote lineages analysed (find_group.py, group.py, og_to_heatmap.py – GitHub repository for these scripts can be found at https://github.com/RDScambler/Ex_OG_code/tree/master/OG_results_pipeline). This required that, at minimum, one member of both lineages under consideration should be present in the OG, with no taxa from any other lineages present. A high number of uniquely shared OGs was taken to signify a potential relationship, indicating that two eukaryote

lineages may form a clade and their OGs may represent synapomorphies.

2.1.2 Search for functional OGs

Using the same criteria as for the pairwise searching, unique OGs shared among three or more lineages were searched within the dataset. This allowed OGs of potential functional importance relating to the taxa under consideration to be identified. In particular, searches were focused on the identification of OGs unique to lineages with common cytoskeletal features, namely those that make up the typical excavate morphology. Lineages targeted for their ventral groove character were *Ancoracysta twista*, Ancyromonadida, Apusomonadida, CRuMs, Discoba, Malawimonadidae and Metamonada. Although the colponemids, belonging to Alveolata, are also known to possess a ventral groove (Tikhonenkov et al., 2014), there were in fact no colponemid species in the dataset being searched, so this lineage was excluded from these searches. Lineages with other cytoskeletal characters of interest were selected using the morphological evidence described in Yubuki and Leander (2013).

2.1.3 Calculation of proportional proteome data

In order to determine total proteome size of each eukaryote lineage, a custom Python script (total_genome.py) was used to count the number of OGs with at least one sequence belonging to a member of the eukaryote lineage in question. The OG total for each lineage was used to calculate the uniquely shared OG data as a proportion of its collective proteome, i.e. the total protein content of a lineage, accounting for the fact that not every species within a lineage has every protein of that lineage (excepting instances where species are the sole representatives of their lineage). This will be the intended meaning when referring to a group's proteome throughout the text. Since the total proteome size of each lineage varies, the number of OGs that two lineages share uniquely will convert to different percentages of the total proteome, depending on which of the two lineages is being considered.

2.1.4 Identification of contaminants and functional annotation

OGs of interest were tested for contamination using eggNOG 2.0 (Huerta-Cepas et al., 2017, 2019), aligned with DIAMOND (Buchfink et al., 2015). The default e-value cut-off of 0.001 was used. If 70% or more of the sequences of an OG recovered best-matching orthologs of bacterial origin, this OG was removed from further analysis. OGs remaining in the analysis were annotated using eggNOG. Functional categories are based on those defined in the COG

database (Galperin et al., 2015). Information from this analysis was corroborated with functional annotation of protein domains inferred using Interproscan 5.44 (Mitchell et al., 2019).

2.1.5 Constraint analysis

Maximum likelihood phylogenies of the 97 taxa were reconstructed using IQ-TREE 2.0 (Minh et al., 2020). An alignment of 42 proteins, totalling 11,836 amino acid sites, was analysed under the LG+F+R8 model. This was found to be the best-fitting model according to the Bayesian Information Criterion (BIC). A second analysis of the same 97 taxa was carried out using an alignment of 116 proteins, totalling 20,272 amino acid sites. The BIC also determined LG+F+R8 to be the most appropriate model for this alignment.

Topological constraint tests were performed with IQ-TREE under the LG+F+R8 model. The approximately unbiased (AU) test (Shimodaira, 2002) was performed, along with the one-sided Kishino-Hasegawa (KH) test (Kishino and Hasegawa, 1989), and the Shimodairo-Hasegawa (SH) test (Shimodaira and Hasegawa, 1999), in order to compare the maximum likelihood tree with alternative tree topologies, using 10,000 RELL replicates (resampling estimated log-likelihoods). This analysis was performed for both the 42 and 116 alignments.

2.1.6 Identification of cytoskeletal proteins by BLAST searching of dataset

In order to identify proteins potentially involved in excavate cytoskeletal structure within the dataset, a number of sequences from public repositories were searched by using BLASTp (Altschul et al., 1997). The following proteins were searched for: SF-assemblin (beta- and delta-giardin), alpha-giardin and gamma-giardin (Dawson, 2010); centrins (Weerakoon et al., 1999); costa proteins (Viscogliosi and Brugerolle, 1994); proteins associated with the cytoskeleton of *T. vaginalis* (Preisner et al., 2016).

Proteins producing significant alignments (with an e-value cut-off of 0.001) to sequences belonging to primarily excavate taxa, as well as non-excavate taxa with excavate-like cytoskeletal characters, were further analysed by inspection of their phylogenies using the Interactive Tree of Life (iTOL) (Letunic and Bork, 2019).

2.2 Results

2.2.1 Analysis of uniquely shared OGs

Uniquely shared OGs were compared between different pairs of taxa, with the eukaryotes in this dataset split into 15 distinct groups (Fig. 3A) and then reanalysed in 18 distinct groups (Fig. 3B). In the 15-way analysis, the greatest number of uniquely shared OGs between two different lineages was 947 (SAR + Haptista), followed by 609 (*Ancoracysta twista* + Telonemia). The 609 of *A. twista* + Telonemia became the highest number of uniquely shared OGs in the 18-way analysis upon the splitting of the SAR and Haptista lineages, followed by Ancyromonadida + CRuMs with 383.

In the 15-way analysis, the *A. twista* + Telonemia relationship is equally pronounced when the total number of OGs are converted to the percentage of each group's proteome (Fig. 4A). The 609 OGs account for 7.55% of the *A. twista* proteome, and 5.23% of the Telonemia proteome. The SAR + Haptista relationship is less pronounced, with the 947 OGs accounting for 6.67% of the Haptista proteome and 3.43% of the SAR proteome. In the 18-way analysis, after *A. twista* + Telonemia, the Ancyromonadida + CRuMs relationship is the most significant when total OGs are converted to percentage of proteome (Fig. 4B). The 383 OGs account for 5.23% of the CRuMs proteome and 3.92% of the Ancyromonadida proteome. Every pair of eukaryote groups shared at least one OG uniquely, with a few exceptions: Apusomonadida + Hemimastigophora, Apusomonadida + Malawimonadidae, and Haptophyta + Malawimonadidae.

The three excavate lineages shared a comparatively small number of unique OGs with one another (Fig. 5). Notably, no OGs were found to be exclusively shared by Metamonada, Discoba and Malawimonadidae together. As a supergroup comparison to the excavates, an Alveolata + Rhizaria + Stramenopiles search was conducted, resulting in 31 uniquely shared OGs (0.17% of Alveolata's, 0.36% of Rhizaria's and 0.22% of Stramenopiles' proteome). Further, megagroup searches of the OGs unique to members of Amorphea (Amoebozoa + Apusomonadida + Opisthokonta) and Diaphoretickes (Alveolata + Archaeplastida + Centrohelida + Cryptista + Haptophyta + Rhizaria + Stramenopiles + Telonemia) resulted in 3 and 4 uniquely shared OGs, respectively.

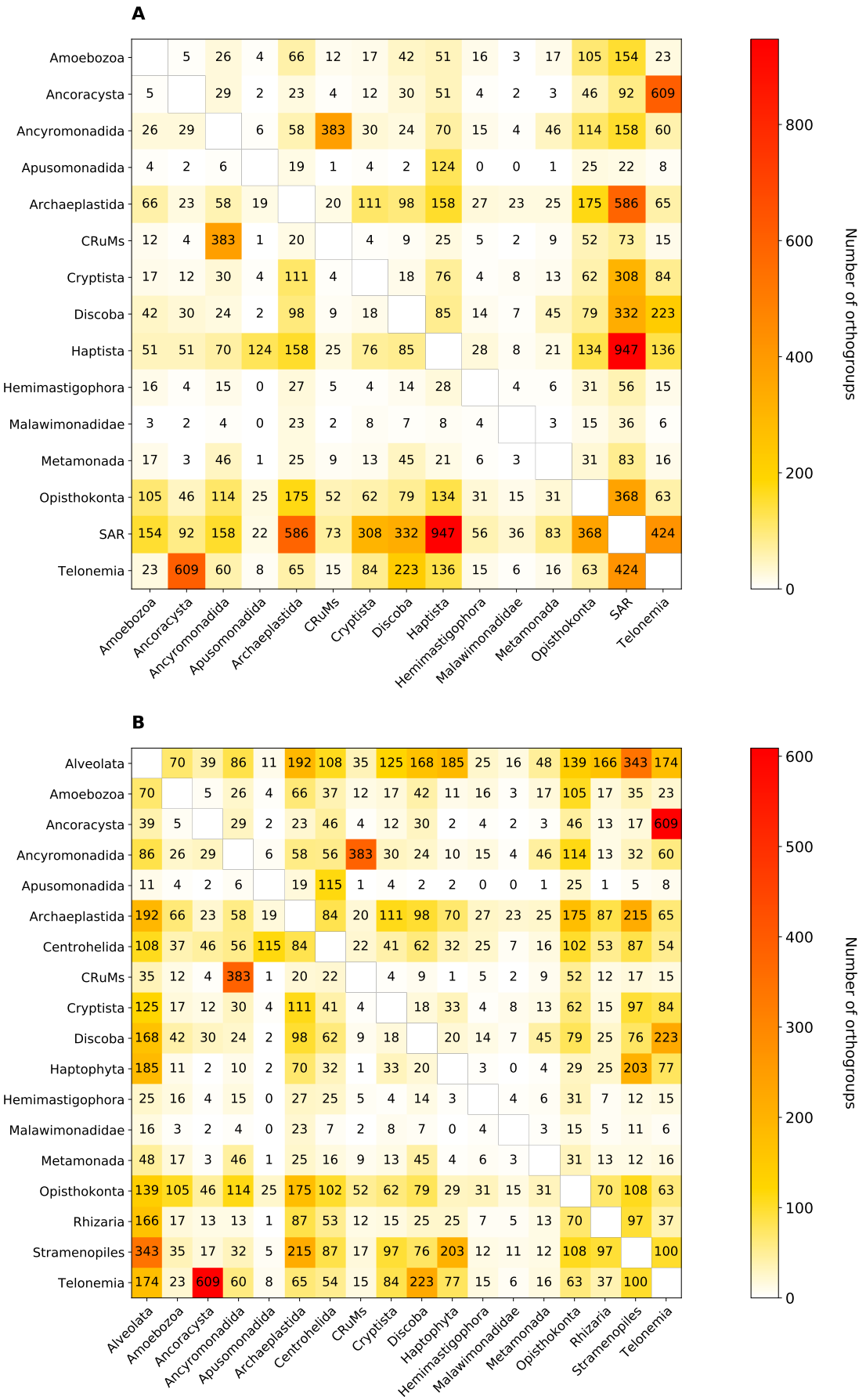

Figure 3 – **Heat maps of uniquely shared OGs.** Eukaryotes are divided into (**A**) 15 and (**B**) 18 distinct groups and their OGs compared.

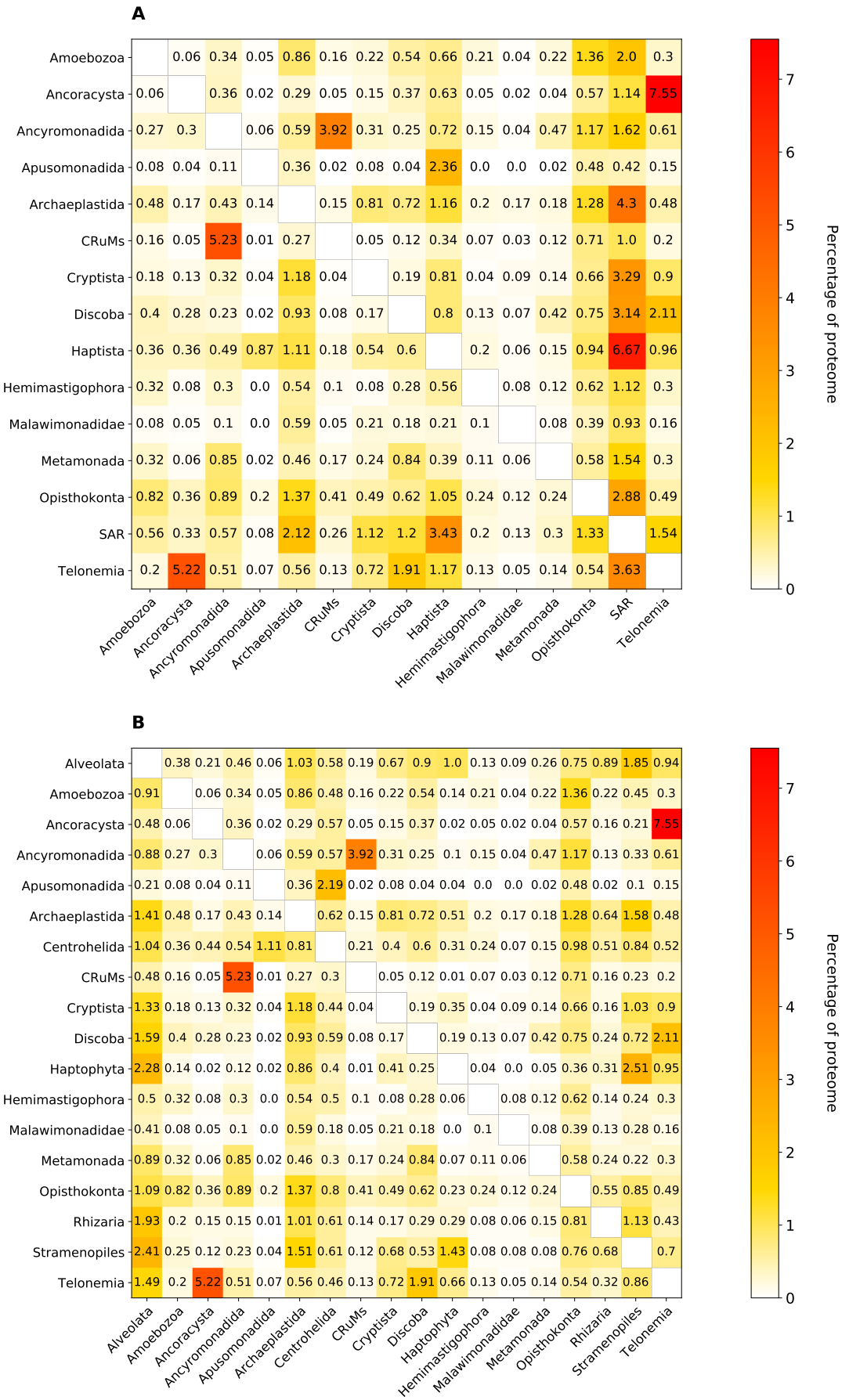

Figure 4 – Heat maps of uniquely shared OGs as a percentage of proteome. The same data from Fig. 3 are used to calculate the percentage of each group's total proteome that is made up of uniquely shared OGs. Percentages correspond to the proteome of the eukaryote group in each row, calculated from the number of OGs shared uniquely with the group in the corresponding column. Analyses are done for the (**A**) 15-way and (**B**) 18-way split.

A search for OGs belonging exclusively to ventral groove-bearing lineages found there to be no OGs common to all of these lineages alone. Another round of searches was performed, allowing for one of each lineage in turn to be missing from the OGs found. Again, no OGs were found in the search.

Searches for lineages sharing other interesting aspects of cytoskeletal organisation did not return any unique OGs when eukaryote 'sets' were defined based on the presence of the following cytoskeletal features: microtubular root R1 with a multi-layered structure, split right root R2, singlet root, microtubular root R3 with an array of superficial microtubules, and microtubular root R4 (presence/absence of these features is based on Yubuki and Leander (2013), although due to uncertainty over the assignment of characters to members of either Apusomonadida, Ancyromonadida or both, since they are sometimes considered together within Apusozoa, searches were conducted including both, and then excluding each subgroup individually).

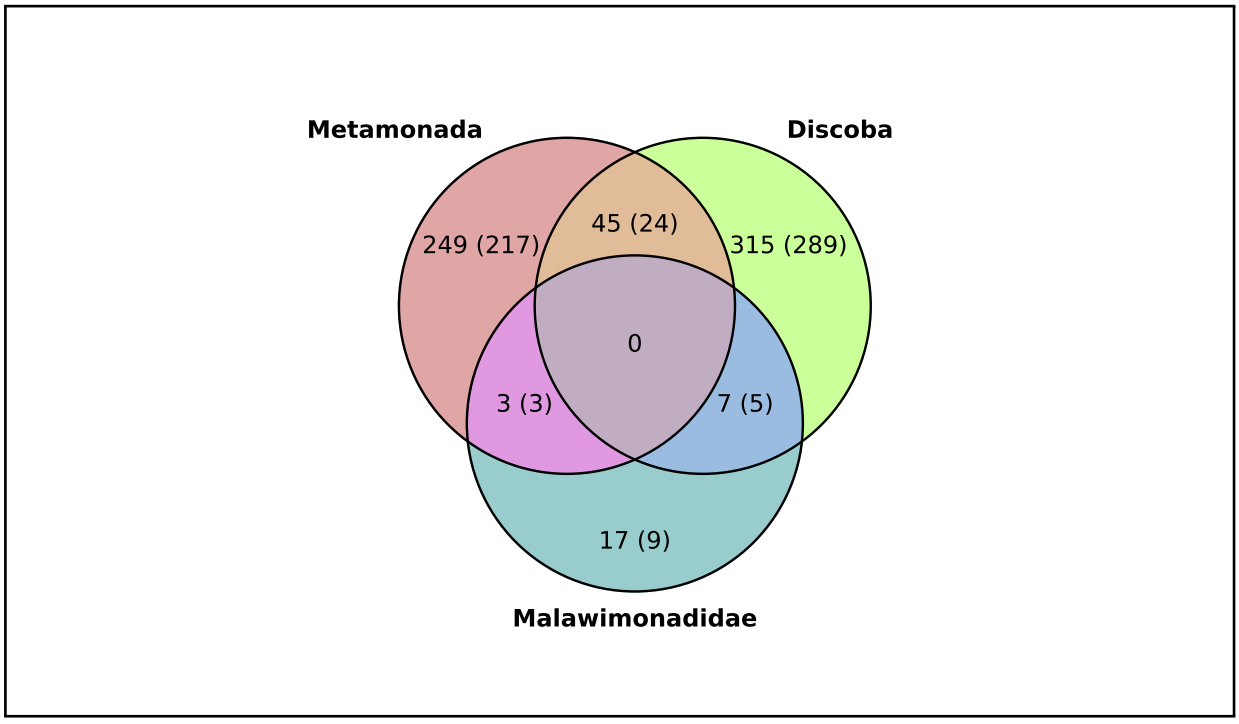

Figure 5 – OGs shared uniquely between the excavate taxa. The venn diagram displays each excavate lineage with the number of OGs unique to itself, as well as unique pairwise OGs and those unique to all three. Numbers in brackets indicate adjusted OG values after the removal of probable bacterial contaminants from each OG set.

2.2.2 Contaminant removal and functional annotation

Ancoracysta twista and Telonemia

Sequences of the 609 OGs shared exclusively between *Ancoracysta twista* and Telonemia were analysed using eggNOG 2.0. A total of 195 OGs returned hits in eggNOG. Of these, 17 were removed as potential bacterial contaminant sequences. Functional categories were inferred from the remaining 178 OGs (Fig. 6). The most frequently occurring category was 'unknown' (37), followed by 'signal transduction mechanisms' (32). The most consistent annotation, found in eight OGs, related to calcium-binding sites/domains and calcium-dependent channels, including voltage-sensitive calcium channels (VSCCs).

Ancyromonadida and CRuMs

Of the 383 OGs shared exclusively between Ancyromonadida and CRuMs, 60 returned hits in eggNOG. Of these, 33 were removed as potential bacterial contaminant sequences. Functional categories were inferred from the remaining 27 OGs (Fig. 6). The most frequently occurring category was 'unknown' (9), followed by 'cell cycle control, cell division and chromosome partitioning', 'transcription', 'signal transduction mechanisms' and 'cytoskeleton' (3).

Excavate taxa

The 45 OGs shared exclusively by Discoba and Metamonada returned a total of 38 hits in eggNOG. Of these, 21 were removed as potential bacterial contaminants. Functional categories were inferred from the remaining 17 OGs (Fig. 6). The most frequently occurring category was 'unknown' (8), followed by 'replication, recombination and repair' (4). The seven OGs shared exclusively between Discoba and Malawimonadidae returned three hits, two of which were suspected of being bacterial contaminants. The remaining OG was annotated as a RAN GTP-ase activating protein. The three OGs shared exclusively between Malawimonadidae and Metamonada returned one hit, not thought to be a bacterial contaminant, annotated as belonging to a multiprotein E3 ubiquitin ligase complex involved in protein quality control.

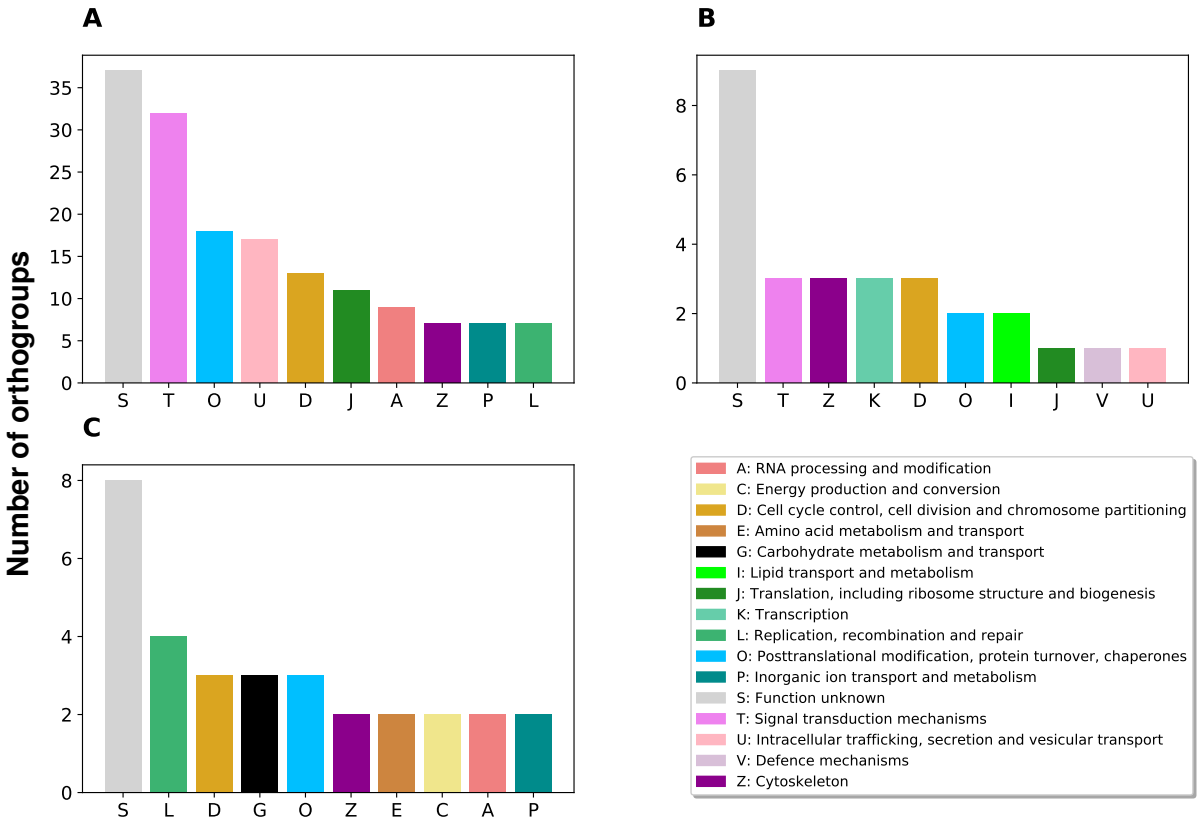

Figure 6 – Functional categories of uniquely shared OGs. Different functional categories are shown as inferred from eggNOG 2.0. Note some OGs are assigned multiple categories, and some did not return any annotation. OGs are shared between (**A**) *Ancoracysta twista* + Telonemia, (**B**) Ancyromonadida + CRuMs, and (**C**) Discoba + Metamonada. The other excavate pairwise comparisons are omitted due to lack of eggNOG annotation.

2.2.3 Distribution of cytoskeletal proteins

The majority of excavate-associated cytoskeletal proteins in the literature examined here (Viscogliosi and Brugerolle, 1994; Weerakoon et al., 1999; Dawson, 2010; Preisner et al., 2016) were not found to be specifically associated with excavate taxa when searched in the dataset, either being ubiquitous or without significant matches in any taxa. However, SF-assemblin and delta giardin did produce significant matches with excavate taxa. Both proteins matched with the same OG, confirming their common evolutionary ancestry. SF-assemblin had significant similarity to sequences in every eukaryote lineage in the dataset with the exception of Centrohelida, Opisthokonta, CRuMs, Apusomonadida and Hemimastigophora. Delta giardin had significant similarity only to sequences of the Metamonada.

2.2.4 Comparison of individual species' OGs within Ancyromonadida and CRuMs

In order to obtain a better idea of potential specific relationships between members of Ancyromonadida and CRuMs, the number of OGs belonging to individual species were counted and compared. These groups were focused on for closer post hoc analysis due to their high number of uniquely shared OGs as well as the high morphological diversity of the constituent species. Unlike *A. twista* and Telonemia, which are comprised of just one genus each, the members of Ancyromonadida and CRuMs are represented here by five species in five different genera. If the 383 OGs shared uniquely between Ancyromonadida and CRuMs are unevenly distributed, specific relationships may be revealed.

Scanning the 383 OGs revealed the following in Ancyromonadida species: *Ancyromonas sigmoides* (312), *Fabomonas tropica* (250), *Nutomonas longa* (319). In CRuMs: *Diphylleia rotans* (267), *Rigifila ramosa* (272).

There does not appear to be any significant bias in number of OGs found in the constituent species of Ancyromonadida and CRuMs. Members of the same taxon can be seen to share a similar number of OGs out of the 383, except in the case of *Fabomonas tropica* which has 62-69 fewer OGs than the other members of Ancyromonadida. However, in neither taxon is there a single standout species with significantly more OGs.

2.2.5 Constraint analysis

In order to further assess the likelihood of true evolutionary relationships between *A. twista* and Telonemia, and between Ancyromonadida and CRuMs, phylogenetic constraint analyses were conducted with IQ-TREE 2.0. Results of the AU, KH and SH test can be seen in Table 1.

Neither the constrained topology of *A. twista* + Telonemia, nor that of Ancyromonadida + CRuMs, could be rejected at a 5% level of significance in any of the tests. In the 42 alignment the *A. twista* + Telonemia topology had the lowest p-value for the AU test at 0.193, whereas in the 116 alignment the Ancyromonadida + CRuMs topology had a lower p-value, at 0.071

Topology	logL	deltaL	p-KH	p-SH	p-AU
42 alignment					
Unconstrained	-782520.4485	0	0.736	1	0.8
((A, Cr) E)	-782539.6124	19.164	0.265	0.491	0.324
((At, T) E)	-782577.7984	57.35	0.166	0.234	0.193
116 alignment					
Unconstrained	-1200255.118	0	0.9	1	0.926
((A, Cr) E)	-1200357.348	102.23	0.0593	0.0806	0.071
((At, T) E)	-1200323.06	67.942	0.0998	0.193	0.134

Table 1 – Constraint analysis of alternative tree topologies. The results from IQ-TREE 2.0 include the Kishino-Hasegawa (KH) test, Shimodaira-Hasegawa (SH) test and approximately unbiased (AU) test. Testing was performed on the maximum likelihood trees inferred from two alignments of 42 and 116 proteins. Results are shown comparing the unconstrained maximum likelihood tree with two alternative topologies. **A** = Ancyromonadida, **Cr** = CRuMs, **At** = *Ancoracysta twista*, **T** = Telonemia, **E** = all other eukaryotes.

Chapter 3

Discussion

The presence of OGs that are unique to any given taxa indicates a shared evolutionary history of some kind. However, patterns of gene gain/loss as well as lateral gene transfer (LGT) – a phenomenon known to occur in eukaryotes as well as prokaryotes (Andersson, 2005) – within and amongst different lineages means it is difficult to say with certainty whether the OGs found in two taxa uniquely are present due to relatively recent genomic innovations, or are instead the remnants of more ancient gene families that have been lost in other lineages. An argument could, in theory, be given for any tree topology if a particular series of gene gain/transfer/loss events is invoked.

Nevertheless, there is a case to be made for taxa that share a significant number of unique OGs, especially when corroborated with other lines of evidence, as it becomes more parsimonious to conclude that they represent relatively recent genomic changes that occurred in a putative common ancestor after its divergence from other lineages but before the divergence of the taxa in question. In other words, they are more likely to be synapomorphies.

3.1 Evolutionary relationships among the excavate taxa

Based on the data used in this study, there does not appear to be strong evidence for a recent evolutionary relationship between any of the three excavate taxa, either when examining each of their pairwise relationships or when considering the three together as a set. The 45 OGs uniquely shared between Discoba and Metamonada is the most significant of the excavate taxon comparisons (equating to 0.84% of Metamonada's and 0.42% of Discoba's proteome), and this is not especially significant relative to the comparisons of other taxa. The fact that no OGs are shared uniquely between Discoba, Malawimonadidae and Metamonada does nothing

to support the idea of a monophyletic Excavata clade. The probable non-monophyly of the excavate taxa suggested by this study is in agreement with recent phylogenomic analyses (Brown et al., 2018; Lax et al., 2018; Price et al., 2019).

One notable issue in identifying relationships using genomic data is the demarcation of OG boundaries. The method for clustering OGs may disfavour sequences with sufficiently low similarity due to rapid evolution. In fact, it has been demonstrated that for at least some genes, assumed to be lineage-specific, orthologs evolving at a constant rate will likely fail homology searches even if present in outgroup lineages (Weisman et al., 2020). If sequences found in the three excavate lineages are fast-evolving, their orthologous sequences are more likely to have diverged to a point where they no longer cluster in the same OG. Certainly in the case of Metamonada species have undergone rapid evolution, and this is evidenced in the long-branched sequences seen in molecular phylogenies that include them (Simpson et al., 2006a).

3.2 Comparison of Uniquely shared OGs in other supergroups and megagroups

The absence of OGs shared uniquely between the excavates is notably lower than the uniquely shared OGs of other supergroups analysed. This analysis recovered Obazoa (Apusomonadida + Opisthokonta) with 25, SAR (Alveolata + Rhizaria + Stramenopiles) with 31, and Haptista (Centrohelida + Haptophyta) with 32 uniquely shared OGs.

When supergroups are further lumped into even larger assemblages – Amorphea and Diaphoretickes – the results are closer to those of the excavate analysis. Amorphea share just 4 OGs uniquely, and Diaphoretickes just 3. These data appear to be compatible with the idea that the excavates diverged at a comparatively deep node in the eToL to these megagroups (i.e. the excavates may be monophyletic, but are a very ancient branch) which have all similarly experienced the loss of most/all of their shared OGs over extended lengths of evolutionary time. However, it should be noted that such inferences are highly speculative when based on a shared absence of supporting data. The fact that Diaphoretickes' subgroups show, for the most part, far greater affinity to one another than the excavate subgroups do, strengthens the case for the monophyly of the former relative to the latter. SAR + Cryptista have 308 uniquely shared OGs, SAR + Archaeplastida have 586, and SAR + Haptista have 947 – some of the most significant relationships in the analysis. The Amorphea subgroup comparisons are less significant, and fall in the same ballpark as the excavates: Amoebozoa + Apusomonadida share 4 OGs

uniquely, Apusomonadida + Opisthokonta share 25, and Amoebozoa + Opisthokonta share 105.

The lack of uniquely shared OGs within both Amorphea and Diaphoretickes reflects the lack of unifying morphological characters in each of these groups. The same cannot be said for the excavates.

3.3 Asymmetry in proteome size

The consequence of different eukaryote lineages having different proteome sizes is that the total number of OGs shared between two (or more) lineages may appear more or less significant depending on perspective, i.e. depending on which lineage's proteome is being considered.

The conversion of total OGs to proportion of proteome that those OGs make up allows for the relative significance for each lineage to be inferred (in Fig. 4 the percentage of each group's proteome corresponds to the lineage in the row). These apparent differences in significance are essentially due to differences in the extent of genomic diversity found within each group of eukaryotes. Those with a greater number of species can be expected to retain a greater number of gene families over the course of evolution, on average, and in this analysis only one member of a lineage is required to signify a potential relationship. This may explain the relatively high number of OGs that SAR (a species-rich lineage) shares exclusively with a number of other lineages in the 15-way analysis (Fig. 3A), since so long as SAR is regarded as a monophyletic group, the OGs it shares exclusively with each lineage cannot all be recent genomic innovations.

At the other extreme, Malawimonadidae has just one representative in this dataset, *Gefionella okellyi*. In addition to having low total numbers of exclusively shared OGs with other lineages, Malawimonadidae also has consistently low proportional proteome data. This suggests it has no close evolutionary relationship with any other lineages and is in line with the idea that *G. okellyi* is one of the few remaining extant species of a very deeply diverged lineage, perhaps one of the earliest-diverging lineages in the eToL. In future comparative studies it would be interesting to add the genomic data of members of the only other known genus in Malawimonadidae, *Malawimonas jakobiformis* and *Malawimonas californiana*, and compare their OGs with those of *G. okellyi*.

Malawimonadidae also has just 17 OGs that are unique to itself (9 after adjusting for probable bacterial contaminants), which suggests that, if it is a deeply diverged lineage, it is in a state of

genomic stasis, or otherwise it may have had a larger genome ancestrally that has, over time, undergone a reduction in size.

3.4 Consequences of lineage-splitting

Unsurprisingly, in the 18-way analysis, the number of exclusively shared OGs in Stramenopiles, Alveolata and Rhizaria, as well as in the Haptista subgroups Haptophyta and Centrohelida, is reduced. This is due to there being fewer representative species in each lineage – though this is accounted for when adjusting for proportion of proteome. OG numbers will also be reduced due to a multi-lineage masking effect, i.e. OGs that were unique to SAR + Haptista will be lost from the 18-way analysis whenever three or more subgroups are present, as the OG will no longer be considered unique to any given pair. As such it is expected that the sum of each subgroup's pairwise comparisons within SAR + Haptista will be less than the parent groups' total (this is observed: combined subgroup pairwise total = 661, SAR + Haptista = 947 OGs).

Although SAR and Haptista have each been recovered as monophyletic (Burki et al., 2007, 2016), it is nevertheless interesting to examine nodes closer to the tips of the eToL. Here the relationships within SAR are reaffirmed, to an extent: Alveolata + Stramenopiles share 343 OGs uniquely (1.85% of Alveolata's and 2.41% of Stramenopiles' proteome). Alveolata + Rhizaria is less significant with 166 (0.89% of Alveolata's and 1.93% of Rhizaria's proteome) as is Rhizaria + Stramenopiles with 97 unique OGs (1.13% of Rhizaria's and 0.86% of Stramenopiles' proteome).

When all three SAR lineages are analysed together only 31 OGs are found to be uniquely shared (0.17% of Alveolata's, 0.36% of Rhizaria's and 0.22% of Stramenopiles' proteome). Based on this the data support an Alveolata + Stramenopiles relationship more strongly than an overall SAR relationship. This is reflected in phylogenomic analyses, which tend to recover an Alveolata + Stramenopiles clade as the crown group of SAR (Burki et al., 2012, 2016; Strassert et al., 2019). A similar result is found between the constituent lineages of Haptista: Centrohelida and Haptophyta share 32 OGs uniquely (0.31% of Centrohelida's and 0.4% of Haptophyta's proteome). These results demonstrate that the comparative genomics approach used in this study is not always congruent with conventional phylogenomic analyses, in which Haptista is found to be monophyletic (Burki et al., 2016; Strassert et al., 2019) – although this is not always the case (Price et al., 2019).

3.5 Lack of reliability in morphological characters

As the quantity of genomic information has exploded in recent years and phylogenetic studies have incorporated more and more loci with broader taxon sampling, eukaryote phylogenies have transformed into trees comprised of a small number of 'supergroups'. Members of the supergroups as they are currently defined do not, in most cases, have shared derived characters (Burki et al., 2019), although this depends on how the supergroups are delimited (e.g. Opisthokonta has shared derived characters; Amorphea does not).

It may be the case that, at the supergroups level, identifying shared morphological/cell biological characters may not be as crucial for inferring evolutionary relationships as once thought. There is great variability within these groups, and often similarities with members of other groups (Baldauf, 2003) that are now thought to result from independent evolutionary events. For instance, the red plastids of chromalveolates were once thought to share a common origin (Cavalier-Smith, 1999), but it is now clear that this is not likely to be the case, and the chromalveolates do not form a natural clade (Burki et al., 2016). The ever-transforming eToL has necessitated a reinterpretation of the origin of characters such as this, identifying many as evolutionary red herrings. The amoeboid state of some protists is another example of a morphological character once thought to have a single evolutionary origin, but since soundly disproven (Pawlowski, 2008).

3.6 Ancestral and vestigial characters

Did the excavate cytoskeleton evolve multiple times? An important aspect of the excavate hypothesis is the possibility that the excavates are a paraphyletic group – an evolutionary scenario that has been suggested by various different authors (Simpson, 2003; Yubuki and Leander, 2013; Cavalier-Smith et al., 2014). If the hypothesis is correct, and the excavate cytoskeleton is homologous, then a paraphyletic Excavata would appear to be a plausible scenario given the results in this study. An alternative explanation is that the excavate cytoskeleton is homologous amongst eukaryotes but has been inherited via LGT. This may explain its presence in a number of scattered lineages, and does not necessarily imply that the condition is ancestral for eukaryotes. Yet another hypothesis is that the excavate cytoskeleton is polyphyletic. Multiple independent origins are a possibility that shouldn't be dismissed out of hand – the results of this study certainly cannot disprove it – although in the case of centrioles/basal bodies and the axoneme, at least, convergent evolution seems highly unlikely, as these structures are

remarkably conserved across eukaryote lineages and were thus likely present in the LECA (Carvalho-Santos et al., 2011). These structures act as the cell's microtubule organising centre (MTOC) (Yubuki and Leander, 2013). The microtubular roots and fibres emerging from the MTOC, as well as the ventral groove character, are less well conserved and present in fewer lineages, so a greater degree of uncertainty surrounds their evolutionary origin. Homologous structures are, by definition, comprised of homologous proteins, which this study has largely failed to identify both amongst the three main excavate taxa as well as in the non-excavate lineages possessing 'typical excavate' morphological characters (this is discussed in more detail in 3.7).

Simpson (2003) suggested that if a 'typical excavate' were proven, by phylogenetic analysis, to have diverged from within a group whose common ancestor was inferred to be a non-excavate-like organism, the excavate hypothesis would be proven incorrect. To date such a phylogeny has not been consistently recovered, however certain other species with excavate-like characters have arisen from within what are decidedly non-excavate lineages. For instance, Colponemidia, whose members possess a ventral groove, a vane on the posterior flagellum and a split right microtubular root (R2), has been identified as the deepest-branching lineage within Alveolata (Tikhonenkov et al., 2020). Yet the group are still nested within SAR and other lineages within Diaphoretickes, so if the excavate cytoskeleton is homologous in Colponemidia it must have been lost several times in its neighbouring lineages.

Interestingly, a sister genus to *Colponema* (Colponemidia), *Acavomonas*, has been described as being highly similar to *Colponema* except in its lack of a ventral groove (Tikhonenkov et al., 2014), which suggests this character can be readily lost (or gained) over relatively short evolutionary distances. Loss of this character may be in response to a change in ecological role, i.e. feeding behaviour, in one of the genera, although both are described as obligate predators of smaller flagellates (Tikhonenkov et al., 2014). A more ancient ventral groove loss event has been proposed by Cavalier-Smith (2018), who has suggested that Rhizaria evolved from a *Colponema*-like ancestor, gaining filopodia as an adaptation to benthic feeding and eventually losing the ventral groove. This character may not therefore be a prerequisite for a free-living heterotrophic lifestyle, however all known ventral groove-bearing cells do appear to be free-living flagellated heterotrophs. It has also been suggested that the ventral groove was an optimal morphological arrangement for enabling the engulfment of algal cells and thus facilitating endosymbiosis events and plastid evolution (Tikhonenkov et al., 2020).

Taking a broader view of the tree, it can be seen that a number of other organisms, occurring

in a number of different lineages, possess excavate-like characters in some form or another. Table 2 lists putatively homologous excavate-like characters in non-excavate lineages. In some cases there is evidence that typical excavate characters are homologous but have undergone differentiation. For instance, the split right microtubular root character is thought to have completely separated into two roots in diphylleids (Cavalier-Smith and Chao, 2010). As the number of taxa identified as having characters homologous to the excavate cytoskeleton increases, the hypothesis that the excavate condition is ancestral to most if not all extant eukaryotes becomes more plausible. Various combinations of cytoskeletal characters are present throughout the eukaryotes, but typical excavate taxa alone have a complete set of the major flagellar apparatus components (Yubuki and Leander, 2013). If this condition is ancestral, it would imply that the LECA was already a fairly complex organism (opposing the traditional evolutionary view of simple life forms giving rise to those of ever-increasing complexity). Recent genomic analyses give support to this idea (Fritz-Laylin et al., 2010). It has been suggested that the LECA was as complex as it was, with no evidence of any earlier intermediate forms remaining extant, *because* this organism's complexity enabled it to comprehensively outcompete simpler forms that may have lacked a mitochondrial symbiont and advanced cytoskeleton (Wickstead and Gull, 2011).

Notably, some lineages within the excavate taxa themselves, confirmed as excavates through phylogenomic analysis, have undergone rapid differentiation and loss of typical characters. For example, diplomonad and parabasalid genera such as *Giardia* and *Trichomonas*, confirmed as belonging to Metamonada (Simpson et al., 2002b, 2006a), have morphologically adapted to parasitic lifestyles, thus losing the ventral groove character and associated cytoskeletal components. In what appears to be something of a transitory state between typical excavate and differentiated parasite, commensal oxymonads possess a vestigial, almost disappeared ventral groove character (Simpson et al., 2002a). Within Metamonada there seems to be a pattern of reductive evolution linked to a change in lifestyle. Similar lifestyle change and morphological differentiation can be seen in Discoba. Euglenozoans, many of which are parasitic (Vesteg et al., 2019), do not possess the typical excavate morphology (Simpson, 2003), yet it is predicted that these taxa evolved from ancestors morphologically and ecologically similar to jakobids (Simpson et al., 2006b), i.e. organisms that *do* possess typical excavate characters.

If it is possible for different excavate taxa to readily lose their excavate traits in response to changing ecological conditions, it is conceivable that all eukaryote lineages could, in theory, have evolved from an ancient excavate-like organism. In this scenario the LECA would have been a heterotrophic flagellate feeding via the ventral groove. During subsequent eukaryote

Taxa	Excavate Characters				References
	VG	SRR	SR	MF (R3)	
Ancoracysta twista	present				Janouškovec et al. (2017)
Ancyromonadida	present	present	present	present	Heiss et al. (2011)
Colponemidia	present	present		present	Tikhonenkov et al. (2014)
CRuMs	present	present		present	Brugerolle (2006)
Stramenopiles		present	present	present	Yubuki and Leander (2013)

Table 2 – Cytoskeletal characters proposed as homologous to those of the excavate taxa. The characters of non-excavate eukaryotes are compared. **VG** = ventral groove, **SRR** = split right root, **SR** = singlet root, **MF (R3)** = microtubular fan associated with R3. Note: Brugerolle (2006) provides a morphological examination of *Sulcomonas lacustris* but does not imply that the cytoskeletal characters are homologous to those of excavates. This inference is made by subsequent authors (Cavalier-Smith and Chao, 2010; Yubuki and Leander, 2013).

evolution, as various ecological niches became available, different lineages would have diversified and, to a greater or lesser extent, lost the groove and associated flagellar apparatus in favour of new features – as with the filopodia of Rhizaria (Cavalier-Smith, 2018). Such a scenario would make those organisms still retaining the excavate cytoskeleton examples of 'living fossils', experiencing an extremely long period of morphostasis.

3.7 Absence of OGs linking taxa that have 'typical excavate' characters

In each of the set searches based on the flagellar apparatus components as recognised in Yubuki and Leander (2013), no set is recovered as having any uniquely shared OGs. Likewise, no uniquely shared OGs were returned in the search for ventral groove-bearing taxa (those that are reported in the literature – Table 2). This would seem to imply that there is either no conserved genetic basis for the components under scrutiny, or there are other lineages within the dataset that possess the components and have been overlooked. Regarding the former explanation, it does appear to be the case that proteins responsible for assembly of the flagellar apparatus exhibit a great deal of diversity given the structure's conserved morphology (Nabais et al., 2020). In some cases, proteins may have diverged in sequence to the point where they are no longer detectable and comparable (Hodges et al., 2010). This may be especially true in very ancient proteins – those that could have been present in the LECA.

Due to the vast evolutionary timescales that have elapsed since the origin of eukaryotes, it may be that each lineage has its own diverged set of proteins (to the point where they are unrecognisable by homology searches) that nevertheless maintain structurally and functionally similar flagellar apparatuses across lineages. It is also possible that although the macromorphological components of the flagellar apparatus have remained highly conserved throughout eukaryote evolution, individual proteins have been lost whilst the same general structure and function have been retained (Nabais et al., 2020). Intermediate filaments – important components in a variety of cytoskeletons – are not thought to be homologous in different eukaryote groups, but nevertheless display a high degree of structural similarity (Preisner et al., 2016). In this scenario excavate cytoskeletal proteins would be patchily distributed across lineages. Alternatively it may be that the same core set of proteins are conserved throughout eukaryotes (i.e. not unique to 'typical excavates') but can be utilised in different ways by the excavate taxa (Dawson and Paredez, 2013).

Looking specifically at the excavates, the uniquely shared OGs between Discoba and Metamonada do in fact include two proteins with a putative function in cytoskeleton formation and organisation. One of these is a regulator of chromosome condensation (RCC1) domain, which functions to inhibit the premature initiation of mitosis (Dasso, 1993), and therefore is not actually involved in the flagellar apparatus. The other is annotated as ARP1, a gene which is a component of the dynactin complex (Clark and Rose, 2006). ARP1 is more or less ubiquitous throughout the eukaryotes (Wickstead and Gull, 2011), so the fact that this OG contains only members of Discoba and Metamonada suggests it is some kind of divergent form. The dynactin complex is only known to interact with cytoplasmic dynein, as opposed to axonemal dynein – the type associated with flagellar activity (Carvalho-Santos et al., 2011) – yet although cytoplasmic dynein functions in vesicular transport, organelle positioning and spindle microtubule organisation (Raaijmakers et al., 2013), it has also been implicated in other cellular processes depending on its structural variants, and may recruit other non-dynein proteins and function to directly anchor microtubules (Schroer, 2004). This protein is therefore a candidate component of the typical excavate cytoskeleton, although notably it is not present in Malawimonadidae.

Among the OGs unique to each individual excavate lineage there was a lack of identified cytoskeletal proteins in the eggNOG analysis, with the exception of one OG unique to the Metamonada. This was annotated as a microtubule-binding, calmodulin-regulated, spectrin-associated protein (CAMSAP). Analysis of this protein's domains in Interproscan revealed that one end was comprised of a CKK domain, which functions to bind CAMSAP proteins to microtubules (Baines et al., 2009), while the rest of the sequence was unannotated. CAMSAPs

regulate microtubule dynamics, binding to microtubule minus-ends to prevent depolymerisation (Hendershott and Vale, 2014).

3.8 Excavate cytoskeletal proteins within the dataset

Excavates, along with all other eukaryotes, are known to possess a core set of conserved cytoskeletal structural proteins, including tubulins, and their associated motor protein families (Dawson and Paredez, 2013; Nabais et al., 2020). Beyond these, however, protein composition appears to be highly variable in the cytoskeletons of eukaryotes (Wickstead and Gull, 2011).

Certain proteins have been attributed to particular excavate structures, for example B-type costa proteins have been linked to the costae of trichomonads (Viscogliosi and Brugerolle, 1994). However, in the UniProt database, the protein 'Costa, isoform B' (A0A0B4KED9) is composed of kinesin-like domains, suggesting it is a motor-protein rather than a structural protein. A BLASTp search of the dataset confirmed that this protein is present in all eukaryote lineages. A protein of the same molecular weight (TVAG_339450, 118 kDa) was identified in *T. vaginalis* by Preisner et al. (2016), although this did not retrieve significant hits to the same OG in the current dataset, and in fact did not have significant matches with any metamonads (note *T. vaginalis* is absent from this dataset).

Giardin – a protein known to localise to the ventral disc of *Giardia lamblia* – has a common evolutionary origin with SF-assemblin (Dawson, 2010), so it is not surprising that the most significant sequence matches of both proteins belonged to the same OG. However, the fact that delta giardin here only finds significant matches with Metamonada species within this OG serves as an example of how a family of homologous proteins can diverge into lineage-specific functions whilst remaining in the same OG. This highlights the potential masking effect of lineage- and function-oriented bioinformatics searches.

The cytoskeletal proteins searched for here have been studied in medically-important parasites – i.e. excavates that no longer possess the typical excavate cytoskeleton, and can therefore be reasonably expected to have proteins that are divergent from those of typical excavates. However, typical excavate species are known to be the closest relatives of these organisms (Simpson et al., 2002b), and so sequence comparisons may still be useful in attempting to infer the molecular composition of the excavate cytoskeleton. Ultimately, progress in this area will be made with more proteomic studies, such as that of Preisner et al. (2016), that focus not on divergent parasites but instead on species with the characteristic excavate cytoskeleton.

3.9 Evolutionary position of *Ancoracysta twista* and Telonemia

We identified a potential relationship between *A. twista* and Telonemia based on OG comparisons and phylogenetic analyses. The 609 OGs (592 after subtracting probable bacterial contaminants) that the two taxa share uniquely is the highest overall number of any pairwise relationship (in the 18-way analysis), and furthermore the topological constraint analysis forcing the two taxa into a clade could not be rejected by the AU test. This putative relationship is interesting since there is no mention of it in the literature up to this point (the supplementary material of Janouškovec et al. (2017) does include an *A. twista* + Telonemia clade with a phylogeny based on the rDNA operon sequence, however this node is not well supported).

Recently, Telonemia has been recovered as a sister lineage to the SAR group, leading to the proposition that the supergroup ought to be renamed 'TSAR' and the assertion that Telonemia is a true eukaryotic phylum in its own right (Strassert et al., 2019). The 424 OGs recovered in this analysis that are uniquely shared between SAR + Telonemia (15-way split) lends support to this – after *A. twista*, this is the most uniquely shared OGs Telonemia has with any group. Meanwhile, *A. twista* has been recovered branching as the sister group to Haptista (Janouškovec et al., 2017). Both taxa are therefore confidently recovered within the megagroup Diaphoretickes.

In terms of morphology, *A. twista* + Telonemia have two notable cellular features in common, the presence of extrusomes (known as ancoracysts in *A. twista*) and structures resembling cortical alveoli (Klaveness et al., 2005; Janouškovec et al., 2017; Cavalier-Smith et al., 2018). These characters are not unique to these taxa, however. Cortical alveoli are found in members of Alveolata, and these may be homologous to those found in glaucophytes (Archaeplastida) (Cavalier-smith, 2003). Extrusomes are found in a variety of protists, and may not all be homologous (Kugrens et al., 1994), however those found in *A. twista* and *Colponema marisrubri* (Alveolata) have been assessed as being nearly identical, to the extent that *C. marisrubri* has been reassigned to *Ancoracysta marisrubri* and a new family established, Ancoracystidae, by Cavalier-Smith et al. (2018) – although currently there is no molecular data available to support this.

In another case of organelle affinity, the extrusomes of *Telonema* spp. and *Metromonas simplex* have been identified as being highly similar (Yabuki et al., 2013a). *M. simplex* belongs to Monadofilosa (Rhizaria), however as with *C. marisrubri* this species lacks publicly available genomic data, and so both of their assignments are not currently supported by phylogenomic analyses. It would be interesting to include these taxa in future genomic analyses to determine

if their current placement within SAR is inaccurate and they in fact cluster with either Telonemia or *A. twista* or both in a new lineage. Such a coalescence of newly-sequenced species around orphan taxa is not unheard of (though note this prediction is based on a small number of morphological characters), nor is it unknown for species with a previously assumed phylogenetic position to be drastically reallocated in light of new molecular analyses. For instance, Hemimastigophora has been previously assigned to Euglenozoa (Simpson, 1997), but is now known to occupy its own deeply diverged lineage (Lax et al., 2018).

Though there is some morphological evidence for a relationship between *A. twista* and Telonemia, this is not especially strong, and the implications of the high number of uniquely shared OGs between the two taxa seen in this study are not reflected in recent phylogenies. An alternative explanation is that one taxon has received a portion of the other's genome via LGT. In eukaryotes this mainly occurs, aside from via plastid acquisition, following the ingestion of genetic material as a result of phagotrophy (Andersson, 2005) – although eukaryote-to-eukaryote LGT is relatively rare (Keeling and Palmer, 2008). Both taxa are eukaryotrophs, however it seems unlikely that one would prey upon the other since both are similar-sized organisms occupying similar trophic levels (Klaveness et al., 2005; Janouškovec et al., 2017) – but note this makes the assumption that both taxa have remained in the same ecological role over the course of their evolution.

The presence of an MIR domain, found in Inositol 1,4,5-triphosphate receptor (IP3R) and ryanodine receptor (RyR) – both of which are calcium release channels (Witcher et al., 1991; Kaftan et al., 1997) – suggests there is some kind of calcium-mediated signal transduction system in place in the two taxa, given the presence of other calcium-activated proteins in their uniquely shared OGs. Calpain domain III and EF-hand motifs are present (the EF-hand is a component of calpain). The EF-hand binds to Ca2+, as do sites in calpain domain II, together activating the molecule (Khorchid and Ikura, 2002). Two OGs with voltage-sensitive calcium channel proteins are also present.

3.10 Evolutionary position of Ancyromonadida and CRuMs

As with *A. twista* and Telonemia, there appears to be a relationship between Ancyromonadida and CRuMs, which was inferred on the basis of the uniquely shared OGs among these two groups. The 383 OGs (350 after subtracting probable bacterial contaminants) shared uniquely between the two taxa is the second-highest of any pair of taxa in the 18-way analysis, and the topological constraint analysis could not be rejected by the AU test.

There is no mention of a specific relationship between these two lineages in the literature, though it is worth noting that CRuMs has only recently been formally recognised (Brown et al., 2018). A *Mantamonas* + Ancyromonadida clade was recovered by Cavalier-Smith et al. (2014), however this was not well supported. The study of Brown et al. (2018) strongly supports CRuMs as the sister lineage to Amorphea, recovering Ancyromonadida as a deeper branch forming a clade with *Malawimonas*. Lax et al. (2018) similarly recover CRuMs + Amorphea, but instead find Ancyromonadida in a clade with Metamonada (though note here only one Metamonada representative is present). Evidently the phylogenetic position of Ancyromonadida is still uncertain.

Morphologically, Ancyromonadida and CRuMs are similar in having a ventral groove – this is absent only in *Mantamonas* (Glücksman et al., 2011). However, as discussed in previous sections this character alone is not indicative of a specific relationship, being found in various different lineages throughout the eToL. Likewise, both taxa possess the split right microtubular root and dorsal root characters both thought to be derived from an ancestral excavate body plan, though neither character is particularly similar in either taxon (Cavalier-Smith and Chao, 2010). Only if the above characters are inferred to have evolved multiple times could they be potential synapomorphies for an Ancyromonadida + CRuMs clade.

Interestingly, in the original description of *Mantamonas*, this genus is assigned to Apusozoa (containing both Apusomonadida and Ancyromonadida) as these taxa have in common highly acronematic cilia and an exclusively gliding lifestyle, though *Mantamonas* was deemed distinct enough to warrant placement in a new family, Mantamonadidae (Glücksman et al., 2011).

Another genus thought to be related to Apusozoa, and more specifically the ancyromonads, is *Micronuclearia*, which shares a similar pellicular structure, and has a similar mode of ingesting bacteria through a depression at the cell surface (Cavalier-Smith et al., 2008). This mode of feeding sounds very much like that of a typical ventral groove-bearing cell, though note the morphology is described as a 'pocket' rather than a groove (Cavalier-Smith et al., 2008). More recently *Micronuclearia* has been inferred to have a specific relationship with *Rigifila ramosa*, with the two genera described as having a body plan distinct from all other protists (Yabuki et al., 2013b). This is somewhat different to the other species within CRuMs. For this reason it was interesting to single out the number of OGs shared uniquely by *R. ramosa* and Ancyromonadida, to compare with the number shared uniquely by *Diphylleia rotans* (the other CRuMs member in this dataset) and Ancyromonadida, with *D. rotans* being a less morphologically divergent species (Yabuki et al., 2013b). There was no significant difference between the

two, however – *D. rotans* shared 267, and *R. ramosa* shared 272.

R. ramosa has been confidently recovered as a CRuMs member in phylogenomic analyses (Brown et al., 2018). The same cannot be said for *Micronuclearia*, which has very little publicly available molecular data, although analysis of rDNA has recovered a *Micronuclearia* + *Rigifila* clade (Yabuki et al., 2013b), and a phylogeny of rDNA plus actin recovered a *Micronuclearia* + *Ancyromonas* clade (Parfrey et al., 2010). More genomic data for *Micronuclearia* are clearly needed to clarify its evolutionary position, however the putative morphological and phylogenetic links to both CRuMs members and Ancyromonadida are intriguing given the relationship between Ancyromonadida and CRuMs indicated by the results in this study. Future phylogenomic analyses including *Micronuclearia* may be key to understanding the specific relationships between it, the CRuMs lineage, and Ancyromonadida.

It is interesting to consider the possibility that LGT may have occurred between Ancyromonadida and CRuMs, and may account for their high number of uniquely shared OGs. This seems highly unlikely to have occurred by an ancyromonad phagocytosing another eukaryote, since they are known to be bacterivorous (Cavalier-Smith et al., 2008). Conversely *D. rotans* has been described as ingesting prey as large as itself (Brugerolle and Patterson, 1990), so in principle this may include members of Ancyromonadida.

It is difficult to say if any of the 383 OGs shared uniquely between the two taxa are attributable to their morphological similarities. The 'typical excavate' characters are not unique to these two lineages, so one would expect OGs related to this cytoskeletal arrangement to contain representatives from the excavate taxa also. This is unless an ancient protein family (or families) diverged to the point where its sequence homology was no longer detectable in other lineages with the same characters (Hodges et al., 2010). Most of the 383 OGs did not return functional annotation in eggNOG. Of the three that were assigned the functional category 'cytoskeleton' (Fig. 6B), only one included details of a particular cytoskeletal function – an OG containing RCC1, i.e. a domain involved in the regulation of mitosis (Dasso, 1993).

In the case of Ancyromonadida in particular, the comparative genomics approach used in this study has proved valuable in highlighting a potential relationship for a group whose placement in conventional phylogenomic analyses has been largely inconsistent. Searching for uniquely shared OGs may not be fruitful in many cases, however it is worth considering, at least for taxa whose phylogenetic placement is uncertain, especially in an era where genomic data are so abundant.

3.11 Missing data

It is worth noting that the OGs analysed in this study are a subset of the total genomic data available in public repositories, which in turn are a subset of all the genomic data that exists in nature. New sequencing efforts will add new sequences that can be included in clustering analyses and will add new sequences to existing OGs, as well as lead to new OGs. Additionally, more taxa will likely be discovered, especially if current sampling biases are addressed (Keeling and Burki, 2019). This has the potential to significantly alter the number of OGs that are unique to a particular set of taxa.

When scanning sequences in eggNOG, attention was drawn to the fact that seed orthologs used in the analysis usually didn't belong to taxa in the same lineage as the query sequences. As a result, those OGs that were unique to pairs of taxa in this dataset would not be considered unique when incorporating this external data, thus reducing the actual number of uniquely shared OGs. This does not significantly impact the main relationships identified in this study – *A. twista* + Telonemia and Ancyromonadida + CRuMs – since in both these cases the majority of OGs simply did not find any matches in eggNOG.

3.12 Future work

Though steps are being made towards understanding the molecular components of the excavate cytoskeleton, it is still not particularly well understood. Future proteomic analyses focused specifically on the microtubular roots and fibres of the excavate taxa may provide answers where current bioinformatics efforts have fallen short. The lack of uniquely shared OGs among the excavates, as well as other eukaryotes in possession of excavate characters, is evidence that there are no straightforward genetic markers mapping genotype to phenotype. Rather, the genetic basis for the excavate cytoskeleton is likely more cryptic, involving either diverged proteins with similar functions in different taxa or proteins conserved throughout most or all of the eukaryotes that fulfil different functions in the excavates. Alternatively, it may simply be that the excavate hypothesis is incorrect – that the excavate cytoskeleton and the proteins that comprise it have multiple independent evolutionary origins.

Armed with a better understanding of how cytoskeletal proteins function in different eukaryotes, the construction of phylogenies for these proteins may allow further insight into the evolution of the eukaryote cytoskeleton. In the case of proteins unique to the individual lineages of excavates, characterisation of the many that are currently unknown in function may also prove

informative.

3.13 Conclusion

The present study's large-scale comparison of genomic data is interesting as it points to potential evolutionary relationships between taxa that hitherto have not been considered, namely that between *A. twista* and Telonemia, and Ancyromonadida and the CRuMs lineage. Though it is difficult to draw concrete conclusions due to the inherent uncertainty surrounding patterns of gene gain, loss and transfer deep in the evolutionary history of eukaryotes, the large number of OGs shared uniquely between the above taxa should certainly encourage future phylogenomic and comparative studies to focus on these relationships. In particular, the addition of *Colponema marisrubri*, *Metromonas simplex* and *Micronuclearia podoventralis* to datasets should help to elucidate the phylogenetic position of these species, as well as the position of *A. twista* and the species within Telonemia, Ancyromonadida and CRuMs.

The lack of uniquely shared OGs among the excavate taxa makes it difficult to draw specific conclusions about what their evolutionary relationships are. However, the absence of evidence does point towards what their relationships, in all likelihood, are *not*. That is, the excavates do not appear to be a monophyletic group (notwithstanding the possibility that the group are a very deeply diverged monophyly whose extant lineages bear limited genomic resemblance to one another). Furthermore, when considered with morphological evidence from the literature (Table 2), it seems reasonable to assert that the excavates may well be an ancient assemblage of protists from which most if not all extant eukaryotes evolved. The strength of this assertion rests on the inferred homology of the various cytoskeletal components, recognised as being 'typical excavate' characters, that are scattered throughout the various eukaryote lineages. If it is confirmed that some of these characters are in fact not homologous, the ancestral excavate argument will be weakened. Otherwise, the LECA may well have been an excavate-like organism, with its excavate relatives living on in a very long period of morphostasis.

Future attempts to root the eToL should make sure to include representatives from Metamonada. These are a crucial piece of the eukaryote puzzle, missing from previous rooted analyses that found the other two excavate lineages, Discoba and Malawimonadidae, branching deeply on either side of the root (Derelle et al., 2015). Such studies may bring us closer to understanding how exactly the excavate taxa fit into the evolutionary history of eukaryotes.

References

Adl, S. M., Simpson, A. G., Lane, C. E., Lukeš, J., Bass, D., Bowser, S. S., Brown, M. W., Burki, F., Dunthorn, M., Hampl, V., Heiss, A., Hoppenrath, M., Lara, E., Gall, L. L., Lynn, D. H., McManus, H., Mitchell, E. A., Mozley-Stanridge, S. E., Parfrey, L. W., Pawlowski, J., Rueckert, S., Shadwick, L., Schoch, C. L., Smirnov, A., and Spiegel, F. W. (2012). The revised classification of eukaryotes. *Journal of Eukaryotic Microbiology*, 59(5):429–514.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402.

Andersson, J. O. (2005). Lateral gene transfer in eukaryotes. *Cellular and Molecular Life Sciences*, 62(11):1182–1197.

Baines, A. J., Bignone, P. A., King, M. D., Maggs, A. M., Bennett, P. M., Pinder, J. C., and Phillips, G. W. (2009). The CKK domain (DUF1781) binds microtubules and defines the CAMSAP/ssp4 family of animal proteins. *Molecular Biology and Evolution*, 26(9):2005–2014.

Baldauf, S. L. (2003). The deep roots of eukaryotes. *Science*, 300(5626):1703–1706.

Baldauf, S. L., Roger, A. J., Wenk-Siefert, I., and Doolittle, W. F. (2000). A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science*, 290(5493):972–977.

Barrett, M. P., Burchmore, R. J., Stich, A., Lazzari, J. O., Frasch, A. C., Cazzulo, J. J., and Krishna, S. (2003). The trypanosomiases. In *Lancet*, volume 362, pages 1469–1480. Elsevier Limited.

Bernard, C., Simpson, A. G., and Patterson, D. J. (2000). Some free-living flagellates (protista) from anoxic habitats. *Ophelia*, 52(2):113–142.

Bernard, C., Simpson, A. G. B., and Patterson, D. J. (1997). An Ultrastructural Study

of a Free-living Retortamonad, Chilomastix cuspidata (Larsen & Patterson, 1990) n. comb.(Retortamonadida, Protista). *European Journal of PROTISTOLOGY*, 33:254–265.

Brown, M. W., Heiss, A. A., Kamikawa, R., Inagaki, Y., Yabuki, A., Tice, A. K., Shiratori, T., Ishida, K. I., Hashimoto, T., Simpson, A. G., and Roger, A. J. (2018). Phylogenomics Places Orphan Protistan Lineages in a Novel Eukaryotic Super-Group. *Genome Biology and Evolution*, 10(2):427–433.

Brown, M. W., Sharpe, S. C., Silberman, J. D., Heiss, A. A., Lang, B. F., Simpson, A. G., and Roger, A. J. (2013). Phylogenomics demonstrates that breviate flagellates are related to opisthokonts and apusomonads. *Proceedings of the Royal Society B: Biological Sciences*, 280(1769).

Brugerolle, G. (2006). Description of a New Freshwater Heterotrophic Flagellate Sulcomonas lacustris Affiliated to the Collodictyonids. Technical report.

Brugerolle, G., Bricheux, G., Philippe, H., and Coffe, G. (2002). Collodictyon triciliatum and Diphylleia rotans (= Aulacomonas submarina) Form a New Family of Flagellates (Collodictyonidae) with Tubular Mitochondrial Cristae that is Phylogenetically Distant from other Flagellate Groups. *Protist*, 153:59–70.

Brugerolle, G. and Patterson, D. J. (1990). A cytological study of Aulacomonas submarina Skuja 1939, a heterotrophic flagellate with a novel ultrastructural identity. *European Journal of Protistology*, 25(3):191–199.

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(1):59–60.

Burki, F. (2014). The eukaryotic tree of life from a global phylogenomic perspective. *Cold Spring Harbor Perspectives in Biology*, 6(5):1–18.

Burki, F., Inagaki, Y., Brate, J., Archibald, J. M., Keeling, P. J., Cavalier-smith, T., Sakaguchi, M., Hashimoto, T., Horak, A., Kumar, S., Klaveness, D., Jakobsen, K. S., Pawlowski, J., and Shalchian-tabrizi, K. (2009). Large-Scale Phylogenomic Analyses Reveal That Two Enigmatic Protist Lineages , Telonemia and Centroheliozoa , Are Related to Photosynthetic Chromalveolates. *Genome Biology and Evolution*, pages 231–238.

Burki, F., Kaplan, M., Tikhonenkov, D. V., Zlatogursky, V., Minh, B. Q., Radaykina, L. V., Smirnov, A., Mylnikov, A. P., and Keeling, P. J. (2016). Untangling the early diversification

of eukaryotes: A phylogenomic study of the evolutionary origins of centrohelida, haptophyta and cryptista. *Proceedings of the Royal Society B: Biological Sciences*, 283(1823).

Burki, F., Okamoto, N., Pombert, J.-F., and Keeling, P. J. (2012). The evolutionary history of haptophytes and cryptophytes : phylogenomic evidence for separate origins. *Proceedings of the Royal Society B: Biological Sciences*, 279(February):2246–2254.

Burki, F., Roger, A. J., Brown, M. W., and Simpson, A. G. (2019). The New Tree of Eukaryotes. *Trends in Ecology and Evolution*, 35(1):43–55.

Burki, F., Shalchian-Tabrizi, K., Minge, M., Skjæveland, Å., Nikolaev, S. I., Jakobsen, K. S., and Pawlowski, J. (2007). Phylogenomics reshuffles the eukaryotic supergroups. *PLoS ONE*, 2(8):1–6.

Carlton, J. M., Hirt, R. P., Silva, J. C., Delcher, A. L., Schatz, M., Zhao, Q., Wortman, J. R., Bidwell, S. L., Cecilia, U., Alsmark, M., Besteiro, S., Sicheritz-Ponten, T., Noel, C. J., Dacks, J. B., Foster, P. G., Simillion, C., Van De Peer, Y., Miranda-Saavedra, D., Barton, G. J., Westrop, G. D., Müller, S., Dessi, D., Fiori, P. L., Ren, Q., Paulsen, I., Zhang, H., Bastida-Corcuera, F. D., Simoes-Barbosa, A., Brown, M. T., Hayes, R. D., Mukherjee, M., Okumura, C. Y., Schneider, R., Smith, A. J., Vanacova, S., Villalvazo, M., Haas, B. J., Pertea, M., Feldblyum, T. V., Utterback, T. R., Shu, C.-L., Osoegawa, K., De Jong, P. J., Hrdy, I., Horvathova, L., Zubacova, Z., Dolezal, P., Malik, S.-B., Logsdon, J. M., Henze, K., Gupta, A., Wang, C. C., Dunne, R. L., Upcroft, J. A., Upcroft, P., White, O., Salzberg, S. L., Tang, P., Chiu, C.-H., Lee, Y.-S., Martin Embley, T., Coombs, G. H., Mottram, J. C., Tachezy, J., Fraser-Liggett, C. M., and Johnson, P. J. (2007). Draft Genome Sequence of the Sexually Transmitted Pathogen Trichomonas vaginalis. 315:207–212.

Carvalho-Santos, Z., Azimzadeh, J., Pereira-Leal, J. B., and Bettencourt-Dias, M. (2011). Tracing the origins of centrioles, cilia, and flagella. *Journal of Cell Biology*, 194(2):165–175.

Cavalier-Smith, T. (1993). Kingdom Protozoa and Its 18 Phyla. *MICROBIOLOGICAL REVIEWS*, 57(4):953–994.

Cavalier-Smith, T. (1999). Principles of protein and lipid targeting in secondary symbiogenesis: Euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. *Journal of Eukaryotic Microbiology*, 46(4):347–366.

Cavalier-Smith, T. (2002). The phagotrophic origin of eukaryotes and phylogenetic classification on protozoa. *International Journal of Systematic and Evolutionary Microbiology*, 52(2):297–354.

Cavalier-smith, T. (2003). Protist phylogeny and the high-level classification of Protozoa. *European Journal of PROTISTOLOGY*, 39:338–348.

Cavalier-Smith, T. (2003). The excavate protozoan phyla Metamonada Grassé emend. (Anaeromonadea, Parabasalia, Carpediemonas, Eopharyngia) and Loukozoa emend. (Jakobea, Malawimonas): Their evolutionary affinities and new higher taxa. *International Journal of Systematic and Evolutionary Microbiology*, 53(6):1741–1758.

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

Cavalier-Smith, T. (2018). *Kingdom Chromista and its eight phyla: a new synthesis emphasising periplastid protein targeting, cytoskeletal and periplastid evolution, and ancient divergences*, volume 255. Protoplasma.

Cavalier-Smith, T. and Chao, E. E. (2010). Phylogeny and Evolution of Apusomonadida (Protozoa: Apusozoa): New Genera and Species. *Protist*, 161(4):549–576.

Cavalier-Smith, T., Chao, E. E., and Lewis, R. (2015). Multiple origins of Heliozoa from flagellate ancestors: New cryptist subphylum Corbihelia, superclass Corbistoma, and monophyly of Haptista, Cryptista, Hacrobia and Chromista. *Molecular Phylogenetics and Evolution*, 93:331–362.

Cavalier-Smith, T., Chao, E. E., and Lewis, R. (2018). Multigene phylogeny and cell evolution of chromist infrakingdom Rhizaria: contrasting cell organisation of sister phyla Cercozoa and Retaria. *Protoplasma*, 255(5):1517–1574.

Cavalier-Smith, T., Chao, E. E., Snell, E. A., Berney, C., Fiore-Donno, A. M., and Lewis, R. (2014). Multigene eukaryote phylogeny reveals the likely protozoan ancestors of opisthokonts (animals, fungi, choanozoans) and Amoebozoa. *Molecular Phylogenetics and Evolution*, 81(August):71–85.

Cavalier-Smith, T., Chao, E. E., Stechmann, A., Oates, B., and Nikolaev, S. (2008). Planomonadida ord. nov. (Apusozoa): Ultrastructural Affinity with Micronuclearia podoventralis and Deep Divergences within Planomonas gen. nov. *Protist*, 159(4):535–562.

Clark, S. W. and Rose, M. D. (2006). Arp10p Is a Pointed-End-associated Component of Yeast Dynactin. *Molecular Biology of the Cell*, 17(February):738–748.

Dacks, J. B., Silberman, J. D., Simpson, A. G. B., Moriya, S., Kudo, T., Ohkuma, M., and

Redfield, R. J. (2001). Oxymonads Are Closely Related to the Excavate Taxon Trimastix. *Mol. Biol. Evol*, 18(6):1034–1044.

Dasso, M. (1993). RCC1 in the cell cycle: the regulator of chromosome condensation takes on new roles. *Trends in Biochemical Sciences*, 18(3):96–101.

Dawson, S. C. (2010). An insider's guide to the microtubule cytoskeleton of Giardia. *Cellular Microbiology*, 12(5):588–598.

Dawson, S. C. and Paredez, A. R. (2013). Alternative cytoskeletal landscapes: Cytoskeletal novelty and evolution in basal excavate protists. *Current Opinion in Cell Biology*, 25(1):134–141.

Derelle, R. and Lang, B. F. (2012). Rooting the eukaryotic tree with mitochondrial and bacterial proteins. *Molecular Biology and Evolution*, 29(4):1277–1289.

Derelle, R., Torruella, G., Klimeš, V., Brinkmann, H., Kim, E., Vlček, Č., Lang, B. F., and Eliáš, M. (2015). Bacterial proteins pinpoint a single eukaryotic root. *Proceedings of the National Academy of Sciences of the United States of America*, 112(7):E693–E699.

Desser, S. S., Hong, H., Siddall, M. E., and Barta, J. R. (1993). An ultrastructural study of Brugerolleia algonquinensis gen. nov., sp. nov. (Diplomonadina; Diplomonadida), a flagellate parasite in the blood of frogs from Ontario, Canada. *European Journal of Protistology*, 29(1):72–80.

Edgcomb, V. P., Roger, A. J., Simpson, A. G. B., Kysela, D. T., and Sogin, M. L. (2001). Evolutionary Relationships Among " Jakobid " Flagellates as Indicated by alpha-and-beta-tubulin phylogenies. pages 514–522.

Embley, T. M. and Hirt, R. P. (1998). Early branching eukaryotes? *Current Opinion in Genetics and Development*, 8:624–629.

Emms, D. M. and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, 16(1):1–14.

Escobar-Zepeda, A., De León, A. V. P., and Sanchez-Flores, A. (2015). The road to metagenomics: From microbiology to DNA sequencing technologies and bioinformatics. *Frontiers in Genetics*, 6(DEC):1–15.

Fritz-Laylin, L. K., Prochnik, S. E., Ginger, M. L., Dacks, J. B., Carpenter, M. L., Field, M. C., Kuo, A., Paredez, A., Chapman, J., Pham, J., Shu, S., Neupane, R., Cipriano, M., Mancuso,

J., Tu, H., Salamov, A., Lindquist, E., Shapiro, H., Lucas, S., Grigoriev, I. V., Cande, W. Z., Fulton, C., Rokhsar, D. S., and Dawson, S. C. (2010). The Genome of Naegleria gruberi Illuminates Early Eukaryotic Versatility. *Cell*, 140(5):631–642.

Galperin, M. Y., Makarova, K. S., Wolf, Y. I., and Koonin, E. V. (2015). Expanded Microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Research*, 43(D1):D261–D269.

Glücksman, E., Snell, E. A., Berney, C., Chao, E. E., Bass, D., and Cavalier-Smith, T. (2011). The Novel Marine Gliding Zooflagellate Genus Mantamonas (Mantamonadida ord. n.: Apusozoa). *Protist*, 162(2):207–221.

Hampl, V., Horner, D. S., Dyal, P., Kulda, J., Flegr, J., Foster, P. G., and Embley, T. M. (2005). Inference of the phylogenetic position of oxymonads based on nine genes: Support for metamonada and Excavata. *Molecular Biology and Evolution*, 22(12):2508–2518.

Hampl, V., Hug, L., Leigh, J. W., Dacks, J. B., Lang, B. F., Simpson, A. G., and Roger, A. J. (2009). Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". *Proceedings of the National Academy of Sciences of the United States of America*, 106(10):3859–3864.

Hampl, V., Silberman, J. D., Stechmann, A., Diaz-Triviño, S., Johnson, P. J., and Roger, A. J. (2008). Genetic evidence for a mitochondriate ancestry in the 'amitochondriate' flagellate Trimastix pyriformis. *PLoS ONE*, (1).

He, D., Fiz-Palacios, O., Fu, C. J., Tsai, C. C., and Baldauf, S. L. (2014). An alternative root for the eukaryote tree of life. *Current Biology*, 24(4):465–470.

Heiss, A. A., Kolisko, M., Ekelund, F., Brown, M. W., Roger, A. J., and Simpson, A. G. (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).

Heiss, A. A., Walker, G., and Simpson, A. G. (2011). The ultrastructure of Ancyromonas; a eukaryote without supergroup affinities. *Protist*, 162(3):373–393.

Hendershott, M. C. and Vale, R. D. (2014). Regulation of microtubule minus-end dynamics by CAMSAPs and Patronin. *Proceedings of the National Academy of Sciences of the United States of America*, 111(16):5860–5865.

Hodges, M. E., Scheumann, N., Wickstead, B., Langdale, J. A., and Gull, K. (2010). Recon-

structing the evolutionary history of the centriole from protein components. *Journal of Cell Science*, 123(9):1407–1413.

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., Von Mering, C., and Bork, P. (2017). Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Molecular Biology and Evolution*, 34(8):2115–2122.

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., Von Mering, C., and Bork, P. (2019). EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, 47(D1):D309–D314.

Janouškovec, J., Tikhonenkov, D. V., Burki, F., Howe, A. T., Rohwer, F. L., Mylnikov, A. P., and Keeling, P. J. (2017). A New Lineage of Eukaryotes Illuminates Early Mitochondrial Genome Reduction. *Current Biology*, 27(23):3717–3724.e5.

Kaftan, E. J., Ehrlich, B. E., and Watras, J. (1997). the Dynamic Range of InsP 3 Receptor-dependent Calcium Signaling. 110(November).

Karnkowska, A., Vacek, V., Zubáčová, Z., Treitli, S. C., Petrželková, R., Eme, L., Novák, L., Žárský, V., Barlow, L. D., Herman, E. K., Soukal, P., Hroudová, M., Doležal, P., Stairs, C. W., Roger, A. J., Eliáš, M., Dacks, J. B., Vlček, Č., and Hampl, V. (2016). A eukaryote without a mitochondrial organelle. *Current Biology*, 26(10):1274–1284.

Karpov, S. A. (2007). The flagellar apparatus structure of Apusomonas proboscidea and apusomonad relationships. *Protistology*, 5(2/3):146–155.

Katz, L. A. and Grant, J. R. (2015). Taxon-rich phylogenomic analyses resolve the eukaryotic tree of life and reveal the power of subsampling by sites. *Systematic Biology*, 64(3):406–415.

Keeling, P. J., Burger, G., Durnford, D. G., Lang, B. F., Lee, R. W., Pearlman, R. E., Roger, A. J., and Gray, M. W. (2005). The tree of eukaryotes. *Trends in Ecology and Evolution*, 20(12):670–676.

Keeling, P. J. and Burki, F. (2019). Progress towards the Tree of Eukaryotes. *Current Biology*.

Keeling, P. J., Burki, F., Wilcox, H. M., Allam, B., Allen, E. E., Amaral-Zettler, L. A., Armbrust, E. V., Archibald, J. M., Bharti, A. K., Bell, C. J., Beszteri, B., Bidle, K. D., Cameron, C. T., Campbell, L., Caron, D. A., Cattolico, R. A., Collier, J. L., Coyne, K., Davy, S. K., Deschamps, P., Dyhrman, S. T., Edvardsen, B., Gates, R. D., Gobler, C. J., Greenwood, S. J., Guida, S. M., Jacobi, J. L., Jakobsen, K. S., James, E. R., Jenkins, B., John, U., Johnson, M. D., Juhl,

A. R., Kamp, A., Katz, L. A., Kiene, R., Kudryavtsev, A., Leander, B. S., Lin, S., Lovejoy, C., Lynn, D., Marchetti, A., McManus, G., Nedelcu, A. M., Menden-Deuer, S., Miceli, C., Mock, T., Montresor, M., Moran, M. A., Murray, S., Nadathur, G., Nagai, S., Ngam, P. B., Palenik, B., Pawlowski, J., Petroni, G., Piganeau, G., Posewitz, M. C., Rengefors, K., Romano, G., Rumpho, M. E., Rynearson, T., Schilling, K. B., Schroeder, D. C., Simpson, A. G., Slamovits, C. H., Smith, D. R., Smith, G. J., Smith, S. R., Sosik, H. M., Stief, P., Theriot, E., Twary, S. N., Umale, P. E., Vaulot, D., Wawrik, B., Wheeler, G. L., Wilson, W. H., Xu, Y., Zingone, A., and Worden, A. Z. (2014). The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLoS Biology*, 12(6).

Keeling, P. J. and Palmer, J. D. (2008). Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics*, 9(8):605–618.

Khorchid, A. and Ikura, M. (2002). How calpain is activated by calcium. *Nature Structural Biology*, 9(4):239–241.

Kishino, H. and Hasegawa, M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence .... *Journal of Molecular Evolution*, 29:170–179.

Klaveness, D., Shalchian-tabrizi, K., Thomsen, H. A., Eikrem, W., and Jakobsen, K. S. (2005). Telonema antarcticum sp . nov ., a common marine phagotrophic flagellate. pages 2595–2604.

Kugrens, P., Lee, R. E., and Corliss, J. O. (1994). Ultrastructure , biogenesis , and functions of extrusive organelles in selected non-ciliate protists. *Protoplasma*, 181:164–190.

Lax, G., Eglit, Y., Eme, L., Bertrand, E. M., Roger, A. J., and Simpson, A. G. (2018). Hemimastigophora is a novel supra-kingdom-level lineage of eukaryotes. *Nature*, 564(7736):410–414.

Leipe, D. D., Gunderson, J. H., Nerad, T. A., and Sogin, M. L. (1993). Small subunit ribosomal RNA of Hexamita inflata and the quest for the first branch in the eukaryotic tree. Technical report.

Letunic, I. and Bork, P. (2019). Interactive Tree of Life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Research*, 47(W1):256–259.

McInally, S. G. and Dawson, S. C. (2016). Eight unique basal bodies in the multi-flagellated diplomonad Giardia lamblia.

Minge, M. A., Silberman, J. D., Orr, R. J., Cavalier-Smith, T., Shalchian-Tabrizi, K., Burki, F., Skjæveland, Å., and Jakobsen, K. S. (2009). Evolutionary position of breviate amoebae and the primary eukaryote divergence. *Proceedings of the Royal Society B: Biological Sciences*, 276(1657):597–604.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., Lanfear, R., and Teeling, E. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, 37(5):1530–1534.

Mitchell, A. L., Attwood, T. K., Babbitt, P. C., Blum, M., Bork, P., Bridge, A., Brown, S. D., Chang, H. Y., El-Gebali, S., Fraser, M. I., Gough, J., Haft, D. R., Huang, H., Letunic, I., Lopez, R., Luciani, A., Madeira, F., Marchler-Bauer, A., Mi, H., Natale, D. A., Necci, M., Nuka, G., Orengo, C., Pandurangan, A. P., Paysan-Lafosse, T., Pesseat, S., Potter, S. C., Qureshi, M. A., Rawlings, N. D., Redaschi, N., Richardson, L. J., Rivoire, C., Salazar, G. A., Sangrador-Vegas, A., Sigrist, C. J., Sillitoe, I., Sutton, G. G., Thanki, N., Thomas, P. D., Tosatto, S. C., Yong, S. Y., and Finn, R. D. (2019). InterPro in 2019: Improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Research*, 47(D1):D351–D360.

Moestrup, O. (2000). The flagellate cytoskeleton: introduction of a general terminology for microtubular flagellar roots in protists. In Leadbeater, S. C. B. and Green, J. C., editors, *The Flagellates: Unity, Diversity and Evolution*, chapter 4, pages 69–94.

Nabais, C., Peneda, C., and Bettencourt-Dias, M. (2020). Evolution of centriole assembly. *Current Biology*, 30(10):R494–R502.

O'Kelly, C. J. (1993). The Jakobid Flagellates: Structural Features of Jakoba, Reclinomonas and Histiona and Implications for the Early Diversification of Eukaryotes. *Journal of Eukaryotic Microbiology*, 40:627–636.

O'Kelly, C. J. (1997). Ultrastructure of Trophozoites, Zoospores and Cysts of Reclinomonas americana Flavin & Nerad, 1993 (Protista incertae sedis: Histionidae). Technical report.

O'Kelly, C. J., Farmer, M. A., and Nerad, T. A. (1999). Ultrastructure of Trimastix pyriformis (Klebs) Bernard et al.: Similarities of Trimastix species with retortamonad and jakobid flagellates. *Protist*, 150(2):149–162.

O'Kelly, C. J. and Nerad, T. A. (1999). Malawimonas jakobiformis n. gen., n. sp. (Malawimonadidae n. fam.): A Jakoba-like heterotrophic nanoflagellate with discoidal mitochondrial cristae. *Journal of Eukaryotic Microbiology*, 46(5):522–531.

Ortega, Y. R. and Adam, R. D. (1997). Giardia: Overview and Update. *Clinical Infectious Diseases*, 25(3):545–549.

Paps, J., Medina-Chacón, L. A., Marshall, W., Suga, H., and 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(1):2–12.

Parfrey, L. W., Barbero, E., Lasser, E., Dunthorn, M., Bhattacharya, D., Patterson, D. J., and Katz, L. A. (2006). Evaluating support for the current classification of eukaryotic diversity. *PLoS Genetics*, 2(12):2062–2073.

Parfrey, L. W., Grant, J., Tekle, Y. I., Lasek-Nesselquist, E., Morrison, H. G., Sogin, M. L., Patterson, D. J., and Katz, L. A. (2010). Broadly sampled multigene analyses yield a well-resolved eukaryotic tree of life. *Systematic Biology*, 59(5):518–533.

Patterson, D. J. (1990). Jakoba Libera (Ruinen, 1938), A Heterotrophic Flagellate From Deep Oceanic Sediments. *Journal of the Marine Biological Association of the United Kingdom*, 70:381–393.

Pawlowski, J. (2008). The twilight of Sarcodina: a molecular perspective on the polyphyletic origin of amoeboid protists. *Protistology*, 5(4):281–302.

Preisner, H., Karin, E. L., Poschmann, G., Stühler, K., Pupko, T., and Gould, S. B. (2016). The Cytoskeleton of Parabasalian Parasites Comprises Proteins that Share Properties Common to Intermediate Filament Proteins. *Protist*, 167(6):526–543.

Price, D. C., Goodenough, U. W., Roth, R., Lee, J.-h., Kariyawasam, T., Mutwil, M., Ferrari, C., Facchinelli, F., Ball, S. G., Cenci, U., Chan, C. X., Wagner, N. E., Yoon, H. S., and Weber, A. P. M. (2019). Analysis of an improved Cyanophora paradoxa genome assembly. *DNA Research*, 26(4):287–299.

Raaijmakers, J. A., Tanenbaum, M. E., and Medema, R. (2013). Systematic dissection of dynein regulators in mitosis. *The Journal of Cell Biology*, 201(2):201–215.

Rodríguez-Ezpeleta, N., Brinkmann, H., Burger, G., Roger, A. J., Gray, M. W., Philippe, H., and Lang, B. F. (2007). Toward Resolving the Eukaryotic Tree: The Phylogenetic Positions of Jakobids and Cercozoans. *Current Biology*, 17(16):1420–1425.

Roger, A. and Simpson, A. G. B. (2009). Evolution: Revisiting the Root of the Eukaryote Tree. *Current Biology*, 19(4).

Roger, A. J. (1999). Reconstructing Early Events in Eukaryotic Evolution. Technical report.

Roger, A. J., Muñoz-Gómez, S. A., and Kamikawa, R. (2017). The Origin and Diversification of Mitochondria. *Current Biology*, 27(21):R1177–R1192.

Schroer, T. A. (2004). Dynactin. *Annual Review of Cell and Developmental Biology*, 20:759–779.

Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. *Systematic Biology*, 51(3):492–508.

Shimodaira, H. and Hasegawa, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution*, 16(8):1114–1116.

Silberman, J. D., Simpson, A. G. B., Kulda, J., Cepicka, I., Hampl, V., Johnson, P. J., and Roger, A. J. (2002). Retortamonad Flagellates are Closely Related to Diplomonads-Implications for the History of Mitochondrial Function in Eukaryote Evolution. *Mol. Biol. Evol*, 19(5):777–786.

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

Simpson, A. G. (2003). Cytoskeletal organization, phylogenetic affinities and systematics in the contentious taxon Excavata (Eukaryota). *International Journal of Systematic and Evolutionary Microbiology*, 53(6):1759–1777.

Simpson, A. G., Bernard, C., and Patterson, D. J. (2000). The ultrastructure of Trimastix marina kent, 1880 (Eukaryota), an excavate flagellate. *European Journal of Protistology*, 36(3):229–251.

Simpson, A. G., Inagaki, Y., and Roger, A. J. (2006a). Comprehensive multigene phylogenies of excavate protists reveal the evolutionary positions of ”primitive” eukaryotes. *Molecular Biology and Evolution*, 23(3):615–625.

Simpson, A. G. and Patterson, D. J. (1999). The ultrastructure of Carpediemonas membranifera (Eukaryota) with reference to the ’excavate hypothesis’. *European Journal of Protistology*, 35(4):353–370.

Simpson, A. G. and Patterson, D. J. (2001). On core jakobids and excavate taxa: The ultrastructure of Jakoba incarcerata. *Journal of Eukaryotic Microbiology*, 48(4):480–492.

Simpson, A. G., Radek, R., Dacks, J. B., and O'Kelly, C. J. (2002a). How oxymonads lost their groove: An ultrastructural comparison of Monocercomonoides and excavate taxa. *Journal of Eukaryotic Microbiology*, 49(3):239–248.

Simpson, A. G., Roger, A. J., Silberman, J. D., Leipe, D. D., Edgcomb, V. P., Jermiin, L. S., Patterson, D. J., and Sogin, M. L. (2002b). Evolutionary history of "early-diverging" eukaryotes: The excavate taxon Carpediemonas is a close relative of Giardia. *Molecular Biology and Evolution*, 19(10):1782–1791.

Simpson, A. G., Stevens, J. R., and Lukeš, J. (2006b). The evolution and diversity of kinetoplastid flagellates. *Trends in Parasitology*, 22(4):168–174.

Sogin, M. L., Gunderson, J. H., Elwood, H. J., Alonso, R. A., and Peattie, D. A. (1989). Phylogenetic meaning of the kingdom concept: an unusual ribosomal RNA from Giardia lamblia. *Science*, 243:75–77.

Stechmann, A. and Cavalier-Smith, T. (2002). Rooting the eukaryote tree by using a derived gene fusion. *Science*, 297(5578):89–91.

Stechmann, A. and Cavalier-Smith, T. (2003). The root of the eukaryote tree pinpointed. *Current Biology*, 13(17).

Strassert, J. F., Jamy, M., Mylnikov, A. P., Tikhonenkov, D. V., and Burki, F. (2019). New phylogenomic analysis of the enigmatic phylum Telonemia further resolves the eukaryote tree of life. *Molecular Biology and Evolution*, 36(4):757–765.

Tikhonenkov, D. V., Janouškovec, J., Mylnikov, A. P., Mikhailov, K. V., Simdyanov, T. G., Aleoshin, V. V., and Keeling, P. J. (2014). Description of Colponema vietnamica sp.n. and Acavomonas peruviana n. gen. n. sp., Two New Alveolate Phyla (Colponemidia nom. nov. and Acavomonidia nom.nov.) and Their Contributions to Reconstructing the Ancestral State of Alveolates and Eukaryotes. *PLoS ONE*, 9(4).

Tikhonenkov, D. V., Strassert, J. F., Janouškovec, J., Mylnikov, A. P., Aleoshin, V. V., Burki, F., and Keeling, P. J. (2020). Predatory colponemids are the sister group to all other alveolates. *Molecular Phylogenetics and Evolution*, 149(February).

Turmel, M., Gagnon, M. C., O'Kelly, C. J., Otis, C., and Lemieux, C. (2009). The chloroplast genomes of the green algae Pyramimonas, Monomastix, and Pycnococcus shed new light on the evolutionary history of prasinophytes and the origin of the secondary chloroplasts of euglenids. *Molecular Biology and Evolution*, 26(3):631–648.

Vesteg, M., Hadariová, L., Horváth, A., Estraño, C. E., Schwartzbach, S. D., and Krajčovič, J. (2019). Comparative molecular cell biology of phototrophic euglenids and parasitic trypanosomatids sheds light on the ancestor of Euglenozoa. *Biological Reviews*, 1704:1701–1721.

Viscogliosi, E. and Brugerolle, G. (1994). Striated Fibers in Trichomonads: Costa Proteins Represent a New Class of Proteins Forming Striated Roots. *Cell Motility and the Cytoskeleton*, 29:82–93.

Weerakoon, N. D., Harper, J. D., Simpson, A. G., and Patterson, D. J. (1999). Centrin in the groove: Immunolocalisation of centrin and microtubules in the putatively primitive protist Chilomastix cuspidata (Retortamonadida). *Protoplasma*, 210(1-2):75–84.

Weisman, C. M., Murray, A. W., and Eddy, S. R. (2020). Many but not all lineage-specific genes can be explained by homology detection failure. *bioRxiv 2020.02.27.968420 [Preprint]*, pages 1–29.

Wickstead, B. and Gull, K. (2011). The evolution of the cytoskeleton. *Journal of Cell Biology*, 194(4):513–525.

Witcher, D. R., Kovacs, R. J., Schulman, H., Cefali, D. C., and Jones, L. R. (1991). Unique phosphorylation site on the cardiac ryanodine receptor regulates calcium channel activity. *Journal of Biological Chemistry*, 266(17):11144–11152.

Yabuki, A., Eikrem, W., Takishita, K., and Patterson, D. J. (2013a). Fine Structure of Telonema subtilis Griessmann, 1913: A Flagellate with a Unique Cytoskeletal Structure Among Eukaryotes. *Protist*, 164(4):556–569.

Yabuki, A., Ishida, K. I., and Cavalier-Smith, T. (2013b). Rigifila ramosa n. gen., n. sp., a Filose Apusozoan with a Distinctive Pellicle, is Related to Micronuclearia. *Protist*, 164(1):75–88.

Yabuki, A., Kamikawa, R., Ishikawa, S. A., Kolisko, M., Kim, E., Tanabe, A. S., Kume, K., Ishida, K. I., and Inagki, Y. (2014). Palpitomonas bilix represents a basal cryptist lineage: Insight into the character evolution in Cryptista. *Scientific Reports*, 4:1–6.

Yubuki, N. and Leander, B. S. (2013). Evolution of microtubule organizing centers across the tree of eukaryotes. *Plant Journal*, 75(2):230–244.

Zhao, S., Burki, F., Brte, J., Keeling, P. J., Klaveness, D., and Shalchian-Tabrizi, K. (2012). Collodictyon-an ancient lineage in the tree of eukaryotes. *Molecular Biology and Evolution*, 29(6):1557–1568.