Understanding the evolution of the membrane trafficking system in diverse eukaryotes through comparative genomics and transcriptomics

by

Emily Katherine Herman

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Cell Biology

University of Alberta

Abstract

Single-celled organisms represent the majority of eukaryotic diversity. Recent advances in sequencing technologies have been critical for understanding the evolutionary biology and cell biology of microbial eukaryotes. Comparative genomic analyses have shown that many genes that underlie fundamental eukaryotic features (e.g. membrane trafficking, cytoskeleton) are conserved across the diversity of eukaryotes, suggesting that they have also maintained a similar function. However, many microbial eukaryotes have specialized lifestyles or behaviours, the evolutionary pressures of which may be observed in changes to gene content in a lineage; in either gene family expansion, divergence, or loss. Building on analyses of gene presence and absence, gene expression changes in relation to a specific cellular behaviour gives even more insight into the underlying cell biology of that process.

The focus of this thesis is the membrane trafficking system, specifically the cellular machinery that underlies intracellular transport, endocytosis, and exocytosis. In the first Results chapter, comparative genomics is used to identify membrane trafficking components in three related organisms, one of which is free-living, while the other two are gut-associated endobionts and/or parasites. The purpose was to determine whether host-association contributes to sculpting of the trafficking system, as is the case in other eukaryotic parasites. In the second Results chapter, comparative genomics and transcriptomics are used to study how membrane trafficking underlies the process of encystation in the gut pathogens *Entamoeba invadens* and *E. histolytica*, which is critical for pathogenesis. The third results chapter looks at the biology of a unique behaviour in the haptophyte lineage: the secretion of large organic or calcium carbonate scales. Again, both comparative genomics and transcriptomics are used to understand how the membrane trafficking system contributes to this extensive secretory process.

The last Results chapter takes a whole-genome approach to understanding pathogenesis in the free-living neuropathogenic amoeba *Naegleria fowleri*, as compared with its harmless relative, *Naegleria gruberi*. The purpose of comparative genomics and transcriptomics of *N. fowleri* and *N. gruberi* was primarily to identify pathogenicity factors. Although no single factor was identified that fully explains the difference in pathogenicity between the two *Naegleria* spp., two major outcomes were achieved. First, this analysis has generated a comprehensive look at the cell biology of *N. fowleri* during host infection. Secondly, it has produced a list of dozens of potential pathogenicity factors that can now be experimentally tested. An example of how *in silico* analyses can support functional work concludes this chapter, where evidence of a Golgi body in *N. gruberi* is shown for the first time.

These –omics analyses have contributed significantly to understanding the biology of these lineages. They highlight patterns of retention, loss, and expansion of membrane trafficking machinery that may be related to unusual trafficking pathways or even novel organelles. They also allowed for comparisons of gene complement and expression between different lineages that have similar lifestyles, for example gut-associated parasites or endobionts (*Entamoeba* spp. and *Blastocystis* sp., *Proteromonas lacertae*), or organisms with a heavy secretory load (haptophytes and *Entamoeba* sp.). Common to all three transcriptomic analyses is the finding that transcriptional responses are complex, often involving differential regulation of paralogous genes.

The data presented here have paved the way for future functional work in microbial eukaryotes, improving our depth of knowledge of membrane trafficking function in eukaryotes, and allowing us to fully appreciate unique cell biology in ecologically and medically relevant organisms.

Preface

(Mandatory due to collaborative work)

The works presented in this thesis are the products of several research collaborations.

Comparative genomics of the membrane trafficking system of three *Blastocystis* strains, found in Chapter 3 of this thesis, is the result of a collaboration with Andrew Roger and his lab at Dalhousie University. It has been published in Gentekaki E, Curtis BA, Stairs CW, Klimeš V, Eliáš M, *et al.* (2017) Extreme genome diversity in the hyper-prevalent parasitic eukaryote *Blastocystis*. PLOS Biology 15(9): e2003769. E. Gentekaki, B. Curtis, C. G. Clark, and A. Roger were responsible for conceptualization and administration of the *Blastocystis* sp. genome project. The membrane trafficking system complement was analysed and described for the genome paper by Emily Herman and Alexander Schlacht, under the supervision of Joel Dacks. Other authors were responsible for other aspects of the genome project. The *Proteromonas lacertae* and *Cafeteria roenbergensis* data presented in Chapter 3 are the result of an ongoing collaboration with Andrew Jackson and his lab at the University of Liverpool. E. Herman performed all comparative genomic analyses of membrane trafficking and autophagy machinery in the genomes of *P. lacertae* and *C. roenbergensis*.

The work performed in Chapter 4 is the result of a collaboration with Mark van der Giezen and his lab at the University of Exeter. It has been published as Herman E, Siegesmund MA, Bottery MJ, van Aerle R, Shather MM, Caler E, Dacks JB, and van der Giezen M. (2017). Membrane trafficking modulation during *Entamoeba* encystation. Scientific Reports 7:12854. M. van der Giezen and J. Dacks designed and supervised experiments; Maria Siegesmund and Mohammed Shather performed laboratory experiments; E. Herman, M. Siegesmund, Michael Bottery, Ronny van Aerle, and Elisabet Caler contributed to data analysis; E. Herman, M. Siegesmund, and M. Bottery organized, designed, and wrote the paper; and E. Herman, J. Dacks, and M. van der Giezen critically revised the manuscript. Specifically, E. Herman analysed the membrane trafficking system complement in *E. invadens* and *E. histolytica*, interpreted gene expression patterns for the trafficking complement, and helped write and revise the manuscript.

Analysis of the membrane trafficking system of the haptophytes in Chapter 5 is part of a long-standing collaboration with Betsy Read and her lab at California State University San Marcos. The original analysis of the membrane trafficking machinery in *Emiliania huxleyi* was published as part of a genome project: Read BA, Kegel J, Klute MJ, Kuo A, Lefebvre SC, Maumus F, Mayer C, Miller J, Monier A, Salamov A, Young J, Aguilar M, Claverie JM, Frickenhaus S, Gonzalez K, Herman EK, Lin YC, Napier J, Ogata H, Sarno AF, Shmutz J, Schroeder D, de Vargas C, Verret F, von Dassow P, Valentin K, Van de Peer Y, Wheeler G; Emiliania huxleyi Annotation Consortium, Dacks JB, Delwiche CF, Dyhrman ST, Glöckner G, John U, Richards T, Worden AZ, Zhang X, Grigoriev IV. (2013) Pan genome of the phytoplankton Emiliania underpins its global distribution. Nature 499:209-213. B. Read coordinated the genome project and I. Grigoriev coordinated genome sequencing and analysis at the US DOE Joint Genome Institute. E. Herman and Mary Klute analysed the membrane trafficking system machinery and described the results of this analysis for the genome paper, under the supervision of J. Dacks. Other authors were responsible for other aspects of the genome project.

Part of the comparative work in Chapter 5 with the haptophytes *Gephyrocapsa oceanica* and *Isochrysis galbana* specifically regarding adaptor protein evolution is published as Lee LJY, Klute MK, Herman EK, Read B, and Dacks JB. (2015) Losses, Expansions, and Novel Subunit Discovery of Adaptor Protein Complexes in Haptophyte Algae. Protist 166:585-597. L. Lee and M. Klute performed comparative genomic analyses, B. Read generated sequence data and performed qPCR experiments. J. Dacks and B. Read supervised the project. E. Herman supervised Laura Lee and aided in data analysis and interpretation. As the corresponding author, E. Herman was heavily involved in manuscript writing and editing along with J. Dacks and B. Read, and corresponded with the journal during the manuscript review and publication process.

Additional analysis of the membrane trafficking machinery of the haptophytes (including *Chrysochromulina tobin*), and gene expression under biomineralizing conditions, is part of an ongoing collaboration with B. Read and Xiaoyu Zhang. B. Read and X. Zhang produced genomic and transcriptomic data associated with *E. huxleyi*, *G. oceanica* and *I. galbana*. Analysis of membrane trafficking complement was performed by L. Lee and E. Herman, and analysis of biomineralization transcriptomic data specific to the membrane trafficking system

was performed by E. Herman. Elisabeth Richardson and E. Herman performed a comparative genomic analysis of membrane trafficking machinery in the publicly available genome of *C. tobin*.

The data presented in Chapter 6 is the result of several collaborations. The *N. fowleri* V212 genome sequence and transcriptomic analysis was the result of a collaboration between our lab, Charles Chiu and his lab at the University of California San Francisco, Francine Marciano-Cabral and her lab at Virginia Commonwealth University, and Govinda Visvesvara at the Centers for Disease Control and Prevention. Part of the work from this collaboration has been published as Herman EK, Greninger AL, Visvesvara, GS, Marciano-Cabral F, Dacks JB, and Chiu CY. (2013) The mitochondrial genome and a 60-kb nuclear DNA segment from *Naegleria fowleri*, the causative agent of primary amoebic meningoencephalitis. Journal of Eukaryotic Microbiology 60:179-91. F. Marciano-Cabral, C. Chiu, and Alex Greninger were involved in organism culturing and genome data generation. J. Dacks, C. Chiu, F. Marciano-Cabral, and G. Visvesvara supervised the project. A. Greninger performed genome assembly and PCR. E. Herman analysed the resulting mitochondrial sequence and a nuclear fragment, and wrote and revised the manuscript with the help of the other authors.

The data in Chapter 6 that includes analysis of the two additional *N. fowleri* strains, ATCC 30863 and 986, are the result of ongoing collaborations with Norbert Mueller, Matthias Wittwer, and Denise Zysset-Burri (sequenced and assembled the genome of strain 30863) at the Labor Spiez and University of Bern; and with Geoffrey Puzon and Tom Walsh (sequenced and assembled the genome of strain 986) at the Commonwealth Scientific and Industrial Research Organisation. All bioinformatic analysis of the three *N. fowleri* genomes and the genome of *N. gruberi* presented in this thesis was performed by E. Herman, with the exception of part of the protease family analysis, which was performed by Inmaculada Ramírez-Macías.

The data in Chapter 6 regarding the visualization of the Golgi body in *Naegleria gruberi* is part of an ongoing collaboration with Anastasios Tsaousis and his lab at the University of Kent. Lyto Yiangou and Diego Cantoni optimized the Western blots and microscopy experiments, and therefore some of their work is included alongside the initial work of E. Herman.

To Chris

For your love, understanding, and unwavering support

Acknowledgements

With the combination of two years of undergraduate research projects and courses, and six years of doctoral studies, I have spent the last eight years working in the Dacks Lab. In that time, I have had the pleasure of working with so many exceptional individuals, both at the University of Alberta and in the field of evolutionary protistology.

First, I have to express my deepest gratitude to my supervisor, Joel Dacks. Joel is an incredible scientist, teacher, and friend, and has profoundly influenced me as a scientist and a person. I am so thankful for his willingness to let me "steer the ship" on my various projects, and his encouragement to explore teaching and leadership opportunities outside of the lab. I cannot express how grateful I am for his mentorship over the years, and I seriously doubt I would be as successful as I have been in my PhD without his guidance and support.

Thank you to my supervisory committee, who have helped shape the course of my PhD: Andrew Mason, Vivian Mushahwar, and Gary Eitzen. I would also like to thank Bart Hazes, both for co-supervising my Biology 499 research project that grew into my main PhD project, and for serving as an examiner during my thesis defense. A special thank you goes to Jane Carlton for agreeing to be the external examiner for my thesis defense. I very much appreciate my committee's critical reading of this mammoth thesis.

I am so incredibly grateful for my fellow Dacklings. Over the years, the lab has changed and matured, and grown and shrunk, but the atmosphere of creativity, curiosity, and cooperation has been constant. The first Dacks Lab member I must thank is my very dear friend Mary Klute. Working in the lab together from the beginning, I have always been inspired by Mary's scientific rigour and strong work ethic. We made a great team working on genome projects together, and I will always cherish the time we spent in the lab. Thank you for your friendship and support both in science and in life. To the current Dacklings – Inmaculada Ramírez-Macías, Lael Barlow, Chris Klinger, Beth Richardson, Shweta Pipaliya – thank you for making it a pleasure to come in to work every day. I have always been able to count on you for help and advice in my work, and for making research fun. Thanks especially to Beth, who has not only been a wonderful colleague, but also an amazing friend. To past Dacklings: I am especially grateful to Jeremy Wideman for compelling me to think philosophically about science, and to Alex Schlacht and

Maria Aguilar, both for stimulating scientific discussions and for helpful advice. Thank you to Fred Mast for his critical insight and for being an excellent scientific role model. Thanks also to undergraduates that I have worked with in the lab over the years: Jelena Borovac, Kevin Mowbrey, Casey Francisco, Ben Pi, Nadeem Ali, Kelly Zerr, Meagan Chiu, Nerissa Nankissour, Alex Filbert, and Raegan Larson. I have also had the distinct pleasure of supervising two talented and inquisitive undergraduate students: Serah Jacob and Laura Lee. Thank you for the hard work, enthusiasm, and for helping me to become a better teacher.

Thank you to the Department of Cell Biology; to fellow students, thank you for the camaraderie and friendship through the years; to the faculty, thank you for the excellent scientific training and support; and to the staff, thank you for providing the administrative support that is critical for a successful PhD. Special thanks to the IST personnel for managing our computer cluster.

No bioinformatics lab is an island. I must therefore thank our collaborators for sharing their data with me, although they are too numerous to list here. However I would like to particularly thank John Archibald and Betsy Read for allowing me – as an inexperienced undergraduate summer student – to contribute to their genome projects. I would also like to express my gratitude to the Protistology community, including collaborators and friends of the Dacks Lab: Anna Karnkowska, Lucas Paoli, Richard Dorrell, Giselle Walker, Kina Sena, Ed Glücksman, Mark Field, Nicola Baker, and Anastasios Tsaousis. Thank you to the entire Tsaousis Lab for welcoming me into your lab for three weeks to learn some molecular biological techniques, and especially to Lyto Yiangou and Chris Miller, who very kindly spent those three weeks sharing their knowledge and expertise with me.

Finally, I have to thank my friends and family. Thank you to Serah, who has been such a fantastic friend and source of strength during difficult times. I truly appreciate your love and support. To my parents, JoAnne and Lee: thank you for your love and generosity, your constant encouragement, and your eternal enthusiasm about my research. To my partner Chris, the person who changed my life forever, opened my eyes to new ideas, gave me newfound courage and confidence, thank you. Life is so much bigger with you in it, and I've never been so excited to see what comes next. You constantly inspire, enrich, and enliven me. This thesis would not have been possible without you.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

CDD: Conserved Domain Database

CEG: Core eukaryotic gene

Cvt: Cytoplasm-to-vacuole targeting

DE: Differential expression

ER: Endoplasmic reticulum

ERAD: ER-associated degradation

EST: Expressed sequence tag

FECA: First eukaryotic common ancestor

GO: Gene Ontology

HMM: Hidden Markov model

HSP: High-scoring segment pair

IEM: Immuno-electron microscopy

IFM: Immuno-fluorescence microscopy

IFT: Intra-flagellar transport

LE: Late endosome

LECA: Last eukaryotic common ancestor

LGT: Lateral gene transfer

LRO: Lysosome-related organelle

ML: Maximum-likelihood

MRO: Mitochondria-related organelle

MVB: Multi-vesicular body

OPH: Organelle Paralogy Hypothesis

PAS: Pre-autophagosomal structure

PSSM: Position-specific scoring matrix

SAR(H): Stramenopiles, Alveolates, Rhizaria, and Haptophyta

TGN: *trans*-Golgi network

UPR: Unfolded protein response

1. Introduction

1.1 Introduction

Microbial eukaryotes are diverse and abundant, capable of occupying nearly every environment on earth. Genomic complexity is often associated with multicellularity, but with ever-increasing genome and gene expression data from diverse microbial eukaryotes, it is clear that they are no less complex or sophisticated than multicellular organisms. *In silico* analysis of high-throughput genomic and transcriptomic data is critical in understanding the evolutionary biology and cell biology of protists for three major reasons. First, –omics approaches allow predictions to be made about the function of a cellular system without relying on genetic techniques that are currently not developed in many eukaryotes. Second, studying the evolution of gene content and expression across a lineage that has evolved a certain trait (e.g. parasitism) is a discovery-based approach that enables a thorough understanding of cellular system function and trait evolution. Finally, it allows for the evolution of a cellular system to be studied across the diversity of eukaryotes, both giving context to results of functional work performed in a handful of model systems, and giving a more accurate view of the general biology of the system in eukaryotes. Therefore, this thesis explores the biology of a selection of microbial eukaryotes across the eukaryotic tree through genomic and transcriptomic analysis. The particular cellular system of interest is the membrane trafficking system, which underpins endomembrane organelle function, a defining feature of all eukaryotic cells. This chapter begins with an introduction to eukaryotic diversity and the membrane trafficking system, followed by a discussion of how comparative genomics and transcriptomics can inform evolutionary protistology, and specifically, how these techniques have been used in this thesis to understand membrane trafficking system function in microbial eukaryotes with specialized lifestyles and cellular processes.

1.2 Eukaryotic Diversity

All extant eukaryotes have evolved from a last eukaryotic common ancestor, or LECA. By using genomic data, it is possible to reconstruct the cellular systems present in the LECA. However, the LECA can be thought of as an 'event horizon', as earlier evolutionary events are much more difficult to infer (by definition, there can be no extant eukaryotes that diverged prior to the LECA). Pre-LECA evolution is essentially the process of eukaryogenesis, which describes changes in the first eukaryotic common ancestor (FECA, with one or more eukaryotic features) that led to the LECA. Eukaryogenesis involved the acquisition of a membrane-bounded nucleus, a basic phagocytosis-like mechanism and internal membranes, and the engulfment and retention of a α-proteobacterial endosymbiont as the mitochondrion. Several models have been proposed to explain the order of these events, and are hotly debated in the field of early eukaryotic evolution.[1–7] Regardless, these features were well in place by the time of the LECA.[1,5,6]

It is now accepted that the LECA was complex and resembled a modern eukaryotic cell, with a fixed mitochondrion, and a complex nucleus, endomembrane system, cytoskeleton, and intracellular signaling system, among others.[8] Evidence for this is discussed below. Fossil and molecular clock phylogenetic analysis has estimated the age of the LECA to range widely from approximately 1-2 billion years ago, but the supergroups that descended from the LECA diverged from it within only 300 million years.[9] In other words, extant eukaryotes are the result of a 'Big Bang' of rapid diversification.[10] The work in this thesis is chiefly concerned with reconstructing membrane trafficking system evolution from this point onward; how diverse eukaryotic lineages have evolved and what more this can tell us about general membrane trafficking, and the evolvability of membrane trafficking components.

Evolutionary analysis is necessarily comparative, and it relies on knowledge of the taxonomic relationships and cell biological features of the lineages of interest. This section of the introduction provides an overview of our current understanding of the eukaryotic supergroups, including general cell biology and the taxonomic relationships of the organisms studied herein, and genomes available for comparative analysis. A tree summarizing our current knowledge of eukaryotic supergroup phylogeny is shown in Figure 1.1. Supergroup identity and characteristic features are defined here as in Adl *et al.* (2012), the most recent classification of eukaryotes, put forth by the Committee on Systematics and Evolution of The International Society of Protistologists.[11] There are currently five eukaryotic supergroups: the Opisthokonta, Amoebozoa, Excavata, Archaeplastida, and the SAR clade. The CCTH clade was a sixth supergroup previously thought to exist, but more recent work has shown that this clade is not monophyletic, effectively dissolving it. Furthermore, since the Adl *et al*. (2012) publication, additional phylogenomic evidence has clarified intra-supergroup relationships, notably within the Fungi, Radek *et al.* (2017); the former CCTH clade, Burki *et al.* (2012); the Amoebozoa, Kang *et al.* (2017); the Rhizaria, Krabberød *et al.* (2017); and the Stramenopiles, Derelle *et al.* (2016).[12–16] In the following sub-sections, the general features of the five eukaryotic supergroups are described, and included are examples of high quality sequenced genomes for sampling in comparative genomic analyses.

1.2.1 *Opisthokonta*

The Opisthokonta is comprised of two major lineages, which are the most studied in cell biology: the Holozoa (animals and their unicellular relatives) and the Fungi.[17] Multicellularity

Figure 1.1. Phylogeny of eukaryotes.

Cartoon schematic of evolutionary diversity within the five eukaryotic supergroups and their relative relationships. It is based on a version published in Walker *et al.* (2011),[19] which was the result of numerous large-scale genomic and phylogenetic analyses and comparative ultrastructural data. This figure has been updated with more current phylogenetic and phylogenomic data, specifically from Kang *et al.* (2017),[14] Radek *et al.* (2017),[12] Burki *et al.* (2012),[13] Krabberød *et al.* (2017),[15] and Derelle *et al.* (2016).[16] The two current rooting hypotheses (Derelle *et al.* 2015 and He *et al.* 2014)[20,21] are indicated by a dashed line. Line lengths are arbitrary. Well-studied model organisms are shown in bold, as are the organisms that are of focus in this thesis. In the SAR clade, two large Stramenopile lineages are further dissected; the Bigyra includes Labyrinthulea and Opalozoa, while the Gyrista includes Pelagophyceae, Diatomeae, Bolidophyceae, Dictyochophyceae, Chyrista, and Oomycota.

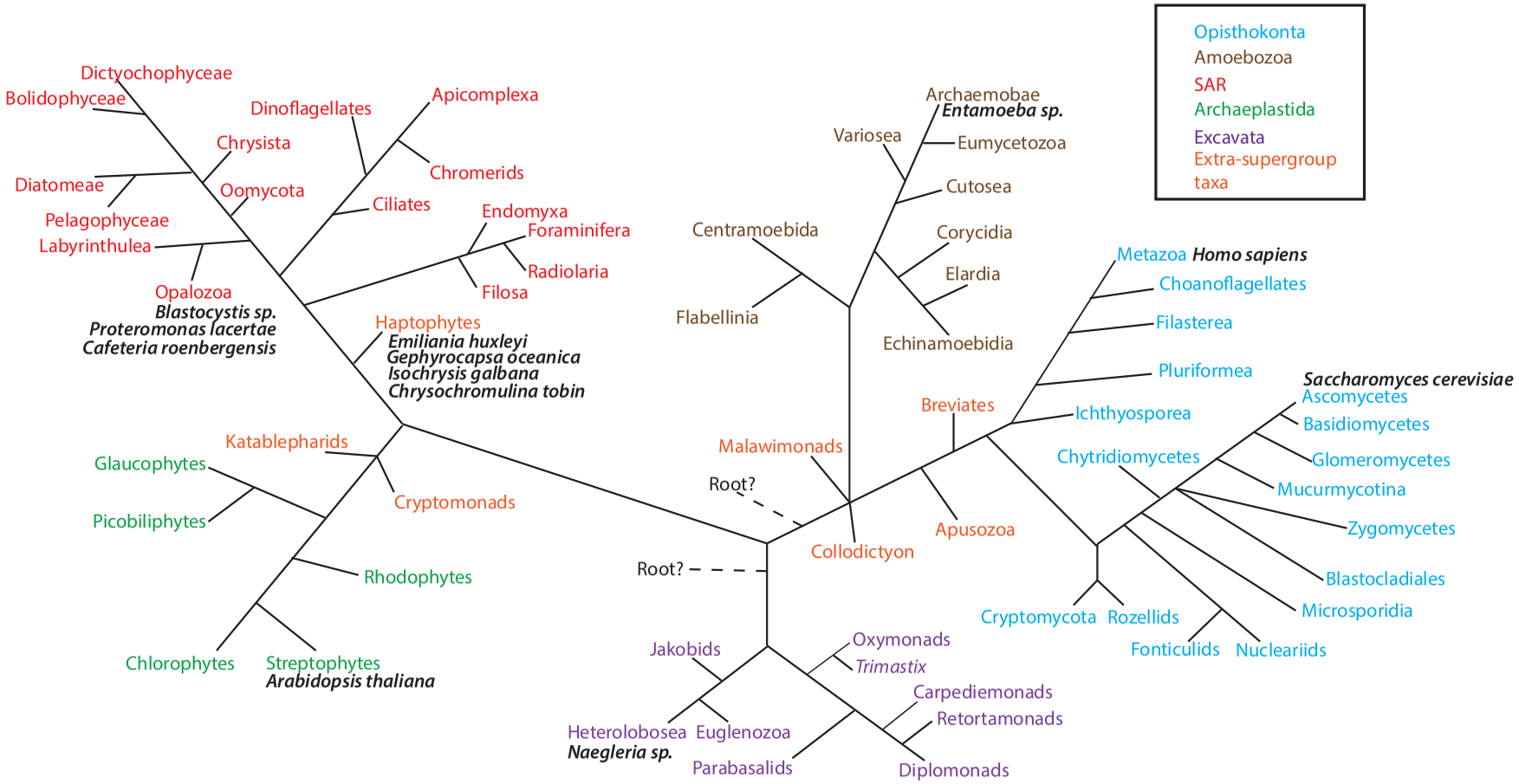

Opisthokonta
Amoebozoa
SAR
Archaeplastida
Excavata
Extra-supergroup taxa
Dictyochophyceae
Bolidophyceae
Dinoflagellates
Apicomplexa
Chrysista
Chromerids
Diatomeae
Oomycota
Ciliates
Endomyxa
Pelagophyceae
Foraminifera
Labyrinthulea
Radiolaria
Filosa
Opalozoa
Blastocystis sp.
Proteromonas lacertae
Cafeteria roenbergensis
Haptophytes
Emiliania huxleyi
Gephyrocapsa oceanica
Isochrysis galbana
Chrysochromulina tobin
Katablepharids
Glaucophytes
Cryptomonads
Picobiliphytes
Rhodophytes
Chlorophytes
Streptophytes
Arabidopsis thaliana
Archaemobae
Entamoeba sp.
Variosea
Eumycetozoa
Cutosea
Centramoebida
Corycidia
Elardia
Flabellinia
Echinamoebidia
Metazoa Homo sapiens
Choanoflagellates
Filasterea
Pluriformea
Breviates
Ichthyosporea
Saccharomyces cerevisiae
Ascomycetes
Basidiomycetes
Malawimonads
Chytridiomycetes
Glomeromycetes
Root?
Mucurmycotina
Zygomycetes
Apusozoa
Collodictyon
Root?
Blastocladiales
Microsporidia
Cryptomycota
Rozellids
Fonticulids
Nucleariids
Jakobids
Oxymonads
Trimastix
Carpediemonads
Retortamonads
Heterolobosea
Euglenozoa
Naegleria sp.
Parabasalids
Diplomonads

Figure 1.1

has arisen independently in this lineage at least four times: once at the base of animals, in the fonticulid slime molds, and in the ascomycete and basidiomycete Fungi.[18] Both *Homo sapiens* and *Saccharomyces cerevisiae* are found in this supergroup – representing roughly one sixth of the diversity of eukaryotes – implying that the phrase "conserved from yeast to man" does not actually describe a particularly ancient or well-conserved eukaryotic feature. Numerous vertebrate and invertebrate genomes are available, as model systems abound in these groups: *Mus musculus*, *Ciona intestinalis*, *Danio rerio*, *Drosophila melanogaster*, *Caenorhabditis elegans*, et cetera. Genome sequencing efforts have also focused on major transition points, such as the origin of multicellularity in animals and fungi,[22] the origin of the bilateral body plan,[23] and the origin of the vertebrates.[24] Basal taxa, diverging prior to the Metazoa (multicellular animals), include the filasterian *Capsaspora owczarzakii*, *Sphaeroforma arctica* (Ichthyosporea), and the choanoflagellate *Monosiga brevicollis*. Collectively, the Metazoa and their unicellular relatives are termed the Holozoa.

The other major branch of the Opisthokonta is the Fungi. Listed here by taxonomic relationship are organisms representing different fungal clades with publicly available genomes.[25] Beginning with the most widely used fungal model organism, *S. cerevisiae* is a member of the Ascomycota. The Ascomycota, together with the Basidiomycota (e.g. *Ustilago maydis*), form the Dikarya: unicellular or filamentous Fungi. Diverging prior to this clade are organisms that tend to be saprotrophic, parasitic, or symbiotic, including the glomeromycete *Rhizophagus irregularis* and *Phycomyces blakesleeanus* (Mucormycotina). Other parasitic Fungi include members of the Chytridiomycotina, such as *Batrachochytrium dendrobatidis* and *Spizellomyces punctatus*. The most basal Fungi are the Microsporidia; intracellular parasites whose genomes are notable for their incredible reduction.[26–28] As in the Holozoa, there are

several taxa that are basal to the Fungi, such as *Rozella allomycis* and *Fonticula alba*. Together, they and the Fungi form the Nucletmycea.

There are also organisms basal to the Opisthokonts, such as the apusomonads (*Thecamonas trahens*). Basal to the apusomonads are the Breviates, which include *Breviata anathema* and *Pygsuia biforma*,[29] although they do not have publically available genomes. Together, these extra-supergroup taxa and the opisthokonts make up the Obazoa group, a strongly supported major group that is sister to the Amoebozoa.[29]

1.2.2 *Amoebozoa*

The Amoebozoa includes cells that have an amoeboid life stage, although not all amoebae are members of the Amoebozoa. A detailed understanding of the taxonomic relationships within the Amoebozoa has largely remained elusive until a recent phylogenomic analysis by Kang and colleagues (2017).[14] This supergroup is comprised of two major clades, the Discosea and the Tubulinea+Evosea (Tevosa). Within the Discosea are the groups Flabellinia and Centramoebia; the latter contains the organism *Acanthamoeba castellanii*, which was the first free-living, solitary (non-social) amoebozoan organism to be sequenced.[30] *A. castellanii* is also a human pathogen, as it causes amoebic keratitis and granulomatous amoebic encephalitis, and is an opportunistic pathogen in immunocompromised patients.[31–34] The Evosea contains several of the most well-known amoebozoans. One is the slime mold *Dictyostelium discoideum*, a social amoeba, which under conditions of nutrient scarcity, aggregates to form a multicellular fruiting body that disperses a bolus of spores.[35] Because of this, multicellularity, cell motility, signaling, and cell-cell interaction have been studied in depth in *Dictyostelium*, as well as numerous other cellular systems.[36] While *Dictyostelium* is a non-pathogenic soil organism,

another member of the Evosea is the human gut parasite *Entamoeba histolytica* (Archamoebae). *Entamoeba* is often asymptomatic, although roughly 40,000-100,000 people die annually from amoebic dysentery and its complications, including amoebic colitis and perforation of the bowel, and extra-intestinal infections of the liver, lungs, and brain.[37,38] *E. histolytica* and its relative *Entamoeba invadens*, a reptile gut pathogen, are of special focus in this thesis.

1.2.3 *Archaeplastida*

The Archaeplastida is a highly diverse supergroup including glaucophytes, green and red algae, and multicellular plants. It is defined by an ancestral primary endosymbiosis with a cyanobacterium. Within this lineage, multicellularity has arisen several times.[39,40]

The three basal algal groups of the Archaeplastida are the glaucophytes, rhodophytes and chloroplastida. The glaucophytes and rhodophytes both contain only chlorophyll a, while the chloroplastida contain chlorophylls a and b. Unlike other algae, the glaucophytes have retained the prokaryotic peptidoglycan wall between the outer plastid membrane and the host encapsulating membrane; one such example of this taxon is *Cyanophora paradoxa*.[41] Within the rhodophytes are the Cyanidiales, which live in high temperature and acidic environments, and include the taxa *Galdieria*[42] and *Cyanidioschyzon*.[43] While genome reduction and streamlining is typically associated with parasitism, the harsh environments in which these organisms live have had a similar effect on their genomes. There are several lineages of multicellular red algae, which include organisms such as *Chondrus crispus* and *Porphyra umbilicalis*.

The earliest branching Chloroplastida are the Trebouxiophyceae and Chlorophyceae. The lineages Prasinophyceae, Charophyta, Bryophyta, Angiosperms, and Gymnosperms branch off in

that order. Single-celled algae dominate until the bryophytes, which include mosses, and land plants in the angiosperms and gymnosperms. Genomes commonly used for comparative genomics are *Chlorella* in the trebouxiophytes; and *Volvox carteri* and *Chlamydomonas reinhardtii* in the chlorophyceae.[44] *V. carteri* is a multicellular colony organism. Because of their central importance to agriculture and ecology, the cell biology and membrane trafficking of plants have been studied extensively, and the clade contains well-developed model organisms. However, because the experimental focus has historically been in the context of multicellular developmental processes (e.g. leaf growth) or responses to plant-specific stresses (e.g. abscisic acid stress response-related protein), it can be difficult to extrapolate gene function in other eukaryotes. Nonetheless, plant model systems have provided a wealth of functional knowledge to complement the work done in human and yeast on the other 'side' of the tree.

The most closely related outgroup to the Archaeplastida contains the cryptomonads and katablepharids. The cryptomonads include the organism *Guillardia theta*, which contains a secondary red algal endosymbiont, including the remnant nucleus of the alga, termed a nucleomorph.[45] Along with the rhizarian *Bigelowiella natans*, these were the first nucleomorph-containing organisms to have their nuclear genomes sequenced.[46]

1.2.4 The SAR(H) Assemblage

SAR is a major clade composed of the Stramenopiles, Alveolates, and Rhizaria, and more recently, the Haptophyta. The Alveolates and Rhizaria are sister taxa. Rhizaria are divided into the Cercozoa and the Retaria. Rhizaria are extremely diverse, making it difficult to make generalizations about this group. They can be amoebae, often with filose pseudopodia, while others are covered by silica scales. Some are parasites, such as the potato parasite

Plasmodiophora brassicae.[47] Two publicly available rhizarian genomes are those of the chlorarachniophyte alga *Bigelowiella natans* and the foraminiferan *Reticulomyxa filosa*.[46,48]

The Alveolates contain many important parasites of humans and animals. The most famous is the apicomplexan *Plasmodium falciparum*, the causative agent of malaria. The apicomplexan group is almost entirely comprised of parasites; some examples are *Toxoplasma gondii* (toxoplasmosis), *Theileria parva* (theileriosis), *Babesia bovis* (babesiosis), and *Cryptosporidium parvum* (cryptosporidiosis). The single non-parasitic apicomplexan lineage is *Nephromyces*, a beneficial symbiont of animals (tunicates).[49] Outside the Apicomplexa are free-living or symbiotic groups such as the chromerids, including *Chromera velia* and *Vitrella brassicaformis*, and the perkinsid *Perkinsus marinus*. These taxa have been critical to understanding the gradual evolution of parasitism in the apicomplexa.[50] Basal to the Chromerids are the Dinoflagellates, which often have symbiotic relationships with coral. One such example is *Symbiodinium*.[51]

The third taxon in SAR is the Stramenopiles; another group with an independent development of multicellularity in the brown algae (e.g. *Ectocarpus siliculosus*).[52] Internal relationships of the Stramenopile clade have historically been difficult to resolve.[53–56] However, a recent phylogenomic analysis by Derelle and colleagues (2016) has shown that the stramenopiles can be split into two major clades, the Bigyra and the Gyrista.[16] The Bigyra contain the labyrinthulomycetes (e.g. *Aurantiochytrium limacinum*),[57] and relevant to this thesis, the Opalozoa, which includes *Blastocystis* sp., *Proteromonas lacertae*, and *Cafeteria roenbergensis*.[58] *Blastocystis* is a gut endobiont found in anywhere between 5-60% of a country's population, with some populations reaching 100%.[59,60] It is an enigmatic parasite of humans and animals, causing intestinal symptoms and dysbiota, although it is also found in

healthy individuals. *Proteromonas* is also a gut endobiont of animals, although its status as a pathogen is unclear, while *Cafeteria* is a free-living marine organism. In this thesis, comparative genomics is used to study the evolution of the membrane trafficking system during the transition from free-living to parasite/endobiont. The other major clade of stramenopiles is the Gyrista, which includes the oomycetes (e.g. *Phytophthora infestans,* the causative agent of potato blight) and the ochrophytes (diatoms and the multicellular *Ectocarpus*). Photosynthesis has likely been acquired in the ochrophytes by a tertiary endosymbiosis of a red algae-containing cryptophyte.[61]

Branching basal to the SAR is the Haptophyta, a clade that contains several algae that extrude calcium carbonate scales. As the algae die, the scales sink to the seabed and become part of natural cliff formations such as the White Cliffs of Dover. As such, they play a major role in carbon cycling. Due to the secretory burden involved in scale generation, part of this Thesis explores the membrane trafficking system in the haptophytes *Emiliania huxleyi*,[62] *Gephyrocapsa oceanica, Isochrysis galbana,* and *Chrysochromulina tobin*. *E. huxleyi* and *G. oceanica* are the most closely related of these haptophytes, with *I. galbana* as a sister group. Together, these represent the Isochrysidales clade. Sister to this grouping is *C. tobin*, which is a member of the Prymnesiales. All four organisms secrete scales, however, *C. tobin* and *I. galbana* are only able to make organic scales, while scale calcification occurs in *E. huxleyi* and *G. oceanica*. The membrane trafficking system complements of the scale-forming haptophytes are assessed, as are the trafficking genes that are differentially expressed in response to growth under biomineralizing conditions of calcium and bicarbonate.

1.2.5 *Excavata*

The Excavata occupy an enigmatic position in the eukaryotic tree. As discussed further below, the Excavata are either paraphyletic with the root of eukaryotes, or they are the earliest diverging branch from the LECA.[21] The Excavata can be split into the Metamonada and the Discoba,[63] and both groups contain an array of symbiotic, parasitic, pathogenic – and occasionally free-living – organisms. All metamonads are anaerobic or microaerophilic, and as such, have degenerated mitochondria.[64,65] One metamonad with a recently sequenced genome is *Monocercomonoides* sp., which was shown to have completely lost all remnants of a mitochondrion.[66] Metamonads typically reside in animal guts, or other low-oxygen environments; however, one free-living relative of *Monocercomonoides* sp. has been described (*Paratrimastix* sp.).[67] Also within the metamonada is the parasite *Giardia intestinalis,* and analysis of its genome has shown reductive streamlining, likely in relation to its parasitic lifestyle.[68] However, another metamonad parasite has a vastly expanded genome. *Trichomonas vaginalis* is a sexually transmitted pathogen with over 100,000 new cases annually.[69] Genome sequencing revealed that it has a ~160Mbp genome, comprised of both repetitive regions and massively expanded gene families;[70] evidence that parasitism is not *de facto* reductive.

The other major group of excavates is the Discoba, which include the Discicristata and the Jakobida. Discicristates can further be split into the Heterolobosea and the Euglenozoa. The neuropathogenic free-living amoeba *Naegleria fowleri* is a heterolobosean; organisms in this taxon often have eruptive pseudopodia, as well as an amoeboflagellate stage. *N. fowleri* causes amoebic meningoencephalitis in humans, which, despite being relatively uncommon, kills roughly 98% of patients, and does so in ~2 weeks.[71] In the Euglenozoa, the most infamous members are the trypanosomatids: *Trypanosoma brucei*, *Trypanosoma cruzi*, and *Leishmania major*. Responsible for African Sleeping Sickness, Chagas Disease, and leishmaniasis,

respectively, these hemoflagellates parasitize humans and many other animals, causing major health, social and economic impact around the world. The Jakobida, which include *Jakoba libera* and *Andalucia godoyi,* are more enigmatic. Despite sitting near the likely root of eukaryotes, these organisms have no publicly available genomes.

1.2.6 Rooting the Tree of Eukaryotes

Identifying the root of the eukaryotic tree – the deepest split that bifurcates two monophyletic groups that include all extant eukaryotes – is critical for understanding the evolution of any biological character. Because of the rapid diversification of extant lineages at the LECA, defining supergroups and their relationships, as well as pinpointing this root, has been a daunting task.[72–75] Previous efforts have been hampered by phylogenetic artifact and lack of sequence data. However, increases in computational power have enabled sophisticated tree-building methods, models of sequence evolution, and topology testing algorithms to be developed. In addition to this is the vast increase in available genome sequence from many diverse eukaryotes.[76] While the question of the root is not fully resolved, there are two main hypotheses for its placement, outlined in Derelle and Lang 2012[77] (updated in Derelle *et al.* 2015),[77] and He *et al.* 2014.[21] Both place the root within or near the Excavata supergroup. He *et al*. (2014)[21] propose that the root of eukaryotes lies between the monophyletic clade of Excavata and the rest of eukaryotes, which then branch into two clades: Opisthokonta + Amoebozoa and SAR(H) + Archaeplastida. On the other hand, Derelle and Lang (2012)[77] suggest that the Excavata groups with SAR(H) + Archaeplastida, and the root lies between this clade and the Opisthokonta + Amoebozoa. Using an updated dataset, Derelle *et al.* (2015)[20] show the same relationship as in the 2012 paper, but with the excavate-like *Malawimonas* grouping on the

Opisthokonta + Amoebozoa side instead, effectively making the Excavata paraphyletic. This Thesis remains agnostic about the root; when interpreting the results of comparative genomic analyses, genes proposed to be present in the LECA must be found in at least one member of the Opisthokonta + Amoeobozoa, Archaeplastida + SAR(H), and the Excavata.

1.2.7 Eukaryotic symbiosis and pathogenesis

Defining the terms by which eukaryotes live in association with other eukaryotes is increasingly challenging, due to conceptual shifts brought forth by large-scale sequencing of host-associated organisms and metagenomics of their environment within the host. For the purpose of this thesis, terms describing these relationships are defined as follows. Symbiosis is the close physical association of two organisms, and can be mutualistic (both parties benefit), commensal (one party benefits while the other neither benefits nor is harmed), and parasitic (one party benefits at the expense of the other). Often the line between commensalism and parasitism is very fine, as a commensal organism in one context can be parasitic in another (e.g. certain individuals or populations being asymptomatic). Another criterion of parasitism is that it is an obligatory relationship; a parasite requires on a host for at least one stage of its life cycle. However, this is not necessarily a requirement of commensalism. This is also the criterion by which parasites are distinguished from free-living pathogens, as this latter category includes organisms that infect a host, but *do not require a host for any lifestyle stage.* Free-living pathogens are often (but not always) opportunistic, which means that they take advantage of an opportunity to infect a host. Some examples are a compromised immune system, altered microbiota, or breached integumentary barriers.

Eukaryote-eukaryote symbioses are found between organisms across the tree of eukaryotes, and represent lifestyles of three of the four lineages explored in this thesis. The most well-studied type of symbiotic relationship is parasitism; parasites are found in all supergroups and infect humans, animals, plants, and even other protists.[19,78] Recently, the idea has been raised that some eukaryotes found in human intestines, which were typically thought of as parasites, are simply commensal organisms, or are even beneficial. Two such examples of these are the excavates *Enteromonas hominis* and *Dientamoeba fragilis*.[79] While these are occasionally associated with disease, they are also found in healthy persons, blurring the line between parasitism and other types of symbioses. In some cases, such as *Blastocystis* sp. – which will be studied in detail in this thesis – there is much debate about its lifestyle, as some strains are statistically associated with diarrhea in some populations, but can be mostly asymptomatic in others. It is further complicated by the fact that the mechanism of *Blastocystis* sp. virulence is not well understood, and symptoms are not necessarily acute. The relationships of endobiotic organisms with their host are clearly complex, and likely depend on other factors, such as the gut environment or differences in host genetics. One explanation that has been proposed is that eukaryotic gut organisms are not specifically destructive to further their own survival, but that they secondarily induce dysbiosis, for example, by causing a shift in the lumen microbiota, or perturbing the gut epithelia.[80]

On the other hand, an organism like *Entamoeba histolytica* is a true parasite, as it destroys host tissue and can invade other parts of the body. The earliest work in parasite genomics was aimed at uncovering how close association with a host has sculpted their genomes. Host dependency suggests that the endobiont has lost some cellular function(s), which can be elucidated through comparative genomic analysis. Its ability to survive in the host also requires

factors for host colonization and immune system modulation, and thus parasite-specific gene family expansions are also common. These changes show that host-parasite co-evolution is the major contributor to parasite genome evolution.

Free-living pathogens are another category of infectious eukaryotes. Instead of association with a host, they are found in the environment and only infect under favourable conditions. Two examples of free-living pathogens are *Acanthamoeba castellanii* and *Naegleria fowleri* (a topic of this thesis). While *A. castellanii* is considered an opportunistic pathogen, *N. fowleri* is not; it infects the eyes and brain of healthy, immuno-competent individuals. Because co-evolution with a host does not drive genomic change in free-living pathogens to the same extent as in a parasite reliant on a host, analyses focus instead on identifying pathogenicity factors.[81]

Genomic analyses of both parasitic/endobiotic organisms and free-living pathogens benefit from comparison with the genomes of closely related organisms that are free-living and/or non-pathogenic. This is because of the lineage-specific expansions and losses in cellular machinery that are unrelated to a parasitic or pathogenic lifestyle. With more sampling points and closely related taxa, our view of endobiont/pathogen evolution becomes better informed. The detailed understanding of genomic evolution that this type of comparison generates is only now being adequately appreciated due to the breadth and depth of eukaryotic taxa that have been sequenced. In addition to parasitism and pathogenesis, any other heritable cellular behavior or lifestyle can be treated this way.

In this thesis, evolution of four diverse eukaryotic lineages is assessed by comparative genomic analysis of the membrane trafficking system. The following section is an overview of membrane trafficking in eukaryotes.

1.3 Overview of Membrane Trafficking

Membrane trafficking is a key cellular process that contributes to organellar homeostasis and renewal, nutrient acquisition, waste disposal, and cell-environment interaction, among other processes. Generally, membrane trafficking is the vesicular transport of protein and other macromolecules between intracellular compartments, and by which cells take up and release materials into the extracellular environment. Generally, membrane trafficking can be roughly broken down into the phases of cargo selection and coat recruitment, vesicle formation and scission, transport from donor to acceptor compartment, and tethering and fusion with the target membrane.[82] The next subsection will give an overview of the endomembrane organelles, followed by the vesicle formation and vesicle fusion machinery responsible for trafficking between them.

1.3.1 Endomembrane organelles of eukaryotes

The endomembrane organelles are the endoplasmic reticulum (ER), the Golgi body and *trans*-Golgi network (TGN), all stages and derivations of endo-lysosomal organelle (early, late, and recycling endosomes; lysosomes; lysosome-related organelles), and the multi-vesicular body (MVB). The plasma membrane, although not an organelle, is also part of the endomembrane system, and it is central to endo- and exocytosis. Lipid droplets, while critical ER-derived lipid storage organelles, do not play a role in endocytosis or secretion, and therefore are not dealt with in this thesis. Finally, autophagosomes and peroxisomes are at least partly endomembrane-derived organelles,[83,84] but also do not contribute to endo- or exocytosis. A generalized version of the eukaryotic MTS can be found in Figure 1.2. All endomembrane organelles are thought to have been derived by an autogenous – rather than endosymbiotic – mechanism based on

Figure 1.2. Overview of membrane trafficking in eukaryotic cells.

Cartoon illustrating the general membrane trafficking pathways in eukaryotes, and where various proteins and complexes function. Transport is mediated by budding and fusion of vesicles and tubules, as well as organelle fusion and maturation. This diagram contains components thought to be present in the LECA. Components that are ancient, but whose only known function relies on an animal-specific protein, are marked with an asterisk (TBC-E, DENND1, and FLCN regulate animal-specific Rab35 activity). Several ancient proteins are not shown in this diagram as their precise cellular functions or locations are unclear: ACAP, AGAP, ADAP, and ArfGAPC2 (ArfGAPs), and TBC-G, TBC-H, TBC-K, TBC-L, and TBC-RootA (RabGAPs). Figure modified with permission from L. Barlow.

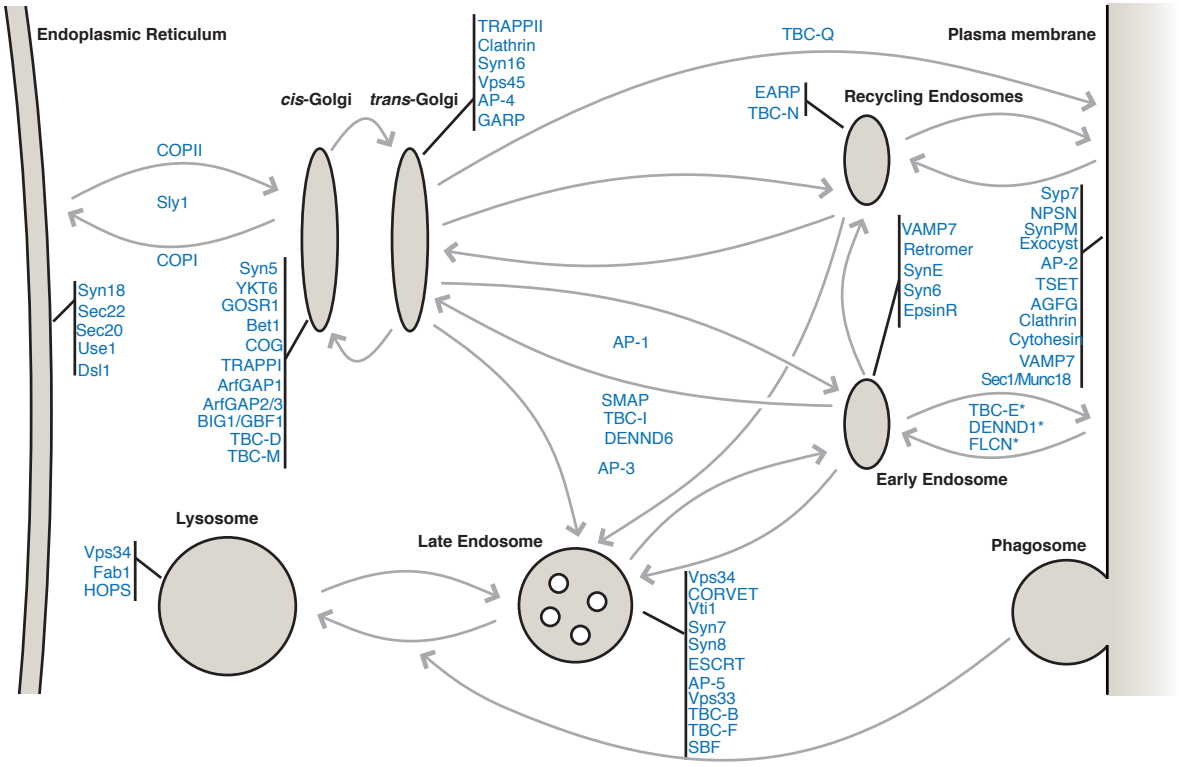

Endoplasmic Reticulum
cis-Golgi
trans-Golgi
TRAPPII
Clathrin
Syn16
Vps45
AP-4
GARP
TBC-Q
Plasma membrane
EARP
TBC-N
Recycling Endosomes
COPII
Sly1
COPI
Syn18
Sec22
Sec20
Use1
Dsl1
Syn5
YKT6
GOSR1
Bet1
COG
TRAPPI
ArfGAP1
ArfGAP2/3
BIG1/GBF1
TBC-D
TBC-M
AP-1
SMAP
TBC-I
DENND6
AP-3
VAMP7
Retromer
SynE
Syn6
EpsinR
Syp7
NPSN
SynPM
Exocyst
AP-2
TSET
AGFG
Clathrin
Cytohesin
VAMP7
Sec1/Munc18
TBC-E*
DENND1*
FLCN*
Early Endosome
Lysosome
Vps34
Fab1
HOPS
Late Endosome
Phagosome
Vps34
CORVET
Vti1
Syn7
Syn8
ESCRT
AP-5
Vps33
TBC-B
TBC-F
SBF

Figure 1.2

paralogous expansion of membrane trafficking gene families.[85] This is termed the 'Organelle Paralogy Hypothesis', and will be discussed further below.

The ER is the location of synthesis of proteins targeted to secretory and endocytic organelles of the membrane trafficking system, and the plasma membrane.[86] These proteins can be inserted into the ER membrane, or translated into the lumen of the organelle. The ER is also the site of initial carbohydrate and lipid modifications of proteins, and synthesis of organellar and plasma membrane lipids.[87–89] The majority of ER-synthesized material is trafficked to the *cis*-Golgi, although an ER-to-plasma membrane trafficking pathway has been discovered in animal cells,[90] and may be more widespread, as the plant syntaxin Syp7 involved in vesicle transport is localized to both the plasma membrane and ER in plant cells.[91]

The Golgi was originally described by Camillo Golgi as an "internal reticular apparatus,"[92,93] and further visualized by Dalton and Felix in 1954.[94] It is often described as having a 'stack of pancakes' morphology; in most eukaryotic cells, the Golgi appears as a set of stacked cisternae. However, independent losses of Golgi stacking throughout the tree of eukaryotes have occurred, most notably in *Saccharomyces cerevisiae,*[95] but also parasites such as *Entamoeba* and *Giardia*, and free-living *Monocercomonoides* sp.[19] In *S. cerevisiae*, the Golgi compartments are dispersed in the cytoplasm, appearing as punctae in immunofluorescent staining of Golgi makers.[96] Other organisms, such as *Entamoeba histolytica,* have unstacked Golgi with dispersed cisternae,[97] while in *Giardia intestinalis*, the Golgi is a set of dynamic structures with tubular connections that arise during encystation.[98] Golgi stacking factors have been defined in mammalian cells, but are not necessarily conserved in other taxa despite the widespread presence of stacked Golgi organelles. It has been proposed that outside of animals, a simpler mechanism exists whereby vesicle fusion proteins promote cisternal adhesion.[99] Nonetheless, the loss of stacking does not

appear to hinder Golgi function, in which proteins transit through it via cisternal maturation, where cisternae mature by way of retrograde vesicles that traffic resident enzymes to earlier stacks. The model of cisternal maturation, as opposed to the model of resident Golgi stacks, is supported by both mathematical modeling[100] and the fact that retrograde vesicular transport occurs in unstacked and stacked Golgi.[101] The Golgi is the site of further post-translational modification and protein sorting, as well as lipid synthesis and polysaccharide synthesis in land plants.[102]

The TGN is a tubulo-vesicular network at the *trans* face of the Golgi, and is a highly dynamic sorting organelle in both endocytic and exocytic transport.[103–105] Trafficking pathways from the TGN are numerous, including both to and from early and recycling endosomes, to the plasma membrane, late endosomes, lysosomes, and lysosome-related organelles. In mammalian cells, the TGN is organized by sub-domains,[106,107] from which vesicles can bud by means of a coat complex, or by uncoated vesicles or tubules that fuse extend and fuse with target organelles.[108] From here, constitutively secreted proteins are trafficked to the plasma membrane, while those destined for regulated secretion are trafficked to a lysosome-related secretory organelle. Lysosome-related organelles (LROs) are modified lysosomes, and are found throughout eukaryotes, with some examples being melanosomes in human cells, the rhoptry invasion organelle of the Apicomplexa,[109,110] and acidocalcisomes in organisms across the eukaryotic tree.[111] Another role of TGN trafficking is transport of newly synthesized proteins to endolysosomal organelles; particularly acid hydrolases via mannose 6-phosphate receptors.[112]

Material is trafficked into the cell by endocytosis, which can be classified as receptor-mediated endocytosis, fluid phase endocytosis or macropinocytosis, or phagocytosis. Vesicles generated from receptor-mediated and fluid-phase endocytosis fuse with the early endosome, the

initial sorting compartment for the mixture of endogenous and exogenous material.[113] The early endosome is sometimes called the 'sorting endosome' for this reason. Phagosomes, on the other hand, are formed when the cell membrane fuses around another cell, which is destined for lysosomal degradation.[114] In the early endosome, receptors involved in clathrin-mediated endocytosis can be rapidly trafficked back to the cell surface directly.[115,116] In some cell types, other plasma membrane proteins are recycled via the recycling endosome, a distinct tubulo-vesicular organelle with membrane subdomains.[117,118] Proteins from the early and recycling endosomes can both be trafficked to the TGN for eventual return to the plasma membrane.[119,120] The endosome matures as material is removed by recycling, a loss of trafficking via tubules, a decrease in pH by the vacuolar ATPase, and a shift in the presence of certain organellar markers.[121–124] Transmembrane proteins and other cargo destined for degradation are internalized through the formation of inward-budding vesicles, forming a multi-vesicular body (a species of late endosome). Late endosomes/MVBs fuse with lysosomes, a highly acidified organelle[125] with an array of soluble acid hydrolases, including glycosidases, proteases, nucleases, lipases, phosphatases, and sulfatases.[126] Fusion occurs first by 'kiss-and-run' interactions, followed by complete fusion.[127,128]

In addition to these organelles that are generally found throughout eukaryotes, there are also lineage-specific or less well-conserved organelles. One such organelle has already been mentioned: the acidocalcisome. Acidocalcisomes are acidic phosphorus- and calcium-storage organelles, and although they have been suggested to be evolutionarily related to dense granules of platelet cells,[129] they may well be functionally analogous and of independent origin. Acidocalcisomes are thought to be lysosome-related organelles, and are found in a diversity of eukaryotes, including parasites such as the trypanosomes, the alga *Chlamydomonas reinhardtii*,

and the amoeba *Dictyostelium discoideum*, as well as in bacteria.[111,129,130] Another organelle that is found across eukaryotes – but not in all eukaryotes – is the contractile vacuole. The contractile vacuole is an osmoregulatory organelle that is found in amoebae, ciliates, trypanosomes, and green algae.[131] Because trafficking pathways to these organelles are currently being explored, they are not of focus here. However, they highlight an arm of the endomembrane system that allows for organellar plasticity, either in terms of novelty or loss, which raises the possibility of additional lineage-specific organelles.

Much of our understanding of organellar function comes from work in human and yeast cells. However, the functions described here fit a general model of eukaryotic cell biology, and the underlying molecular mechanisms are overwhelmingly conserved in other eukaryotes.[85,132] The following section introduces the vesicle formation and fusion machinery that traffics between these organelles. While the bulk of functional data comes from human and yeast models, each section will reference comparative genomic or functional work from other eukaryotes where possible. Not only does this generate a more generalized model of membrane trafficking function, but it also sets the stage for the work presented in this thesis, which exclusively focuses on membrane trafficking in microbial eukaryotes.

1.3.2 Membrane trafficking processes

The basis of membrane trafficking is the formation of vesicles from donor compartments or the plasma membrane, shuttling to the appropriate organelle, and fusion with the acceptor compartment. Membrane trafficking machinery is therefore generally divided in the literature into that which is involved in vesicle formation versus vesicle fusion. A general overview of this process is shown in Figure 1.3.

Figure 1.3. Overview of vesicle formation and vesicle fusion in eukaryotes.

Vesicle formation steps are initiation, budding, and scission, and vesicle fusion steps are tethering, docking, and fusion. These trafficking events involve various coat complexes, membrane deforming proteins, Rab and Arf GTPases, ArfGAPs, SNAREs, and tethers, which function together to transport soluble and transmembrane cargo between different subcellular compartments. Cartoon generated by L. Barlow; reproduced from Barlow and Dacks (2017)[133] with permission.

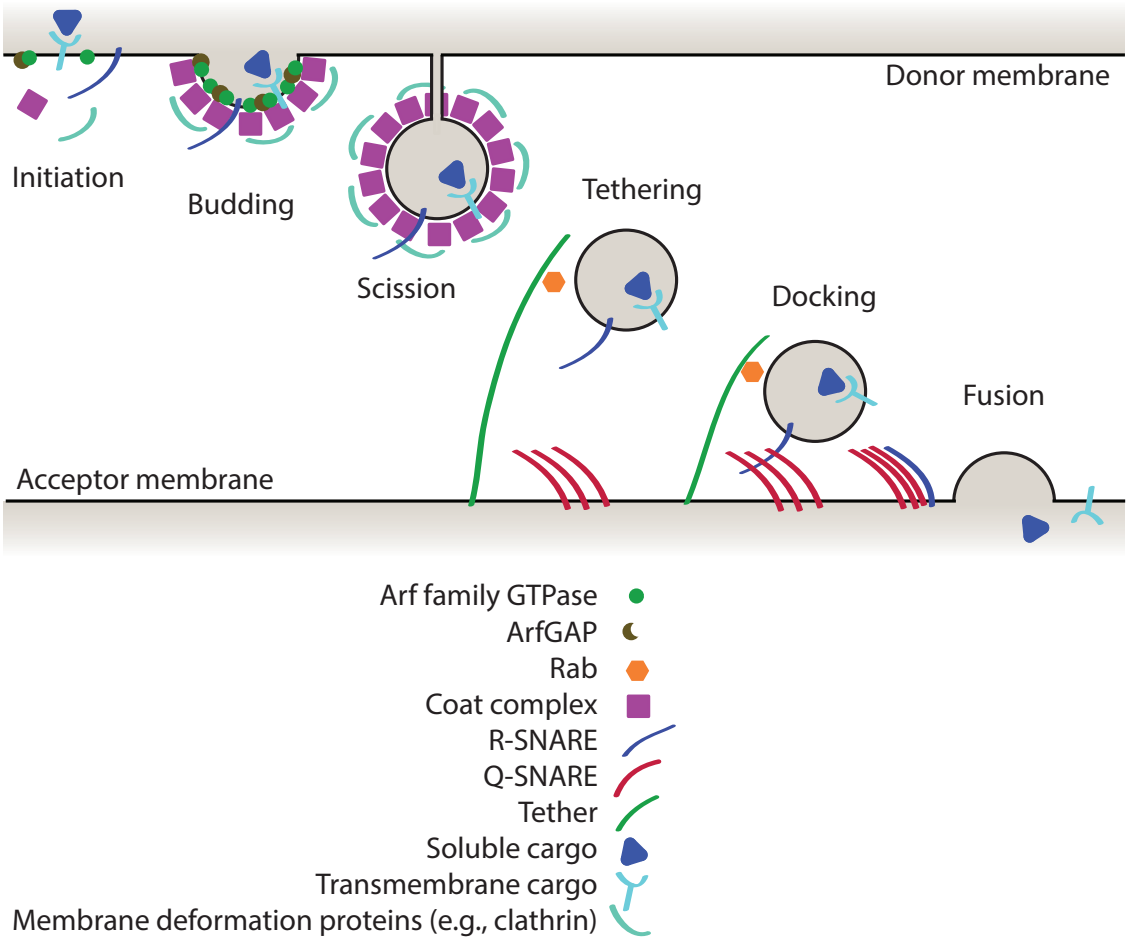

Donor membrane
Initiation
Budding
Scission
Tethering
Docking
Fusion
Acceptor membrane
Arf family GTPase
ArfGAP
Rab
Coat complex
R-SNARE
Q-SNARE
Tether
Soluble cargo
Transmembrane cargo
Membrane deformation proteins (e.g., clathrin)

Figure 1.3

Much of the membrane trafficking machinery is paralogous, and is thought to be the result of a series of gene duplication events beginning with a primordial set of trafficking machinery. This idea is at the heart of the Organelle Paralogy Hypothesis (OPH), first laid out by Dacks and Field (2007).[85] The OPH describes how non-endosymbiotic organelles were acquired in eukaryotes. Gene duplication and coevolution of at least three interacting primordial trafficking factors – an ancestral coat system, a syntaxin SNARE protein, and a Rab GTPase – resulted in the organellar complexity of the endomembrane system, as these trafficking factors together encode pathway specificity and organellar identity. One consequence of the OPH is that the steps in vesicle formation and fusion often occur in the same way at different trafficking pathways and organellar locations. These trafficking steps are broken down in the following sections, as are the trafficking pathways mediated by other machineries that are not the result of paralogous gene expansion.

1.3.2.1 Vesicle formation

Vesicle formation begins with the aggregation of protein cargo in a subdomain of the donor organelle. Soluble cargo is bound by cargo receptors; these, with membrane-associated cargo, bind adaptor or coat proteins via a linear sorting signal.[82] Other subunits of the coat complex are recruited, inducing membrane deformation, vesicle budding, and scission. The energy necessary for these latter processes is provided by GTP hydrolysis.[134,135] Vesicle uncoating occurs following scission from the membrane, although there is evidence that some coats remain until the fusion step, and can participate in tethering.[136,137] Additionally, some vesicle fusion factors are incorporated into the membranes of these vesicles, which help provide trafficking specificity.

Many of the cargo adaptor and coat complexes in the eukaryotic cell belong to a single family, known as the hetero-tetrameric adaptor complex-containing coats, or HTAC-CCs.[138] These are the five adaptor protein (AP) complexes, the TSET complex, and COPI complex. These are part of a larger group of proteins in the cell that are likely related, and they share a similar architecture for membrane deformation: one or two beta-propeller domains followed by an alpha-solenoid domain.[139] Proteins with this architecture are thought to derive from an ancestral protocoatomer complex, and as stated in the OPH, have duplicated giving rise to APs, TSET, COPI, COPII, clathrin, intra-flagellar transport (IFT) proteins, the SEA complex, the HOPS and CORVET multisubunit tethering complexes, nuclear pore complex proteins, and potentially also the retromer coat complex.[140,141]

Within the HTAC-CCs, homology between corresponding subunits is detectable by sensitive bioinformatic methods. Four core proteins make up each HTAC; these are two large subunits, a medium subunit, and a small subunit.[142] In COPI and TSET, there are two outer coat subunits that are recruited as part of the complex (COPA and COPB' in COPI,[143] and TTRAY1 and TTRAY2 in TSET).[144] However, this is variable for the adaptins. Clathrin is the membrane-deforming coat for AP1, AP2, and potentially AP3,[145,146] and the proteins SPG11 ad SPG15 are the coat subunits for AP5.[147] AP-3 does not appear to function with clathrin in mammalian cells,[148,149] while the coat for AP4 is currently unknown.

The HTAC-CCs are an example of the Organelle Paralogy Hypothesis in action; they are the result of paralogous duplication and coevolved divergence of the subunits. In other words, a primordial HTAC-CC (protocoatomer) underwent a series of paralogous gene duplications prior to the LECA, giving rise to the COPI, TSET, and AP complexes that coordinate vesicle formation at various locations in eukaryotic cells. COPI and TSET likely diverged prior to the

adaptins.[144] The TSET complex has retained a patchy distribution in eukaryotes, as only the TCUP (medium) subunit remains in the opisthokonts.[144] In *Dictyostelium*, TSET is associated with the plasma membrane, and in *Arabidopsis*, it is localized to the cell plate that develops during mitosis to divide daughter cells, and also functions in endocytosis.[150] COPI, however, is universally conserved in eukaryotes, and it moves cargo both within the Golgi stack and retrograde from the Golgi to the ER. COPI plays other roles in the cell; it is involved in lipid droplet formation,[151] and COPI-coated vesicles have been observed to bud from the *trans*-Golgi network (TGN) in mammalian cells using a viral protein transport assay.[152] In multicellular plants, the COPI complex is involved in cell plate generation during cytokinesis.[153]

An important factor in COPI cargo sorting, assembly, and budding is the hydrolysis of GTP bound by the small G protein Arf.[135,154] Arf lacks inherent GTP hydrolysis activity, so this action is performed by the Golgi-resident ArfGAPs 1 and 2/3 in mammalian cells.[155,156] Arf regulators, which promote GTP hydrolysis in the case of GTPase Activating Proteins (ArfGAPs), or exchange GDP for GTP in the case of Guanine nucleotide Exchange Factors (ArfGEFs), play an important role in vesicle formation dynamics.[157] The GEFs for mammalian Arf1 are the GBF/BIG family proteins,[158–160] whose activity is mediated by a Sec7 domain. This functionality is likely conserved in eukaryotes, as work in the apicomplexan *Plasmodium falciparum* traced Brefeldin A-resistance in this parasite to a mutation in the Sec7 domain of an ArfGEF protein.[161] It is not clear whether GTP hydrolysis triggers COPI vesicle uncoating, but it does appear to be prerequisite.[162] There is evidence to suggest that at the ER, the multisubunit tethering complex Dsl1 can also trigger this uncoating.[163]

The branching order of the AP complexes is shown in the paper by Hirst and colleagues (2014), in which the related TSET complex is described.[144] AP5 is the most anciently diverging

HTAC, followed by AP3, AP4, and AP1 and AP2. AP3 traffics cargo from tubular endosomes to late endosomes, lysosomes, and LROs,[164–166] while AP5 is thought to function at late endosomes and/or lysosomes.[147] AP4 mediates TGN-endosome trafficking,[167–169] although the directionality is not known. AP1 and AP2 function in tubular endosome-TGN trafficking[170,171] and clathrin-mediated endocytosis,[146,172] respectively. In general, AP1 and AP2 are highly conserved across eukaryotes, with few exceptions (e.g. AP2 is lost in *T. brucei*).[144,173,174] AP3 and AP4 are typically conserved, but have both been lost in at least one major lineage; examples are Apicomplexa and *Saccharomyces*, respectively.[50,175] AP5 is patchily found in eukaryotes, and the complex is often in the process of being lost.[147]

While COPI is responsible for Golgi-ER retrograde trafficking, COPII, made up of Sec23, Sec24, Sec13, and Sec31, traffics cargo in an anterograde fashion from the ER to the Golgi.[176] The small GTPase Sar1, which is related to the Arfs, is the GTPase that drives COPII vesicle formation. Sec23 and Sec24 bind cargo and Sar1, which together form the pre-budding complex.[177] The outer COPII coat is made up of Sec13 and Sec31, which bind this complex and promote membrane deformation. Following vesicle scission, hydrolysis of GTP by Sar1, stimulated by GTPase activating protein (GAP) activity of Sec23, promotes coat disassembly.[137]

The dynamics of clathrin-mediated endocytosis have been studied in depth, leading to the identification of several monomeric clathrin adaptors. Two of these are CALM/AP180 and the Epsins, which, in human cells, have phosphatidylinositol-4,5-bisphosphate-binding ANTH and ENTH domains.[178,179] In mammalian cells and other eukaryotes, a related protein EpsinR is present, however, it functions with clathrin and AP-1 at the Golgi.[178,180,181] AP180 recruits clathrin to the plasma membrane, and plays a role in limiting vesicle size. Epsins interact with the Eps15 or Eps15R proteins at the edge of clathrin-coated pits. Like AP180, Epsin15R interacts

with AP2 in mammals.[182] Lineage-specific clathrin adaptor proteins have also been identified, such as Dab2 in animal cells, and the TBCAP proteins in *T. brucei*.[174,183] Clathrin-coated vesicle scission at the plasma membrane occurs through the GTPase dynamin, which polymerizes around vesicle necks and constricts due to GTP hydrolysis.[184,185] Outside of metazoa, it is not clear that dynamin is involved in clathrin-mediated endocytosis, however, the ciliate *Tetrahymena* does recruit dynamin in this process, which is thought to be the result of convergent evolution.[186] Clathrin has been mostly studied as an endocytic factor, but it also functions at the TGN with the AP-1 complex and EpsinR.[187] In this context, Arf1, rather than dynamin, interacts with the AP1 and clathrin.[188] This is also the case in *Trypanosoma brucei*, which requires Arf1 for endocytosis and trafficking between the Golgi and lysosomes.[189]

Retromer retrieves internalized plasma membrane receptors from endosomes to the TGN, where they can then be recycled to the plasma membrane.[190,191] The cargo-binding subunits of retromer are Vps26 and Vps35, while Vps29 acts as a scaffold protein.[192,193] In humans and yeast, membrane-deforming subunits contain PX and BAR domains. In human cells, these are sorting nexin (SNX) 1 and SNX2 proteins, which may form homodimers or heterodimers. In yeast, they are Vps5 and Vps17.[194,195] The PX domain binds phosphatidylinositol 3-phosphate in the membrane, while the BAR domain induces membrane curvature. Vps17 is not found outside the Opisthokonta,[196] and may be a lineage-specific duplication of Vps5. PX-BAR domain-containing proteins are patchily distributed across eukaryotes; for example *Bigelowiella natans*, which has a complete retromer coat, does not appear to have any PX-BAR domain proteins. Therefore, some eukaryotes may use another membrane-deforming proteins as part of the retromer coat. Yeast uses the dynamin homologue Vps1 to support tubule fission.[197,198] It is not clear whether mammalian dynamin works in the same way with retromer.

As mentioned above, ArfGAPs and GEFs regulate the GTPase function of Arf. Most ArfGAP subfamilies have representatives in members of at least two eukaryotic supergroups, including SMAP, ArfGAP1, ArfGAP2/3, ACAP, AGAP, ADAP, AGFG, and ArfGAPC2.[199,200] More lineage-specific ArfGAP subfamilies have been characterized in the metazoa, including GIT and ARAP. It is likely that other eukaryotes have similar lineage-specific innovations in this protein family. Many ArfGAP proteins have similar or overlapping functions, particularly in the endocytic system. ArfGAP1 and ArfGAP2/3 are involved in COPI trafficking,[154,201,202] SMAP is involved in endocytosis and endosome-TGN recycling,[203,204] AGFG, also known as Hrb, is involved in clathrin-mediated endocytosis,[205] and ACAP is involved in endocytosis and cell movement via Arf6-mediated actin remodelling.[206,207] AGAP and ADAP are also involved in the endocytic system and acting remodeling, respectively.[208,209] The function of ArfGAPC2 is unknown, as it was only recently discovered and is not present in opisthokonts. In other organisms, the ArfGAPs are less well studied. However, an SMAP orthologue was identified as a clathrin-interacting protein in *T. brucei*, consistent with its role in mammalian cells.[183] Outside of the metazoa, ArfGEFs can be divided into the GBF/BIG subfamily and the cytohesin subfamily. Both families contain a Sec7 domain, while cytohesins contain an N-terminal coiled-coil domain for protein interaction, and a C-terminal pleckstrin homology (PH) domain for phosphoinositide interaction.[160] The GBF/BIG family regulates COPI function at the Golgi,[158,210,211] and the cytohesins are involved in endocytosis and actin-based motility.[212] It is not clear how conserved these roles are outside of mammalian cells, as plant ArfGEFs (GNOM proteins) do not appear to have PH domains, and yet have been shown to be involved in endocytosis and endosomal recycling, and the Sec7 GEF in *Tetrahymena* is localized to cilia.[213–215]

Finally, there is a set of non-protocoatomer-derived vesicle formation machinery called the Endosomal Sorting Complexes Required for Transport (ESCRTs). Unlike other coats, the ESCRTs induce negative membrane curvature at late endosomes to form inward-budding vesicles, thereby generating a multivesicular body (MVB).[216–218] ESCRTs are also involved in cytokinesis in mammalian, plant and archaeal cells.[219–223] In mammalian cells, they link membrane scission and microtubule disassembly to separate the two daughter cells. It is likely that this was the ancestral ESCRT function, as the archaeal ESCRT homologues share this role. ESCRTs are made up of five complexes, ESCRT 0, I, II, III, and III-associated (III-A). The ESCRT 0 complex is responsible for binding and clustering ubiquitinated cargo, and is made up of Hrs and STAM in mammalian cells, homologous to Vps27 and Hse1, respectively, in yeast. These proteins are united by the presence of an N-terminal VHS domain (Vps27/Hrs/STAM), GAT (GGA and Tom1), and ubiquitin-interacting motifs (UIMs).[224,225] Vps27 and Hrs also contain an intervening FYVE domain, which binds phosphatidylinositol 3-phosphate in membranes. However, previous evolutionary analyses have shown clear orthologues of these proteins are only found in other members of the opisthokonts.[226] Other VHS domain-containing proteins are more broadly conserved; the Tom1 family of proteins also have an N-terminal VHS domain, followed by a ubiquitin-binding GAT domain, and are able to sort ubiquitinated cargo in both human cells[227,228] and *Dictyostelium discoideum*.[229] A single Tom1 family protein termed Tom1esc has been found across the eukaryotic tree, and was therefore likely present in the LECA.[230] While the role of Tom1esc in ESCRT-specific cargo sorting is not well-described, both it and other ubiquitin-binding VHS domain proteins may be involved in recruiting cargo for MVBs in other eukaryotes. The ESCRT I complex is a heterotetramer of Vps23, Vps28, Vps37, and Mvb12, which together assembles into a stalk and fan-shaped headpiece that interacts with

cargo (via Vps23) at the membrane neck of the inwardly budding vesicle.[231] Mvb12 is opisthokont-specific, suggesting that other organisms, ESCRT I functions either as a trimer or with another protein in its place. ESCRT I co-assembles with ESCRT II, which is made up of Vps22, Vps36, and two copies of Vps25, in a Y-shaped heterotetramer.[232,233] ESCRT II components bind both ESCRT I and ESCRT III proteins, and PI3P lipids in the membrane.[234,235] ESCRT III subunits are involved in membrane scission in both MVB formation and cytokinesis. Scission is performed by Vps20, Snf7/Vps32, and Vps24, with Snf7 multimerising into spirals at the neck of the vesicle.[236] Vps2 and Vps24 bind a multimeric Vps4-Vta1 complex, which recycles other ESCRT subunits from the membrane. Did2 (Vps46) and Vps60 recruit and activate the Vps4-Vta1 complex.[236,237] Finally, the yeast protein Bro1 (ALIX in mammalian cells) promotes the recruitment and activity of Doa4, an enzyme that removes ubiquitin from intraluminal vesicle cargo.[238] ESCRT complexes I and II are occasionally lost in eukaryotic lineages, while at least some subunits of ESCRT III and IIIA are virtually always retained (e.g. the ATPase Vps4).

1.3.2.2 Vesicle fusion

The vesicle fusion machinery includes three main groups of proteins: the multisubunit tethering complexes (MTCs), the Soluble *N*-ethylmaleimide-sensitive Factor Attachment Protein Receptors (SNAREs), the cognate Syntaxin-binding (SM) proteins that function with SNARE complexes, and the Rab monomeric G proteins (Figure 1.3). MTCs are protein complexes on the target organelle membrane that tether incoming vesicles at distances up to 30nm, and interact with both Rab GTPases and SNARE proteins, and in some cases also coat complexes and vesicle lipids.[239] As mentioned in the section on vesicle formation machinery, MTCs such as the HOPS

and CORVET complexes are derived from protocoatomer.[240] Some MTC subunits are also SNARE complex proteins and SM proteins, for example Sec20 of the MTC Dsl1 is a Qb SNARE,[241] and Vps33 of the VpsC core is an SM protein.[242] As vesicles are brought close to the target membrane by MTCs, SNAREs promote vesicle docking. SNAREs are coiled-coil proteins that are found on both the vesicle and target membrane. There are four SNARE families, the Qa, Qb, Qc, and R SNAREs; naming conventions are based the presence of a glutamine or arginine at the 'zero layer'.[243,244] One member of each SNARE family combines to form a SNARE complex during vesicle tethering, and it is the hydrophilic residues at the zero layer that form hydrogen bonds with each other, linking the SNARE complex.[245] In addition to the Q and R SNAREs, there are also Qbc SNARES, which have both Qb and Qc SNARE domains. In mammalian cells, these are involved in secretion, particularly in neurons, but are patchily conserved in eukaryotes.[246,247]

SNARE complex formation is aided by four SM proteins; Sec1/Munc18 (exocytosis),[248,249] Sly1 (intra-Golgi, Golgi-ER transport),[250,251] Vps33 (early endocytosis),[252] and Vps45 (late endocytosis),[253] which arose via gene duplication and are found across the diversity of eukaryotes.[254] They interact with SNAREs either directly or indirectly, stabilizing them and promoting SNARE complex formation.[255] The protein NSF (N-ethylmaleimide sensitive factor) then separates the individual SNAREs after fusion, via its AAA ATPase activity.[256] TBC-N is the GAP for Rab11, at the recycling endosome,[257] while TBC-B and TBC-F are Rab7 GAPs, at the late endosome.[258,259] TBC-E functions as a GAP for Rab35 in mammalian cells,[260] which plays a role in 'fast' recycling from early endosomes to the plasma membrane, and in autophagy.[261,262] While TBC-E is ancient, Rab35 is specific to animal cells,[263] raising the question of its Rab interacting partner in other eukaryotes. Further complicating

matters is the fact that there is promiscuity in Rab-RabGAP interaction *in vitro*.[264,265] There are other ancient TBC RabGAPs, such as TBC-G, TBC-H, TBC-K, TBC-L, and TBC-RootA;[266] however, their functions are either not known or are involved in other cellular processes (e.g. cilia formation).

GEF function can occur through the action of DENN proteins, or through the action of certain MTC subunits.[267,268] Vps9 domain-containing proteins also function as Rab5 and Rab21 GEFs in the early endocytic system.[269] Although Vps9 proteins are ancient components of the membrane trafficking system, their evolution across eukaryotes is not addressed in this thesis (Herman *et al.* in preparation). Like the beta propeller-alpha solenoid domain organization that is shared by protocoatomer-derived factors, the medium and small AP subunits and related zeta subunit of COPI, the DENN RabGEFs, the R-SNAREs, and the MTC TRAPP are united by the presence of a longin domain at the N-terminus of the sequence.[270,271] Most of the DENN domain-containing proteins are GEFs for animal-specific Rabs, with a few exceptions. DENND6 is the GEF for Rab14, which functions in TGN-endosome transport, and work in trypanosomes shows that SBF1 functions at late endosomes with retromer and ESCRT components.[272,273] Similarly to TBC-E, DENND1 and folliculin (FLCN) have also been identified as Rab35 GEFs in mammalian cells; part of the fast recycling pathway.[274,275]

Like the coat complexes, the MTCs, SNAREs, SM proteins, and Rabs work at discrete subcellular locations (Figure 1.2). Trafficking from the ER to the Golgi involves the SNAREs syntaxin 5 (Qa), GS28/Bos1 (Qb), Bet1 (Qc), and Sec22 (R).[276] The R-SNARE Ykt6 can swap out for Sec22 in anterograde and potentially retrograde trafficking in yeast.[277] The TRAPPI MTC promotes the tethering of COPII-coated vesicles via both the interaction of Bet3 (TRAPPI) and Sec23 (COPII), and has GEF activity for Rab1, which in turn promotes COPII vesicle fusion.[278]

Sly1 is the SM protein that works in anterograde and retrograde ER-Golgi trafficking, as well as intra-Golgi trafficking. In human cells, TBC1D20 (TBC-M) is the RabGAP for Rab1.[279] In general, the SNARE complements of microbial eukaryotes include one or more paralogues of conserved, ubiquitous machinery, but can also have lineage-specific SNARE duplications. For example the land plants have large SNARE repertoires (~60-70 members), and the majority of these sequences are lineage-specific, and do not have homologues in other taxa.[280] Another example is the R SNARE family in *Paramecium*.[281] This expansion includes both VAMP7-like sequences, and lineage-specific sequences. Novelty in this arm of the membrane trafficking system appears to be commonplace, and points to lineage-specific trafficking pathways in these organisms.

Retrograde transport from the Golgi to the ER is mediated by syntaxin 18/Ufe1, Sec20, Slt1/Use1, and Sec22; and the MTC Dsl1, comprised of Dsl1, Tip20, and Sec39.[241,282] The Dsl1 complex can interact with both Sec22 and Ykt6 R-SNAREs, and the Dsl1 subunit itself can interact with the alpha and epsilon subunits of COPI to promote vesicle tethering and uncoating.[283] Retrograde Golgi transport is mediated by Rab6.[284] Mammalian cells contain an ER-Golgi intermediate compartment, or ERGIC, however it is not clear that the ERGIC exists outside of animal cells as it has not been identified yeast and plants.[285] Intra-Golgi retrograde trafficking involves syntaxin 5 (Qa), Gs28/GOSR1 (Qb), Gs15/BET1 (Qc), and Ykt6 (R), and both the TRAPPII and COG tethering complexes.[286,287] TRAPPII is made up of the same subunits as the TRAPPI complex, with three additional subunits: Trs120, Trs130, Tca17 and Trs65.[288]

The Qa SNARE proteins that mediate post-Golgi exocytic transport are the plasma membrane syntaxins (SynPM), which include syntaxin 1, 2, 3, 4, 11, and 19 in human cells.[289] A

single Qbc SNARE, a member of the SNAP-25 family of proteins, fills the role of the Qb and Qc SNAREs. However, Qbc SNAREs are patchily distributed in eukaryotes, and the role of the Qbc SNARE may be fulfilled by other Qb or Qc SNAREs in other organisms. *Paramecium* encodes a SNAP25-like Qbc SNARE, which is thought to function in endocytosis, and therefore has a somewhat divergent function from the mammalian orthologue.[290] It is also important to consider that the various plasma membrane syntaxins in mammalian cells also represent lineage-specific expansion, and therefore exocytic diversification is a phenomenon that has occurred multiple times in eukaryotes.[280]

The VAMP7 family of proteins, including Vamp2, Vamp7, and Vamp8 in human cells, are exocytic R-SNAREs, but can also work in the endocytic pathway.[291,292] Recently, two ancient SNAREs were identified: the Qb SNARE Novel Plant Syntaxin (NPSN11) and the Qc SNARE Syntaxin of Plants 7 (Syp7).[91,293] These proteins have both been proposed to work in the plasma membrane in *Arabidopsis*, are highly conserved in eukaryotes, but are lost independently in both Fungi and at the base of the Holozoa. Although their function outside of plants is unclear, it is possible that they are the exocytic Qb and Qc SNAREs in most of eukaryotes, and Qbc SNAREs function in parallel with them or in their place. The MTC responsible for plasma membrane trafficking is the exocyst, and the SM protein is Sec1/Munc18.[294,295] The exocyst subunit Sec6 binds multiple subunits of the exocytic SNARE complex in yeast.[296] Exocyst is a plasma membrane tether in mammalian cells, but it has complex behavior in other eukaryotes. In plants, Exocyst has been shown to function in cell growth in land plants, and furthermore it includes a subunit (Exo70) which is encoded in 8 to 47 copies in these organisms, in an otherwise normal complex.[297,298] Furthermore, in trypanosomes, the exocyst complex includes a ninth subunit, Exo99, and functional work has shown that this complex is involved in endocytosis rather than

exocytosis.[299] These findings indicate another arm of the membrane trafficking system that is comparatively plastic.

Within the endocytic system, SNAREs can promote vesicular trafficking, as in recycling endosome-to-TGN trafficking, as well as the homotypic fusion of early endosome, and the fusion of endosomes and lysosomes. Endosomal recycling to the TGN relies on the SNAREs syntaxin 6, syntaxin 16, Vti1a, and Vamp4 (or general VAMP7-family protein outside of human cells).[300] Syntaxin 6-mediated TGN-endosome trafficking is also conserved in the apicomplexan *Toxoplasma gondii.*[301] The GARP MTC is responsible for tethering vesicles at the TGN, and the related EARP complex tethers early endosome vesicles prior to fusion at recycling endosomes.[302,303] Rab4, Rab11 and Rab35 are associated with recycling endosomes.[304] In human cells, RabGAPs TBC-N (TBC1D12) and TBC-E (TBC1D13) activate Rabs 11 and 35, respectively.[257,260] The TRAPPII complex acts as a GEF for Rab11,[305] while the DENND1 subfamily of proteins regulate Rab35.[275] The early endosomal Rab5 is a characteristic early endosome marker protein, while Rab7 is associated with late endosomes. It is the replacement of Rab5 by Rab7, known as 'Rab conversion', which defines the early-to-late endosome transition.[121] Both rabs interact with the retromer coat complex.[306] The only clear GAP of Rab5 and Rab7 is the TBC1D2 protein, which is animal-specific;[307] therefore the GAP for these Rabs is unknown in the majority of eukaryotes. The Vps9 family of proteins act as GEFs for the Rab5 subfamily.[308] An analysis of the evolution of the Vps9 family in eukaryotes was performed (Herman *et al.* in preparation), but was not included in this thesis. The Vps39 subunit of the HOPS MTC is the GEF for Rab7, as is the Mon1-Ccz1 complex in yeast cells.[309,310] Rab5 to Rab7 conversion is a common process found in organisms across the tree of eukaryotes.

The MTCs that function at early and late endosomes are CORVET and HOPS, respectively.[311,312] These complexes share four subunits, known as the VpsC core. Complex identity comes from two additional subunits via their binding either Rab5 or Rab7: Vps3 and Vps8 in the case of CORVET, and Vps39 and Vps41 in the case of HOPS. Complex assembly likely requires the rab-specific subunits, as the VpsC core cannot be isolated alone. However, intermediate forms have been identified.[240] Homotypic late endosome fusion involves the SNAREs syntaxin 7, syntaxin 8, Vti1b, and VAMP8, which likely interact with the Vps33 subunit of HOPS.[313–315] Vps33 also serves as an SM protein in both late endosome (LE) and LE-lysosome fusion. HOPS facilitates this fusion by binding multiple SNARE proteins.[314]

There are several other components of the membrane trafficking system that have more general function in membrane trafficking. These include p97 and Vps34. p97 was originally thought to be a paralogue of the distantly related NSF, which disassembles SNARE complexes. However, it is now known that p97 does not interact with SNAREs, but rather has multiple functions throughout the cell.[316] In addition to being involved in ER-associated degradation at the ER, p97 associates with the long-range tether EEA1 to control early endosome size.[317] Vps34 is a class III PI 3-kinase, which modulates endocytosis through its interaction with EEA1 and Rab5,[318,319] and late endosome trafficking through Rab7.[320]

1.4 The Role of Genomics and Transcriptomics in Evolutionary Protistology

Now that the topics of eukaryotic diversity and the membrane trafficking system have been introduced, the next question is how eukaryote genome and transcriptome can be used to understand the membrane trafficking system. In this section, the roles of these –omics techniques

are explored in how they contribute to our knowledge of organismal biology and membrane trafficking system evolution.

1.4.1 Filling taxonomic sampling gaps

Genome sequencing is a relatively recent method in cell biology. The first non-viral sequenced genome was of the bacterium *Haemophilus influenzae* in 1995, followed by the first fungal genome (*Saccharomyces cerevisiae*) in 1996, the first plant genome (*Arabidopsis thaliana*) in 2000, and a draft version of the human genome in 2001. The first sequenced protist genomes also appeared at this time, and the focus was on parasites: the fungal parasite of humans *Encephalitozoon cuniculi* was sequenced in 2001, and *Plasmodium falciparum*, the causative agent of malaria, was sequenced in 2002. The desire to understand the genetic blueprint of the organisms that kill millions of people per year no doubt motivated this early research, however the next step was improving the taxonomic breadth of sequencing efforts.

Several major outcomes were born of the work to sequence genomes from across the tree of eukaryotes. First, it provided further evidence that parasitism had evolved independently from many free-living lineages across the eukaryotic tree. Second, genome sequencing and improved phylogenetic methods were instrumental in informing a taxonomy based on many individual phylogenies of eukaryotes (published as Adl *et al*. 2005).[321]

As more eukaryotic genomes from across the tree were sequenced and comparative genomics of cellular systems were performed using these data, it became overwhelmingly clear that the Last Eukaryotic Common Ancestor (LECA) was incredibly complex. Although not an exhaustive list, these systems include a complex nuclear pore complex, cell division machinery,

ubiquitin signaling system, and of course, membrane trafficking system.[8,322–325] This assessment was based on finding homologous cellular machinery in organisms across the tree of eukaryotes, thereby indicating that it was likely present in the LECA. Observing this broad maintenance of trafficking factors through sequencing and comparative genomics raised the question of whether these factors are functionally homologous to those in animal and yeast cells. Assessing to what extent functional homology is preserved across eukaryotes is critical to creating a generalized model of eukaryotic membrane trafficking, which has implications for understanding how this system functions in organisms of ecological and medical relevance, and even understanding human cell biology and disease.

Despite the usefulness of these efforts to resolve the homology-function relationship between evolutionarily conserved membrane trafficking factors, they do not take into account our asymmetric understanding of membrane trafficking system function. There may well be genes important for membrane trafficking function in most eukaryotes that have been lost from animals and yeast, the most well studied model systems. Addressing the problem of asymmetry has only recently begun, but two clear examples of MTS components lost in animals are the TSET complex and the ArfGAPC2 protein mentioned in the previous section.[144,199] Both have patchy distributions in eukaryotes, but are partially or fully lost from the opisthokonts. Another example of a protein that, although present in animal cells, was a previously ignored membrane trafficking factor is Tom1 and related proteins. Tom1 is a patchy protein found across eukaryotes, and work in *D. discoideum* suggests that it functions as an ESCRT 0 analogue in ubiquitin-mediated cargo binding.[229] In addition to this role in non-opisthokonts, it is now thought to act alongside ESCRT 0 in animal cells.[228] It is likely that more examples of ancient membrane trafficking machinery will be uncovered, but this will require functional work in other

model systems. This is again supported by genome sequencing and comparative genomics, as the taxonomic distribution of novel trafficking factors is key to determining whether they are a lineage-specific innovation, or whether they are part of a general model of trafficking in eukaryotes.

At this point, much of the foundational membrane trafficking comparative genomics has been done, in the sense of having multiple sequenced representatives from nearly all of the major taxonomic groups. Some of the last major lineages without representative genomes were the Rhizaria and the cryptophytes. The first rhizarian organism sequenced was the marine alga *Bigelowiella natans*, and the first cryptophyte was *Guillardia theta*, and their genomes were published by Curtis *et al.* in 2012.[46] They were sequenced together, partly because they were from previously unsampled groups, but also because they both have nucleomorph genome-containing secondary algal endosymbionts, to which host proteins must be targeted if they are to be maintained. E. Herman and M. Klute performed a comparative genomic analysis of the membrane trafficking machinery in both organisms, published in Curtis *et al.* 2012[46] and Schlacht *et al.* 2014.[326] While this work is considered supplementary to the thesis, and is not included here, it represents one of the last analyses of the "firsts" of major clades. Another such analysis that is included in this thesis is comparative genomics of the membrane trafficking system of *Emiliania huxleyi*, the first haptophyte to be sequenced. Following the genome project (published in Read *et al.* 2012),[327] work in this lineage has expanded to include the genomes of other haptophytes and address how the membrane trafficking system supports the secretion of large calcium carbonate scales from these cells.

This early genome sequencing work and comparative genomics led to the identification of the complement of membrane trafficking factors that were likely to have been present in the

LECA. These factors and their role in trafficking are shown in Figure 1.2. This LECA complement therefore serves as a null hypothesis, from which lineage-specific variation in the form of losses or duplications can be identified.

1.4.2 Probing lineage-specific innovation

Now that the taxonomic breadth of eukaryotes has been sampled, the focus of sequencing has again shifted. Concomitant with this is an improvement in sequencing technology – a shift from Sanger to Next Generation Sequencing technologies – and better tools for sequence assembly and genome analysis.[328] The capacity to sequence multiple genomes as part of a single experiment allowed questions about lineage-specific biology to be asked.

The type of question being asked determines how much genomic sequencing is necessary to answer it. A comparison of a parasite and its free-living sister taxon, for example, requires a minimum of two sequenced genomes (although sequencing multiple strains could reveal hidden diversity). The Broad Institute's project on the Origins of Multicellularity, on the other hand, involved at least six newly sequenced genomes of basal animals, fungi and their close relatives, in addition to others that were already available.[22] Other groups have undertaken even more sequencing-heavy projects, such as the Joint Genome Institute's 1000 Fungal Genomes Project.[329] The purpose of these is to probe deeply within a single clade, in order to uncover lineage-specific patterns of genome evolution. Projects requiring multiple sequenced genomes run the gamut in terms of taxonomic breadth and depth, and can take advantage of previously generated sequence datasets that can be repeatedly queried as part of new projects.

The projects in this thesis are generally structured as comparative genomics of a large cellular system (~300-400 genes) between several organisms. Of course, the inverse approach to comparative genomics can be taken, where one can ask how a small, single gene family has evolved across many of sequenced eukaryotes. These types of analyses have been performed for highly paralogous families of membrane trafficking system proteins, such as the Rabs, TBC domain-containing RabGAPs, and ArfGAPs.[199,263,330] These works reconstruct the paralogue complement present in the LECA, and chart lineage-specific innovations throughout eukaryotes. Often, they identify previously unknown subfamilies, such as the ArfGAPC2 family described above. Another example of an analysis of a paralogous gene family is that of the Vps9 domain-containing proteins (Herman *et al.* in preparation, not included in this thesis). Vps9 domain-containing proteins are GEFs for the Rab5 subfamily, and are key regulators of the endo-lysosomal system in human and yeast.[269,331] The analysis showed that Vps9 family proteins are present across eukaryotes, and there were at least three ancient Vps9 domain-containing proteins in the LECA. It also tracked the acquisition of domains previously thought to be animal-specific, showing that they are in fact more ancient.

Comparative genomics is used in many fields of evolutionary study outside the membrane trafficking system. Another example of this type of work is the evolutionary analysis of kinetochore proteins published by van Hooff *et al.* (2017).[332] Kinetochores are the protein complexes that attach to centromeres and microtubules during cell division, and ensure that each daughter cell gets a single sister chromatid from each pair. Because kinetochore compositions have diverged while retaining their original function, the group decided to query all 70 kinetochore proteins in 90 genomes. Strangely the complement of kinetochore machinery of the LECA reconstruction did not resemble that of extant eukaryotes; the process of kinetochore

evolution largely took place by gene loss and duplication. However, they did also identify complexes where subunits coevolved, although in different manners. While the pattern of retention of the kinetochore machinery in eukaryotes suggests that it was present in the LECA, the authors state that the wide-scale losses they observed speak to the surprising flexibility of an essential biological process.

Regardless of the cellular system, these types of analyses of gene family evolution give insight into the complement of these genes present in the LECA, while also delving into lineage-specific variation that can be relevant for human cell biology.

1.4.3 Going beyond prediction in functional genomics

Comparative genomic analyses are snapshots of a cellular system in two or more organisms. They are static, and they allow us to make predictions about cell biology within a lineage. However, they do not tell us about how (or whether) the genes are expressed to support different cellular processes. Initially, the expression of specific genes could be tested under various conditions using Northern blots. However, it requires that genes are chosen *a priori*, and cannot be reasonably performed for every single gene in a eukaryote. Other methods that report the transcriptomic landscape of a cell have since been developed, such as microarrays and RNA-Seq. With microarrays, gene expression is detected by the binding of fluorescently labeled sequence (the sample) to oligonucleotide probes. This requires generation of a set of probes based on the organism of interest's predicted proteome. In RNA-Seq, next-generation sequencing technologies (e.g. Illumina's sequencing by synthesis) are used to sequence a cDNA library generated from mRNA, rather than genomic DNA as is the case in genome sequencing. In addition to being a quantitative measure of gene expression, RNA-Seq also reveals isoform

sequence and expression, which can be functionally relevant, fine-tuning of activity in a cellular system.

Transcriptomic analyses provide a wealth of functional data that goes beyond gene presence and absence. Most obviously, significantly differentially expressed genes under one condition but not another can be correlated with the biology of that condition. Transcriptomic analysis of a complex process over time can give understanding to phases of that process that are marked by gene expression changes. This is a major subject of the thesis, as transcriptomics is a common technique used to study gene regulation during a developmental process. One example is a transcriptomic assessment of the development of the aggregative, multicellular form of *Dictyostelium discoideum*.[333] This approach allowed the authors to identify the specific genes that are differentially expressed at each timepoint; DNA processing and mitosis genes are associated with the process of becoming multicellular, while organic signaling molecules and cofactor biosynthesis plays a larger role in the fully formed aggregate. Furthermore, these data gave insight into temporal dynamics and other nuances of the aggregation process. It was found that during aggregation, gradual changes in gene expression were punctuated by instances of rapid gene expression changes, followed by rapid phenotype changes. Gene expression patterns were highly complex, and as different parts of the multicellular aggregate developed, they did so in a synchronous way. This is just one of countless uses of transcriptomics to better understand cellular function.

Furthermore, transcriptomes can help pinpoint pseudogenes, identify splicing isoforms, and find transcription start sites. In some organisms that use an alternate genetic code or unusual splicing boundaries, transcriptomes will have more accurate gene models than *ab initio* predictions, and can aid in gene prediction. Furthermore, transcriptomes can be used as part of

phylogenomic projects, where tens to hundreds of genes are concatenated for phylogenetic analysis in order to improve the resolution of species trees. Because there is a length-limited amount of phylogenetic information in a single gene sequence, reconstructing deep relationships with single genes is generally not possible. However, concatenating many genes increases the information content of the 'supergene', improving phylogenetic resolution.

Genomic and transcriptomic analyses that are performed together work synergistically to provide more computational data than either approach could do alone. The limitations of a genome without any gene expression data have been discussed above. Additionally, the transcriptome of an organism generated for phylogenomics (i.e. not necessarily comprehensive) requires less sequencing power than does generating even a minimal coverage genome. Transcriptomes are typically easier to assemble than genomes, as they lack genomic repeat regions. A transcriptome alone is, by its nature, incomplete, and therefore any comparative analyses cannot speculate on potential gene absence. However, gene expression analyses that take into account genomic data can be highly informative. For example, for a given set of differentially expressed genes, with the corresponding genome, one can know the expression levels of all paralogues of a differentially expressed gene, giving a more complete view of how the system of interest is modulated.

One drawback of transcriptomics is that it measures mRNA abundance, not protein abundance. Factors that influence protein abundance post-transcription can include mRNA translation rate, protein stability, and targeting of proteins for degradation. While there has been debate about the correlation between static mRNA and protein levels,[334] differential gene expression has been shown to be related to changes in protein abundance.[335] Detecting changes

in relative protein content under different conditions must be done using proteomics methods, which are outside the scope of this thesis.

1.4.4 Patterns of evolution within the eukaryotic membrane trafficking system

As the membrane trafficking system is a key eukaryotic cellular feature, genome projects of microbial eukaryotes have often included comparative genomic analyses of this system, some which are included in this thesis. This affords a unique opportunity to observe patterns in how parts of the system evolve across eukaryotes. Wideman and colleagues (2014) published this type of meta-analysis with a focus on the endocytic system.[336] In this work, they explore a generalizable model of the endocytic system, and highlight lineage-specific endocytic mechanisms across the tree of eukaryotes. One of the major conclusions of this work is that across eukaryotes, there are multiple cases of lineage-specific mechanistic novelty within the endocytic system that appear at first glance to be general conservation of function. The authors state that these are examples of convergent evolution and/or coalescence of function, and suggest that they "… may indicate some constraints on how these various trafficking steps can operate, as well as suggest that aspects of cellular physiology may be even more divergent between supergroups than we have assumed or imagined." This type of assessment can only be made by considering the results of comparative genomic and phylogenetic analyses from an extensive repertoire of eukaryote genomes.

Because of the wealth of sequence data available, sensitive comparative genomic and phylogenetic techniques, and the growing understanding of membrane trafficking system function in non-opisthokont model organisms, meta-analyses of trafficking pathways as a whole are now possible, and is a major feature of the Discussion.

1.5 Focus of this thesis

The focus of this thesis is two-fold. First, it is an analysis of the evolution of the membrane trafficking system across eukaryotes. Comparative genomics and transcriptomics are used in tandem to study the array of cellular processes that are underpinned by membrane trafficking, and how trafficking genes are regulated during these processes. Second, this thesis is organized by work of increasing informational content; the first Results chapter contains comparative genomics, the middle chapters contain syntheses of comparative genomics and transcriptomics, and the final Results chapter contains the sequence of multiple genomes and transcriptomes, addressing not just membrane trafficking but a large-scale analysis of the cell biology of pathogenicity in a free-living amoeba. It is structured this way to show how combining more data and multiple types of data gives us a deeper and more complex appreciation of eukaryotic biology and membrane trafficking evolution.

The papers published by Dacks and Doolittle (2001)[5] and Dacks and Field (2004)[337] were the first comparative genomic analyses of membrane trafficking system components across eukaryotes. Since then, our understanding of membrane trafficking evolution has grown by addition of new protein families and new genomes, and includes the work discussed in this thesis, which encompasses organisms and lineages from across the tree of eukaryotes (Figure 1.1). Furthermore, the inclusion of membrane trafficking gene expression data affords insight into how this system functions in eukaryotes with unique lifestyles.

1.5.1 Membrane trafficking evolution and the transition to parasitism/endobiosis

Chapter 3 is a study of the evolution of the membrane trafficking system in three organisms of the Stramenopiles (SAR clade), *Blastocystis* sp., *Proteromonas lacertae*, and *Cafeteria roenbergensis*. *Blastocystis* sp. and its sister taxon *P. lacertae* are enigmatic residents of animal guts, and there is evidence that *Blastocystis* sp. is a parasite of humans and animals.[59] However, because of the unclear nature of their relationship with animals, both organisms can be referred to as 'endobionts', which does not preclude a parasitic lifestyle. *C. roenbergensis* is the outgroup, and is a free-living marine organism. The main focus of this chapter is addressing changes to the complement of membrane trafficking genes in this lineage that may be associated with being endobionts of animals. This includes both changes associated with parasitism or symbiosis (genome streamlining, reliance on host factors), and changes associated with the low-oxygen environment of the gut. It has already been shown that *Blastocystis* sp. has a type of degenerated mitochondria (a mitochondria-related organelle, or MRO), which is typically associated with anaerobic and parasitic organisms. The signature of living in these environments may be visible in the trafficking complement of the gut-associated organisms, when compared with the free-living *C. roenbergensis*.

There are other reasons to study membrane trafficking in this lineage outside of the question of parasitic/symbiotic adaptation. As shown in Figure 1.1, these lineages are part of the Opalozoa, one of the most basal clades of Stramenopiles. Analyses of this lineage can therefore give insight into the origin of Stramenopile-specific losses or gene family duplications in the membrane trafficking system. Secondly, there are morphological and ultrastructural differences between these organisms – one of which a large central vacuole only in *Blastocystis* sp. [338] – that may have implications for membrane trafficking gene complements in these organisms.

In this first project, these questions are addressed using only comparative genomics. The genomes of three *Blastocystis* sp. subtypes were included in analysis, as well as the *P. lacertae* genome, and the *C. roenbergensis* transcriptome. As the transcriptome is the only dataset available for *C. roenbergensis*, the loss of membrane trafficking components cannot be evaluated in this organism. Overall, these five genomes and transcriptomes allow predictions to be made about changes to membrane trafficking function – via gene presence and absence – and how they might be related to the shift to an endobiotic lifestyle or other differences between these organisms.

1.5.2 Microbial eukaryotes with unique behaviours: linking membrane trafficking system genomics and function

Chapters 4 and 5 of this thesis expand to include comparative genomics of the membrane trafficking system and transcriptomics in relation to two distinct secretory events. These are the encystation process in the parasitic amoeba *Entamoeba* spp., and the secretion of calcium carbonate scales and organic body scales by four genera of haptophyte algae.

Entamoeba histolytica is a human gut parasite, while its close relative, *Entamoeba invadens,* infects reptiles and amphibians. *E. invadens* is often used as a proxy to study the infection of *E. histolytica*, as *E. invadens* can be induced to form cysts *in vitro* while *E. histolytica* cannot.[339] Encystation is highly relevant to pathogenesis, as only the cyst form is infective. Studying the dynamics of cyst formation may therefore give insight into how this process can be prevented by therapeutic measures, thus reducing parasite spread. The membrane trafficking system has been studied in relation to virulence in *E. histolytica*, and while there have been analyses of gene expression during encystation,[340,341] the membrane trafficking system has

not been studied specifically during this process. However, it clearly underpins encystation, which involves various lectins and chitin being transported to the cell surface.[342–345] Analysis of the membrane trafficking system gene complement was performed to assess the similarity of the gene complements of *E. invadens* and *E. histolytica*, in order to establish whether *E. invadens* is a good proxy for *E. histolytica* specifically in this cellular system. It also allows comparisons between *Entamoeba* and *Blastocystis* sp., as they are both gut parasites facing similar environmental pressures.

A transcriptomic analysis was performed on *E. invadens* grown under encystation-inducing conditions, at four timepoints. Because only one sample was sequenced per timepoint, comparisons of individual gene expression changes could not be done reliably. Instead, clusters of similarly expressed genes were generated, and the trafficking steps relevant to encystation were deduced by determining which steps had multiple constituent genes found in clusters whose expression increases during encystation.

The third Results chapter includes the work of two projects; first, an analysis of the membrane trafficking system in *E. huxleyi*, the first haptophyte organism with a sequenced genome. The purpose of the second project was to understand the cellular dynamics of biomineralization in the haptophytes. All haptophytes extrude scales which form a type of cell wall, while some species generate calcium carbonate scales.[346] This biomineralization has far-reaching environmental consequences, as haptophytes are critical players in carbon cycling.[347] Micrographic evidence shows that scales are likely formed in a Golgi-related organelle, and in some calcifying haptophytes, a membranous 'reticular body' forms near the growing scale and is involved in calcification.[348] To explore the effect of scale secretion and biomineralization on the membrane trafficking complement, comparative genomics of four haptophyte species was

performed: *Emiliania huxleyi* and *Gephyrocapsa oceanica*, biomineralizing haptophytes, the sister taxon *Isochrysis galbana*, which secondarily lost the ability to biomineralize and secretes organic body scales,[346] and *Chrysochromulina tobin*, which is more distantly related and also secretes organic body scales.

RNA-Seq was performed on *E. huxleyi* and *G. oceanica* grown under scale-forming conditions – the addition of calcium, bicarbonate, and both calcium and bicarbonate – and compared with cells grown with no calcium or bicarbonate. Genes involved in scale formation should be co-regulated in both organisms. Three replicates of each condition were performed, meaning that the expression of individual membrane trafficking genes in scale formation and biomineralization can be assessed. Not only do these analyses give insight into how membrane trafficking processes support scale secretion, but they can also give specific information about which genes and paralogues are involved in this process.

In both cases of encystation and scale formation, the secretory system is likely to be critical. Although there may be gene family expansions of secretory factors that are relevant to these processes, it is the gene expression analyses that show how the trafficking system is regulated. It furthermore pinpoints specific differentially expressed genes, which effectively generates a list of factors that could be targeted in future experimental work (e.g. *Entamoeba*-specific encystation factors as therapeutic targets).

1.5.3 Using comparative genomics and transcriptomics to explore the evolution of pathogenesis in the free-living neuropathogenic amoeba *Naegleria fowleri*

The final chapter of the Results is an analysis of pathogenesis in *Naegleria fowleri*. Rather than focusing solely on the membrane trafficking system, this chapter serves as an example of the information that can be gleaned from analyzing whole genomes – both large-scale features and multiple cellular systems – in order to understand host infection. Unlike the parasites *Blastocystis* sp. and *Entamoeba* spp., *N. fowleri* is a free-living amoeba that does not require a host. However, the cellular factors that give it the ability to infect the brains of humans and animals are only partially known.[349–359] The genome of a non-pathogenic species, *N. gruberi* is compared with three *N. fowleri* strains, in order to identify differences that may be relevant for pathogenesis.

To get a comprehensive picture of the genes differentially expressed in association with pathogenesis, RNA-Seq was performed on *N. fowleri* grown in culture and *N. fowleri* passaged through mice, which is known to be more virulent.[360] Genes that are differentially expressed in highly pathogenic, mouse-passaged *N. fowleri* are likely to be related to pathogenesis, as pathogenicity factors, or regulated in response to host infection. Combining these data with the genomic comparisons can again give insight into the gene families that are involved in pathogenesis and the specific paralogues that could be targeted for therapeutics. These analyses are expanded beyond trafficking machinery, as pathogenesis is likely to be a multi-factorial process that involves multiple cellular systems.

The *Naegleria* genomes and transcriptomes also provide a stepping-stone for functional work to understand the basic cell biology of *Naegleria*. A stacked Golgi has never been observed in *Naegleria*, raising the question of whether this lineage has a Golgi body, and if so, what is its morphology. In the *N. gruberi* genome paper, homologues of Golgi-related membrane trafficking genes were identified.[361] In this chapter, the transcriptomic data in *N. fowleri* is used to show that

orthologues of these genes are expressed. Furthermore, antibodies to a Golgi body marker (NgCOPB) are generated and localized using immunofluorescence and immunoelectron microscopy, showing the presence of Golgi-like organelles.

Together, these comparative genomic and transcriptomic analyses illuminate membrane trafficking evolution in eukaryotes: how it underpins unique, lineage-specific cell biology, and how it can inform a better functional model of this system in eukaryotes.

2. Methods

2.1 Introduction

The methods used in this thesis are largely computational, with the exception of molecular biological work in Chapter 6. As such, the work in this thesis relies on sequence data originally generated by collaborators, who are acknowledged in the Specific Methods sections of each chapter. E. Herman performed downstream data quality checking and analyses, unless otherwise stated. This Methods chapter gives an overview of the techniques and software used throughout the Thesis, with the parameters that are generally used. Any special techniques or changes to these parameters are noted in Specific Methods.

2.2 Homology Searching

2.2.1 BLAST

BLAST, or Basic Local Alignment Search Tool, was first published by David Lipman's team at the National Center for Biotechnology Information in 1990, and is arguably the foundational method of the past three decades of comparative genomics research.[362] Briefly, BLAST works by identifying high-scoring segment pairs (HSPs) between the query sequence and sequences in the database. To maximise both sensitivity and speed, it uses a heuristic that approximates the Smith-Waterman algorithm, which finds the best local alignment given the nucleotide or amino acid substitution matrix and gap penalty scheme. BLAST produces a list of hits potentially homologous to the query sequence, alignments of the HSPs, and a scoring metric called an Expect value (E-value) for each hit and HSP. The E-value is interpreted as the number of hits one would expect to see with the same score purely by chance, given the size of the database. Therefore, as the E-value approaches zero, the likelihood of the query and hit being homologous increases. BLAST+ versions 2.2 through 2.6 were used in this Thesis.[362,363] Each program in the BLAST suite performs a different type of search: BLASTP uses a protein query

to search a protein database, BLASTN uses a nucleotide query to search a nucleotide database, TBLASTN uses a protein query to search a translated nucleotide database, BLASTX uses a translated nucleotide query to search a protein database, and TBLASTX uses a translated nucleotide query to search a translated nucleotide database. Default parameters for word size and gap open and extension costs were used for all BLAST searches, as was the Blosum62 scoring matrix.[364]

To infer homology from BLAST searches, the following methodology was used. Query sequences were used to search a database of interest using one of the BLAST+ programs. Candidate hits were further considered if they were retrieved with and E-value better (less) than 0.05. Candidate hits were then used as queries to search the sequence database of the query organism. To be considered homologous, the reciprocal search must retrieve the query or a clear orthologue as the top hit with an E-value also less than 0.05 with >2 orders of magnitude difference between it and the next non-orthologous protein, and the HSP(s) should extend along most of the protein. However, the length of the HSP relative to the query and subject sequences is often short when the query and subject taxa are distantly related. A hit retrieved with an E-value of 0.05 means that one can expect by random chance 5/100 hits with an equal or better score, given a database of that size. This cutoff is deliberately lax to cast a wide net to collect potential homologues. In cases where homology is unclear because the forward and reciprocal search E-values are near the cutoff, the hits are noted as 'potential' homologues, and further techniques are used to find evidence of homology, such as domain prediction (discussed below) or searching the non-redundant database. The NCBI's non-redundant database is a comprehensive, non-redundant set of sequences from all available sources (e.g. organelle genomes, chromosomes, contigs, mRNAs, proteins, etc.). Using potential homologues to search

the non-redundant database and retrieving clearly annotated homologues of the initial query sequence as the top hits provides additional evidence for sequence homology between the initial query and the potential homologue.

BLAST searches were run both on external websites, mainly the National Center for Biotechnology Information and the Joint Genome Institute, as well as a local computer cluster. J. Barlow, L. Barlow, and E. Herman envisioned and wrote a set of programs collectively called Monkey House to increase standalone BLAST efficiency. Monkey House allows multiple query files and multiple BLAST databases to be searched at once, retrieves the accession numbers of all hits, retrieves their sequences from BLAST databases, and populates the 'query' directory with these sequences for fast reciprocal BLAST searching. Monkey House was used for all standalone BLAST searches. The scripts directing BLAST searches are written using the Bash shell language, while the Hunter-Gatherer script is written in Perl. faSomeRecords is a program that was originally downloaded from the UCSC Genome Browser website (http://hgdownload.cse.ucsc.edu).

Briefly, Monkey House works in the following way. A dispatch script (dpat) executes the BLAST program of choice iteratively for each query file (containing one or more queries) in a 'query' directory searching each genome in a 'database' directory. The results are outputted in tabular format in a 'results' directory (one result file for each query file-genome pair). The Hunter-Gatherer script then extracts the accession numbers for all hits in all files, and uses them to retrieve the corresponding sequence from the initial BLAST database, making individual files with these sequences in the 'query' directory. To do a reciprocal BLAST, the database of interest (i.e. of the original query organism(s)) is added to the 'database' folder, and the dispatch script can then be used to BLAST search with the new queries and new databases. Monkey House

works with all standalone BLAST tools as long as the tabular output format is specified with the options: -outfmt "6 qacc sacc pident evalue". It can also be used with HMMer, pHMMer, and jackHMMer. L. Barlow and J. Barlow wrote the comparative genomic and dispatch scripts, and E. Herman wrote the Hunter Gatherer program. Directory structure and code available in Online Appendix File 2.1.

Up-to-date BLAST databases were downloaded from the NCBI (https://www.ncbi.nlm.nih.gov) in the case of *Homo sapiens*, *Saccharomyces cerevisiae*, *Arabidopsis thaliana, Blastocystis* ST4, *Blastocystis* ST7, and *Chrysochromulina tobin*. The *Emiliania huxleyi* CCMP1516 database and the *Naegleria gruberi* database were downloaded from the JGI (https://jgi.doe.gov/). The *Entamoeba invadens* and *Entamoeba histolytica* genomes were downloaded from AmoebaDB (http://amoebadb.org/amoeba/). All other BLAST databases were generated from unpublished data, or their sources are listed in metadata files for individual analyses.

Query sequences were chosen as functionally characterized membrane trafficking system components in *H. sapiens*, *S. cerevisiae*, or *A. thaliana*, and *N. gruberi*. Occasionally query sequences from a close relative of the organism of interest were used, provided those sequences were clear orthologues of functionally characterized version of the protein.

2.2.2 HMMer

HMMer is a more sensitive homology searching program than BLAST, as it relies on a model generated from a multi-sequence alignment of homologues to search a database.[365–367] The HMMer version 3 was used in all analyses.[367] After aligning a set of homologous sequences, one

arm of the HMMer program, hmmbuild, generates a hidden Markov model (HMM). The HMM is a probabilistic model which takes into account the amino acid frequencies at each position of the multi-sequence alignment, and patterns that occur over multiple columns of the alignment. It also describes insertions and deletions as states between which transitions occur, and the probabilities of these transitions can also be modelled. The HMM is then used by hmmsearch to search a sequence database. Like BLAST, the hmmsearch outputs E-values and scores for each identified domain as well as the full sequence.

The following methods were used to infer homology from HMMer searches. To generate a multi-sequence alignment, sequences determined to be homologous to a functionally characterized sequence (by an initial round of BLAST searches) were aligned using the alignment program MUSCLE.[368] Hmmbuild and hmmsearch of HMMer were then used, respectively, to build the HMM and search a sequence database. Candidate hits were considered if the E-value calculated for the entire sequence was less than 0.05. In order to confirm the candidate hits, these sequences were then used as BLAST queries to search the genome(s) of one or more sequences present in the HMM, preferably those where the sequence has been functionally characterized, as well as the non-redundant protein database. As with the BLAST methodology, the retrieved sequence is considered homologous if it retrieves one of the sequences in the HMM as the top hit, with an E-value less than 0.05, and with at least two orders of magnitude difference between it and the next non-orthologous protein. As HMMer is designed to detect weak sequence homology beyond the scope of BLAST, performing a reciprocal BLAST search may not retrieve one of the query sequences of the multi-sequence alignment, even if the sequences are truly homologous. Therefore, additional information was taken into account when determining homology. Sequences that failed to retrieve the query sequence or a

clear orthologue, but did have corroborating evidence of homology from the NR BLAST search, or the correct domain organization, were considered as 'potential' homologues.

All HMMer searches were run locally using default settings. The Monkey House programs were designed to perform multiple HMMer searches and retrieve sequences for reciprocal BLAST search.

2.2.3 Domain Detection

In order to predict domain architecture, sequences were used to search the Conserved Domain Database (CDD) using the NCBI's CD-SEARCH program,[369,370] or the Pfam database.[371] Pfam identifies domains using HMMer to search a database of HMMs of domains, while the CD-SEARCH uses a version of PSI-BLAST (RPS-BLAST) to scan a set of position-specific scoring matrices (PSSMs). PSSMs are similar to HMMs, in that they use a multiple sequence alignment where amino acids at each position are scored based on the frequencies of residues at that position. With RPS-BLAST, CD-SEARCH effectively uses the query sequence to search a database of PSSMs using a BLAST-based methodology. Searching in both the CDD and Pfam databases often yield similar results, but at times one will be able to identify domains that the other does not. Therefore, these methods were used in tandem when searching for domains in divergent sequences.

2.3 Phylogenetics

While homology searching identifies related proteins by sequence similarity, phylogenetics seeks to reconstruct the evolutionary relationships between gene or protein

sequences. Therefore, phylogenetics requires that all sequences used to generate a tree are homologous. Phylogenetic analysis relies on the inclusion of only the positions within each sequence that are evolutionarily conserved. Therefore, the first step is sequence alignment, and then masking and trimming the alignment to remove all sites of uncertain homology. Model testing is done to determine the best model of sequence evolution that fits the dataset. There are two methods that are considered state-of-the-art for testing evolutionary hypotheses using phylogenetics: Maximum Likelihood (ML) and Bayesian inference. Only clades with statistically supported nodes in phylogenies generated by both methods are considered to be robustly reconstructed. For visualisation purposes, node support values from ML trees are mapped onto trees generated by Bayesian inference to create a consensus tree, since Bayesian inference generates equally or more accurate relative branch length estimates compared to ML methods.[372]

2.3.1 Sequence alignment and model testing

Phylogenetics relies on homologous residues at each position. Because insertions and deletions can occur in homologous proteins that have diverged over evolutionary time, this necessitates alignment of the sequences, and trimming to remove non-homologous sites. All multiple sequence alignments for phylogenetics were performed with either MUSCLE v.3.8[368] or M-COFFEE (online server).[373,374] Generally, M-COFFEE was only run when MUSCLE-built datasets did not appear to be accurately aligned by visual inspection. MUSCLE builds a progressive alignment using a pairwise profile alignment approach, which is then iteratively refined. M-COFFEE also uses an iterative strategy, but unlike MUSCLE, it works by generating both global and local alignments, weighting them relative to sequence identity, and progressively aligning the closest sequences based on their weights. Both programs were run with default

parameters. In some cases when using MUSCLE, previously generated alignments were used to classify new sequences. These new sequences were aligned to the multi-sequence alignments using the –profile option.

Alignments were visualized in MacClade v.4,[375] or in its successor, Mesquite v. 3.03,[376] and manually masked and trimmed to exclude non-conserved positions and indels. The relevant part of alignments are the blocks of sites that are clearly well-aligned; within each site, >75% of sequences have an amino acid at that position, and the amino acids range from being identical to having similar properties. These have intervening indels, and are often flanked by unconserved regions. Alignments are masked and trimmed to remove these regions, and often both the first and last positions of conserved blocks, as well as the positions flanking indels, since there is room for error in aligning residues at these boundaries. Because missing data are not treated as gaps (i.e. deletions), then as long as there is sufficient informational content in the alignment, using a stringent mask will not negatively impact the robustness of the resulting phylogeny. All masked alignments are available upon request.

ProtTest v.2.4[377] and v.3.3[378] were used to determine the best fit model of sequence evolution for the trimmed alignment. The program considers the following substitution model matrices: JTT,[379] DCMut,[380] Dayhoff,[381] WAG,[382] Blosum62,[364] LG,[383] and VT.[384] The model LG[383] is the overwhelming model of choice for alignments in this Thesis; information about model choice is found in tree metadata files. Substitution matrices assume that substitution of amino acids is identical over all positions. However, this does not take into account constraints on sequence evolution. ProtTest is therefore also able to consider whether, in addition to the model, other information should be taken into account, such as the proportion of invariant sites (+I), the estimated rate of evolution of each site as the probability of belonging to a given rate

category (+G), and the observed amino acid frequencies (+F). ProtTest then compares different combinations of models and parameters using the Akaike Information Criterion, which, based on likelihood and number of parameters, will try to identify the model closest to the unknown conditions that generated the data.[385,386] However, AIC may select an over-parameterized model for smaller datasets; therefore, the corrected AIC (AICc) was occasionally used. AICc uses a different relative penalty score for model complexity, but the score converges with that of AIC as dataset size increases.[386] Model and parameter metadata are given for the trees in each chapter.

2.3.2 Bayesian methods

The heart of Bayesian inference is the probabilistic relationship between the hypothesis and the data. The posterior probability, which is the conditional probability of the hypothesis (all model parameters, e.g. the tree topology) given the data, is a measure of how well the observed data and the model agree.[387] Calculating the posterior probability depends on prior probabilities of the hypotheses (e.g. probability distribution for all selected parameters) and the likelihoods of the hypotheses given the data. The Markov chain Monte Carlo algorithm approximates the posterior probability landscape, which includes the tree with the highest posterior probability (at a global maximum) as well as all others.[388] Practically, two analyses are run independently, and should converge to produce two similar trees. These are run in parallel for a specified number of generations; in each generation, changes are made to the trees (e.g. branch length, topology), and accepted ones are those that 'climb uphill' in the posterior probability landscape (i.e. generate trees with a better log likelihood). However, this can lead a run to be stuck in a local maximum. This problem is remedied by swapping between four MCMC chains: one 'cold' MCMC chain and three 'hot' MCMC chains. In this sense, heating refers to flattening the posterior probability

landscape to increase the ability for a chain to find an isolated peak. In other words, this means allowing more unfavourable tree topology changes in order to get to a better overall topology. Swapping between the chains occurs at a specified frequency, which allows the cold chain to move between peaks more easily. At the end of the analysis, only the two cold chains are kept, and for each the initial trees that are sampled which have low log likelihoods are discarded. The two chains are then checked for convergence (i.e. how similar they are).

Bayesian phylogenies were generated using both MrBAYES v.3.2.0-3.2.2[387,389,390] and Phylobayes v.3.3f.[391–393] The main difference between the programs lies in how they handle site-specific evolution. In MrBAYES, one can specify a gamma distribution of the rate of substitution across all sites with four or more discrete categories. Phylobayes uses a more sophisticated CAT model, a non-parametric, site-heterogeneous mixture model that takes into account that different sites in an alignment will have different probabilities of evolving into another character.[391]

In all trees, MrBAYES was run for at least one million generations, with two independent runs and four MCMC chains. Specific information about models and parameters used is associated with each alignment in the supplementary files for each section. Trees were both printed and sampled every 1,000 generations. For each run, the log likelihood plateau was visualized *post hoc* by plotting log likelihood by the number of steps. In all cases, this plateau occurred well within the first 25% of sampled trees; however for consistency, a relative burn-in value of 25% was used. After burn-in, the trees from both cold chains were summarized, and the average standard deviation of the splits frequencies was calculated (i.e. whether or not the runs converged). Unless otherwise noted, all MrBAYES trees presented in this work are the result of runs that have converged, with a value <0.01.

Phylobayes also makes use of independent runs. Depending on the complexity of the dataset, either two or four chains were specified. Phylobayes does not require a specific number of generations to be set *a priori*, since it can be configured to automatically stop if the runs meet certain conditions. For analyses in this thesis, the conditions are running for a minimum of 100 cycles (generations), that the maximum discrepancies of summary variables (combinations of parameters) between the two runs is equal to or less than 0.1, and that the effective sample size for all summary variables is greater than 100. The effective sample size statistic takes into account the autocorrelation of parameter values across the cycles, and therefore a sufficiently high value indicates adequate mixing of chains. The default burn-in value of 20% was used for all trees. Finally, all trees were visualized using FigTree v.1.4.0. In interpreting phylogenies, statistically supported nodes must have posterior probabilities ≥0.8 for MrBAYES and Phylobayes methods. In all cases except for Figure 5.1 (where outgroup is known), trees are arbitrarily rooted.

2.3.3 Maximum-Likelihood methods

Maximum-Likelihood (ML) treebuilding works by identifying the phylogeny that makes observing the data most likely given the parameters, rather than the inverse, using the data to determine the most likely phylogeny.[394] This is because the former can be calculated by considering all possible datasets, while the latter would require considering all possible trees, which is not feasible. First, an initial tree is found using a rapid approximate method, and then initial model parameters are estimated. Then, the model parameters remain constant while search heuristics are used to identify alternative trees (based on rapid hill-climbing methods).[394] Generally, ML algorithms iterate between topology finding and optimizing model parameters

and branch lengths. Once the ML tree is found, confidence is assigned to the clades by bootstrapping. In bootstrapping, a specified number of pseudo-replicate datasets is generated by sampling sites in the alignment with replacement. For each bootstrapped alignment, ML estimation is performed and the clade similarities are summarised to generate an assessment of topology support. For the trees generated in this Thesis, summarising bootstraps is done by the majority rule – extended method. The most highly supported clades that are recovered in more than half of the pseudoreplicates appear in the consensus tree, followed by the next most frequently observed clades, and provided they do not conflict.

RAxML v.7.3-v.8.1.3[395,396] and PhyML v.2.4.4-v.3.1[397,398] were used as ML methods. In general, a gamma distribution is specified rather than CAT, since the CAT model used by RAxML – completely different from the one used by Phylobayes – is not appropriate for alignments with fewer than 50 taxa (RAxML website, https://sco.h-its.org/exelixis/web/software/raxml/hands_on.html). For PhyML, when estimating tree topologies, the -s BEST parameter was used, which uses both nearest neighbour interchange and subtree pruning and regrafting methods. A minimum of 100 bootstraps is specified. Consensus trees are visualised in FigTree, and statistically significant support for a given clade is a bootstrap value ≥50.

2.4 Methods only used in Chapter 6

The *N. fowleri* genome sequencing project was part of a collaborative effort between the Dacks Lab, the Chiu Lab at University of California San Francisco, the Marciano-Cabral Lab at Virginia Commonwealth University, and Dr. Govinda Visvesvara at the Centers for Disease Control and Prevention. Additionally, the *N. fowleri* ATCC 30863 strain was sequenced by D.

Zysset-Burri, M. Wittwer, N. Muller at the Spiez Laboratory, Federal office for Civil Protection, Switzerland. The *N. fowleri* 986 strain was sequenced by G. Puzon and T. Walsh at CSIRO Land and Water, Australia.

2.4.1 Genomics and Transcriptomics

A. Greninger in the Chiu Lab sequenced the *N. fowleri* CDC:V212 strain using a combination of Illumina HiSeq and 454 methods. After pre-processing, the mitochondrial genome and extrachromosomal plasmid were assembled using Geneious software. The final version of the nuclear genome was assembled using CLC Genomics Workbench (https://www.qiagenbioinformatics.com/).

Multiple transcriptomic analyses were performed on RNA extracted from *N. fowleri* V212 and *N. fowleri* LEE (ATCC 30894). To generate transcripts to be used for gene model prediction, *N. fowleri* V212 mRNA was extracted by the Chiu Lab and was sequenced using an Illumina HiSeq sequencer, and ~150 million paired-end reads were aligned to the nuclear genome using the program SNAP.[399] I used transcripts generated from these alignments in downstream gene prediction analyses.

F. Marciano-Cabral grew the *N. fowleri* LEE strain for pathogenicity-related analyses. *N. fowleri* LEE was grown axenically in oxoid media. Three replicates of *N. fowleri* LEE were passaged continuously through 50 $B_6C_3F_1$ male mice. After mouse sacrifice, amoebae were extracted and grown in axenic media for one week to clear the culture of human cells. mRNA was extracted from the three mouse-passaged cultures (high pathogenicity *N. fowleri*), and from three independent axenic cultures (normal pathogenicity *N. fowleri*). Illumina MiSeq sequencing

was performed at The Applied Genomics Centre at the University of Alberta, generating paired-end 2x300 reads.

2.4.1.1 RNA-Seq read processing

Read pre-processing is required to remove reads of low quality and trim low quality ends, and to trim adaptors. This quality control increases the number of reads that can be assembled together or that can map to a reference sequence, as sequence errors can prevent proper assembly. Reads from the pathogenicity-related transcriptomics experiment were pre-processed using Trimmomatic v0.32[400] and visualized using FastQC v0.11.2.[401] For each read, Illumina sequencing reports not only the read sequence, but also quality scores associated with each predicted nucleotide base. Several factors can impact quality scores, such as mRNA quality, fluorophore crosstalk during sequencing, the optics and instrumentation of the sequencer, and DNA polymerase fidelity.[402] Quality is measured by Phred score, which is logarithmically related to base-calling error probabilities.[403] The Trimmomatic parameters and outputs for each forward and reverse read set are found in Supplementary Table ST2.1. In general, adapters were trimmed using the TruSeq3-PE adapter file for reference. Then, the following 15 bp of all reads were trimmed to remove low-quality regions. A sliding window approach was used to trim reads once the average base prediction quality dropped below a Phred score of 20 over a window of 4 bp. Finally, reads that were less than 50 bp in length were discarded. Final FastQC quality analyses are available in Online Appendix Data 2.1-2.12.

2.4.1.2 Read mapping

In general, read mapping software works by performing local or end-to-end read alignment, scoring these alignments based on similarity and mapping quality (i.e. uniqueness in mapping positions), and then selecting read alignments that meet a specified score threshold. Pre-processed reads were mapped to the genome of *N. fowleri* V212 using the program TopHat v.2.0.10,[404] since TopHat is able to align reads that span exon junctions. TopHat is based on the related software, Bowtie2,[405] which handles read mapping to sequences without gaps. Bowtie2 performs end-to-end read alignment by scoring potential alignments with mismatch penalties and accepting alignments that exceed a threshold. The paired nature of paired-end reads is also taken into account during mapping, as is mapping quality (uniqueness of each mapping assignment). The *N. fowleri* LEE reads were able to be mapped to the *N. fowleri* V212 genome with relatively good mapping rates (see Supplementary Table ST6.2 in Chapter 6), since comparative genomic analysis of the three *N. fowleri* strains revealed very high gene sequence similarity. Because the reads from the *N. fowleri* LEE-MP (mouse-passaged LEE strain) and LEE-Ax (axenically grown LEE strain) samples had some genomic contamination, they were also mapped directly to the *N. fowleri* V212 predicted genes using TopHat, in order to get a more accurate gene expression values. For the HiSeq-generated reads, TopHat was run with default parameters. For the MiSeq-generated LEE-Ax and LEE-MP reads, the minimum intron length was changed to 30bp, based on an assessment of predicted genes. Otherwise, default parameters were used.

2.4.1.3 Transcriptome assembly

Transcriptome assembly using a reference sequence is similar to mapping reads to genomes, although exon boundaries generally do not need to be taken into account in the transcriptome. Cufflinks v.2.1.1[406] was used to generate transcripts from TopHat-aligned reads. In the case of the pathogenicity transcriptomic analysis, this was done using the Reference Annotation Based Transcript option,[407] using the predicted genes as a reference dataset. The reference transcripts are tiled with 'faux-reads' to aid in the assembly, and these sequences are added to the final dataset containing the newly assembled transcripts. The purpose of the HiSeq read dataset was to generate transcripts to train the Augustus gene prediction program; therefore, no reference annotation was used for the Cufflinks transcriptome assembly of HiSeq data. Default cufflinks options were used for both experiments.

For the pathogenicity transcriptomic analysis, the datasets of *N. fowleri* LEE reads mapped to the V212 predicted genes, and these comprised the bulk of 'transcripts'. Due to an issue with genomic contamination, and the fact that the *N. fowleri* genome has little intergenic space, attempts to assemble transcripts *de novo* or even using a genome-guided approach generated extremely long fused sequences of multiple genes. Therefore, reads were mapped to the V212 genes. In order to not lose any transcriptomic data that were not part of the V212 genome, two approaches were taken. To capture transcripts that correspond to regions of the V212 genome but were not predicted as genes, Cufflinks was used to generate transcripts based on *N. fowleri* LEE reads mapped to the V212 genome. This results in transcripts corresponding to gene predictions, as well as transcripts assembled *de novo*. Secondly, to capture LEE transcripts that have no corresponding genomic sequence in the V212 genome, the program Trinity (release 2013-02-25)[408,409] was used for purely *de novo* transcriptome assembly. Trinity

was run using a genome-guided approach, with a --genome_guided_max_intron value of 5000. Any LEE transcripts generated by Cufflinks (mapping to genome) and Trinity that did not have corresponding sequence in the V212 genome were added to the transcript dataset.

2.4.1.4 Differential expression analyses

To determine which genes were differentially expressed in *N. fowleri* passaged through mice versus grown in culture, reads from each biological replicate were mapped to an *N. fowleri* V212 gene dataset (plus additional sequence not found in the *N. fowleri* V212 genome), and the difference in gene expression between the two conditions was assessed. Differential expression (DE) analyses were performed for the pathogenicity transcriptomic data using the programs Cuffdiff[410] and Trinity.[409] Cuffdiff was first used to map reads to the final transcriptome assembly. It was run in two ways: (i) treating each replicate as a separate condition in order to get read-mapping data that could be passed to Trinity programs, and (ii) treating replicates as replicates for each condition, in order to get Cuffdiff-generated DE results. The purpose of passing Cuffdiff-generated read count data to Trinity was because Trinity includes of a suite of Perl scripts that can be used to assess quality, and because one of its options for running DE analyses is the program edgeR,[411,412] which performs slightly better than Cuffdiff in detecting true positives in DE analyses.[413]

In order to generate read counts, Cuffdiff was run twice; first with default parameters, and secondly, with a mask file (-M), containing the mitochondrial genome, extrachromosomal plasmid, and any sequence from the first Cuffdiff run with extremely high read mapping values (>10,000). In the second run, reads mapping to these sequences are filtered out from further read quantitation, as they can otherwise skew transcript abundance estimates. Read counts for genes

were used in downstream analyses, rather than transcript read counts. Although this ignores potential splicing isoforms, it reduces the possibility that true positive DE genes are missed because of the difficulties involved in accurately mapping reads to similar isoforms.

The Trinity Perl-to-R (PtR) toolkit was used to assess variation between replicates. This analysis showed that one of the LEE-MP replicates, MP2, was highly dissimilar to the other mouse-passaged samples (as well as the samples grown in culture), and was therefore excluded from further analysis. Cuffdiff was also re-run on the samples excluding the MP2 sample.

For DE analysis, the edgeR[411,412] program was run through the Trinity script run_DE_analysis.pl, using a dispersion value estimated from the dataset. The dispersion is an estimate of biological variability, which is determined using a quantile-adjusted conditional maximum likelihood method, since this is a case of a simple pairwise comparison between two conditions (LEE-Ax vs. LEE-MP). Using the calculated dispersion and a fitted negative binomial model, edgeR uses the exact test (similar to Fisher's exact test) to determine DE genes. These results were considered to be the set of DE genes. They were also compared to the DE results from running Cuffdiff, and found to be very similar, with only a few genes considered to be DE by edgeR but not Cuffdiff.

2.4.2 Gene prediction and annotation

Gene prediction refers to identifying genomic sequence that encodes genes. This is not a trivial process, as there can be high organismal variation in gene structure (e.g. intron boundaries), leading to errors in gene models unless the gene prediction program is trained using gene evidence from that organism or a close relative. A transcriptome can provide this evidence

to the gene prediction program, which was the strategy used here. Gene prediction was performed using the program Augustus v.2.5.5.[414] Augustus can perform both *ab initio* and transcript-guided gene prediction, and is one of the best-performing methods for gene prediction in terms of sensitivity and specificity for exon, intron, and gene prediction, and is one of the best for predicting novel splice sites.[415] Particularly important is the fact that transcript-level gene prediction is improved when Augustus is used with the input of experimental evidence. Augustus works by combining both intrinsic information (genome sequence) and extrinsic information (expressed transcripts) to generate a generalized hidden Markov model used to predict protein-coding regions.[416] The program rewards gene structures that are supported by extrinsic information and penalizes those that are not supported, in a user-defined way (i.e. the user can determine how important different aspects of the extrinsic information are to informing the gene model). The HiSeq dataset of assembled transcripts was used as extrinsic evidence, termed 'hints' in Augustus. Augustus was also trained for the *N. fowleri* genomes using a manually annotated 60kb segment from the *N. fowleri* V212 genome published in Herman *et al.* 2013.[417] This region was annotated by using it as a TBLASTN query to search the non-redundant database. Gene boundaries were identified using the alignments of the top hits, and genes were annotated based on top BLAST hit identities. For the three *N. fowleri* genomes, the parameter --alternatives-from-evidence was set to true, as this reports alternative gene transcripts if there is evidence for them (i.e. from transcriptome dataset). The parameter --alternatives-from-sampling was also set to true, as this outputs additional suboptimal transcripts. Parameters for determining the importance of different hint data were kept as default.

The *N. fowleri* V212 mitochondrial genome and extrachromosomal plasmid were manually annotated. Boundaries of protein coding genes were determined by searching the NCBI

non-redundant database and the *N. gruberi* mitochondrial genome using these sequences as BLASTX queries. tRNAs and rRNAs were predicted using the online programs tRNAscan-SE[418] Infernal,[419] and RNAmmer v.1.2.[420]

2.4.2.1 Core Eukaryotic Gene analysis

The completeness of a genome can be measured by a Core Eukaryotic Gene (CEG) analysis. CEGMA is a pipeline that identifies and annotates core genes in genomic DNA. Parra and colleagues (2007)[421] identified 458 genes that are thought to be CEGs; conserved in nearly all eukaryotic genomes. When each group of CEGs is aligned, they meet the following criteria: protein coverage is at least 75% of the alignment, there are no more than five internal gaps longer than 10 amino acids, and the sequences share at least 10% identity. A more curated dataset of 248 CEGs was later defined to include only those CEGs that meet the previous criteria and have a low number of in-paralogues in each species. Not surprisingly, most CEGs are housekeeping genes, although they vary in function. The typical CEGMA pipeline first performs its own gene prediction using geneid and GeneWise software, but since accurate gene predictions were already generated for the *N. fowleri* genomes, CEGMA was run using the Augustus-predicted proteins.

2.4.3 Genomic comparison of three *Naegleria* strains

2.4.3.1 Synteny analysis

Synteny is the conservation of blocks of sequence between two genomes. Preliminary analysis of a 60 kilobase region of the *N. fowleri* V212 genome suggested that there was little

conservation of sequence organization between *N. fowleri* and *N. gruberi*. To visualize synteny across all four genomes, the program Mauve (build 2015-02-25) was used, with the progressiveMauve alignment method.[422,423] Mauve generates progressive genomic alignments and reports blocks of similarity between them (with similarity profiles), showing rearrangements and transversions. Because the genomes being aligned are from relatively closely related taxa (within the same genus), the match seed weight of 15 was used. Both 'Full Alignment' and 'Iterative Refinement' parameters were selected. These parameters direct Mauve to first perform a recursive anchor search and full genome alignment using MUSCLE, followed by guide tree-independent refinement of the alignment. Default gap open and gap extend scores of -400 and -30 were used, respectively, as was the suggested HOXD scoring matrix.

2.4.3.2 Orthologous group analysis

To identify orthologous groups of sequences between the four *Naegleria* protein datasets, the program OrthoMCL v.2.0.9[424] was used. OrthoMCL starts with an all-versus-all BLAST, which was done locally, followed by filtering by percent match length. The OrthoMCL algorithm then finds protein pairs, making use of an SQL relational database.

To explain the orthogroup clustering methodology, several terms describing different homology relationship must first be defined. Homologues are evolutionarily related sequences, while orthologues are genes that are found in different species and evolved through speciation, and paralogues are genes related by duplication within a genome. The term in-paralogue effectively means 'paralogue', but this is distinct from out-paralogues, which describes genes related by both speciation and gene duplication. For example, the medium subunit of the AP1

complex is an out-paralogue to the small subunit of the AP2 complex, while the small subunits of both AP1 and AP2 are in-paralogues.

As mentioned above protein pairs are identified based on BLAST results. Protein pairs are co-orthologues, connected either by virtue of being orthologues or in-paralogues. To distinguish between orthologues and in-paralogues, OrthoMCL takes into account the reciprocal best BLAST hit pairs and normalized E-values of all other hits above a 1E-5 threshold and an HSP percent length > 50% of the shortest full protein. Normalization is done by averaging the E-values of all in-paralogues (this includes all in-paralogues in in-paralogue pairs, plus in-paralogues that have orthologues in any genome), and then dividing all E-values by this average. Finally, the orthologue, in-paralogue, and co-orthologue pairs and their normalized E-values are used as inputs for the MCL program, which uses the Markov Clustering algorithm to determine groups of orthologues, or orthogroups.[425] The Markov Clustering algorithm is an unsupervised clustering algorithm that clusters graphs (networks) based on pairwise scores (in this case, normalized E-values) and an inflation value. This latter value controls the clustering tightness, and for these analyses, was kept at the suggested 1.5.

2.4.4 Motif prediction

2.4.4.1 Trans-membrane domain prediction

The program TMHMM v2.0[426,427] was used to detect trans-membrane helices in all *N. fowleri* V212 proteins. This software uses an HMM generated from 160 cross-validated membrane proteins, and outputs all predicted helices and protein orientation in the membrane. Because trans-membrane helix prediction was done to generate a list of potential G-protein coupled receptors investigated further by domain prediction, scoring cutoffs were not used.

2.4.4.2 Transcriptional regulatory motif prediction

The programs RegRNA 2.0 and MEME were used to identify potential regulatory motifs upstream of up-regulated genes. For RegRNA 2.0, TRANSFAC TFBS was set to Human. MEME was set to 'normal' mode. Otherwise, for both programs, default settings were used.

2.4.5 Generating organellar markers in *Naegleria gruberi*

E. Herman performed the following functional work in N. gruberi in the lab of Dr. Anastasios Tsaousis at the University of Kent, UK, in November-December 2016.

2.4.5.1 N. gruberi culturing

N. gruberi strain NEG-M (provided by L. Fritz-Laylin) was grown axenically in M7 medium at 27°C.[428] Cells were subcultured every 3-4 days. M7 medium contains L-methionine, glucose, yeast extract, and fetal bovine serum. Some growth can occur in the absence of glucose, but not the other components.

2.4.5.2 Subcellular fractionation

In order to obtain fractions of cytosol, mitochondria, and membrane-bound organelles, *N. gruberi* cells were subjected to differential centrifugation. All centrifugation steps were performed at 4°C. First, *N. gruberi* cell cultures were spun at 1,000 *g* for 10 minutes. Cells were resuspended in SM buffer (250mM sucrose and 10mM MOPS-KOH, pH 7.4), and then spun

again under the same conditions, in order to remove debris. Cells were then resuspended in SM buffer containing protease inhibitor (Complete Mini EDTA-free cocktail tablets, Roche) and DNase (TURBO DNase, Ambion). Cells were lysed by passage through a 33G hypodermic needle five times. Lysed cells were resuspended in the SM buffer with inhibitors and spun at 1,000 *g* for 10 minutes to remove unbroken cells, membrane fragments, and nuclei, and the supernatant was collected. The supernatant was spun twice at 2,000 *g* for 10 minutes to obtain a fraction containing membrane-bound organelles, including the mitochondria. From this, the mitochondrial fraction was obtained by spinning twice at 10,000 rpm for 30 minutes (pellet), while the cytosol is found in the supernatant from this spin. These four fractions – whole cell lysate, membrane-bound organelles, mitochondria, and cytosol – were used for Western blotting.

2.4.5.3 SDS-PAGE

Sodium dodecyl sulfate polyacrylamide gel electrophoresis was performed to separate *N. gruberi* whole cell extracts and protein fractions prior to Western blotting. 12% gels were made for NgCOPB and NgSec31, while and 8% gel was made for NgSynPM. Visualization dye was added to protein samples, which were boiled for 10 minutes, and spun in a microcentrifuge at max speed for 10 minutes.

2.4.5.4 Western Blotting

Protein on the gels was transferred to a PVDC membrane. After transfer, Ponceau S stain was used to check for the presence of protein on the blots. Blots were blocked with a 5% milk PBS-Tween solution for 1 hour at room temperature, or overnight at 4°C. After blocking, blots

were washed in 0.5% milk PBS-Tween three times for 10 minutes each. Following the washes, the primary antibody was added to 1% milk PBS-Tween: α-NgSec31 polyclonal antisera from rat, α-NgCOPB polyclonal antisera from chicken, and α-NgSynPM polyclonal antisera from rabbit (two animals per protein; Davids Biotechnologie GmbH; Germany). Antibody concentrations are given for each blot. The blots were incubated with the primary antisera for one hour, followed by three 10 minute washes with 1% milk. Then, the appropriate secondary antibodies conjugated to peroxidase (Sigma) were added at a concentration of 1:2500 in a 1% milk PBS-Tween, and incubated with the blots for one hour. Finally, the blots were washed with PBS three times for 10 minutes each and visualized using the Syngene G:BOX XT4 machine and the GeneSys software.

2.4.5.5 Immunofluorescence microscopy

N. gruberi cells were grown on LabTek Chamber slides for 24-48 hours prior to the experiment. Cells were incubated with ER-tracker DPX marker (Molecular Probes) for 20 minutes and fixed with 1% formaldehyde. They were then permeabilized with 0.5% Triton-X in 1X PBS. Cells were blocked for one hour in 3% BSA-1XPBS and probed with chicken α-NgCOPB (1:26), rat α-NgSec31 (1:200), and rabbit α-NgSynPM (1:100) antisera. 1:1000 dilutions of the following secondary antibodies were incubated with the slides: Alexa Fluor 488 goat α-chicken IgG (H-L), Alexa Fluor 488 chicken α-rat IgG, and Alexa Fluor 594 donkey α-rat IgG (Molecular Probes). A DAPI-containing anti-fade mounting reagent (Vectashield) was used to mount the cells, and cells were then observed by fluorescence microscopy (observed under an Olympus IX81 fluorescence microscope and a laser scanning Zeiss LSM 880 confocal

microscope). Images were taken with Micromanager 1.4 software fluorescence and Zeiss Zen software for confocal microscope and processed with ImageJ.

2.4.5.6 Immunoelectron microscopy

Aspirated cultures of *N. gruberi* were fixed in a PBS-4% formaldehyde solution for 1 hour, and then were washed with PBS. The samples were dehydrated through an ethanol series, 30%, 50%, 70%, 90%, and 3x100%. The ethanol was aspirated and the samples were suspended in LR white Resin (Agar Scientific). Samples were placed in a vacuum for 2 minutes to improve resin permeation. Fresh resin was added and the samples were transferred into gelatin capsules (Agar Scientific) and hardened in a 60°C oven for 15 hours. The blocks were polished and sectioned by ultra-microtome at a thickness of 70nm. They were then placed 3mm gold grids. Immuno-staining was performed in humidifying chambers. The grids were blocked by incubating with a 2% BSA PBS-Tween solution for 1 hour. The primary antibodies were incubated with the samples for 15 hours in 1:10, 1:50, and 1:100 dilutions at 8°C. Gold-conjugated secondary antibodies were then incubated for 30 minutes at room temperature. Counterstaining was done using 4.5% uranyl acetate in PBS (incubated for 15 minutes) and Reynold's lead citrate (2 minutes). IEM images presented here are the work of *A. Tsaousis*.

3. Membrane trafficking evolution in symbiosis and parasitism in *Blastocystis* sp., *Proteromonas lacertae*, and *Cafeteria roenbergensis*

3.1 Introduction

This chapter focuses on how the membrane trafficking system has evolved in three closely related organisms with distinct lifestyles, ranging from free-living to parasitic. Often, parasite genomes are streamlined, and it is thought that this may be to conserve energy during replication; however, there are also instances of gene family expansion to support specialized interaction with the host. Parasite genomes must be compared with those of closely related free-living species, in order to rule out lineage-specific characteristics that are not related to a parasitic lifestyle.

Blastocystis sp., *Proteromonas lacertae*, and *Cafeteria roenbergensis*, are deep-branching heterokonts (Stramenopiles) (Figure 1.1).[429] *Blastocystis* has long been an enigmatic obligatory endobiont of the guts of human, mammals, reptiles, birds, and insects. It is a commonly found protist in the gut of humans, reaching 100% in some populations (Senegal River Basin), and 20% in industrialized countries.[60] It is estimated that between one and two billion people may be colonized by *Blastocystis* worldwide.[430] It is spread by contaminated food and water, as well as human-to-human contact. Infection with *Blastocystis* sp. has been associated with non-specific gastrointestinal disorders such as nausea, vomiting, and diarrhea, and has also been linked to Irritable Bowel Syndrome and urticarial lesions (case reports reviewed in Roberts *et al.* 2014).[431] However, many people that carry *Blastocystis* sp. are asymptomatic. Eukaryotic gut microbiome studies are only now beginning to unravel the extent of infection versus symptom presentation. *Blastocystis* sp. can be classified into different subtypes, of which there are 17 in total, and nine that have been detected in humans.[432] ST1 through ST4 are most frequently isolated from humans, and have distinct geographic distributions.[433] Human infection with ST5 through ST9 is likely the result of zoonotic transmission, as they are less prevalent in humans but have known

animal hosts. Despite ST3 being a relatively common human infection, those who are infected tend to be asymptomatic.[434] Originally, *Blastocystis* sp. pathogenesis was thought to be dependent on subtype, but epidemiological data suggest that the relationship is more complex; subtype pathogenicity and symptom severity varies by geographical region.[434–437] For example, the ST2 strain has been shown to have high symptom-infection rates in some populations,[436,438] but no link to pathogenicity in others.[439,440] Relevant to this work, infection data have shown that ST1 is statistically related to pathogenicity,[441] and both ST4[442,443] and ST7[444] are considered to be pathogenic strains.

P. lacertae is the closest relative to *Blastocystis* species that has yet been identified.[445] Originally isolated from the cloaca of the sand lizard *Lacerta agilis*, *P. lacertae* has been described as a either a parasite or commensal organism associated with lizards and amphibians[446–448] (and rarely mammals), but it is unclear whether its presence actually causes illness in the host. While both *Blastocystis* sp. and *P. lacertae* are endobionts of animals, the sister taxon this clade is *Cafeteria*, a free-living bacterivorous marine flagellate. It is a suspension feeder found in oceans around the world, typically in coastal waters. It uses its flagella for both anchoring and to aid in feeding,[449] and is a phagotroph, in contrast with *P. lacertae*, which is capable only of pinocytosis.[450,451] While phagocytosis has been observed in a single stock of *Blastocystis* sp.,[452] its feeding strategy is somewhat unclear.

In addition to ranging from free-living to parasitic, these three organisms are also morphologically diverse. *Cafeteria* and *Proteromonas* are heterokont-like, recognizable as such since they contain two flagella of different lengths. *Blastocystis* sp., however, has a highly derived morphology. It is a round, non-motile cell with a large internal vacuole that makes up nearly the entire cell volume (multiple large vacuoles are also observed).[338] Under light

microscopy, it more closely resembles an air bubble than it does a eukaryote, let alone a heterokont.[453] Additionally, avacuolar, amoeboid and cyst forms have also been observed, but the vacuolar form is considered to be the typical *Blastocystis* cell form.[338] A study in rats suggests that, like many parasites transmitted fecal-orally, only the cyst is transmissible.[454] As mentioned above, the morphology of *Proteromonas* is more typical of Stramenopiles, but like *Blastocystis* sp., it is also capable of encystation. A cyst form has not been described for *Cafeteria*.

Another difference between the gut-resident taxa and *Cafeteria* is their metabolism; *Blastocystis* and *Proteromonas* are obligate anaerobes with modified mitochondria, while *Cafeteria* has a fully functional aerobic mitochondrion. Derived mitochondria, or mitochondria-related organelles (MROs), are found across the tree of eukaryotes and are typically – but not always – associated with parasitism and anaerobiasis (reviewed by Makiuchi and Nozaki 2014).[455] Two types of MROs are hydrogenosomes and mitosomes, the latter being the more functionally and morphologically reduced species. The mitochondria of *Blastocystis* and *Proteromonas* have been classified as an intermediate form; not a typical mitochondrion, but less reduced than a hydrogenosome.[456,445,457] Despite living under anaerobic conditions, this modified mitochondria is reportedly functional in *Blastocystis*, although no activity of mitochondrial enzymes has been detected.[458] As *Cafeteria* is free-living, it likely has a fully functional mitochondrion.

A frequently observed feature of parasites is membrane trafficking system specialization, including both gene losses as part of genomic streamlining (e.g. in the Apicomplexa, *Giardia*, and to some extent the kinetoplastids)[50,68,459] as well as novel or specialized trafficking biology stemming from gene family expansions associated with host association (e.g. in *Entamoeba* and

Trichomonas).[70,460] Examining the membrane trafficking system in the endobiotic *Blastocystis* sp. and *P. lacertae* in comparison with that of the free-living *C. roenbergensis* will give insight into how living in a host gut has impacted the basic intracellular biology of these organisms. Both *Giardia* and *Entamoeba* are human gut parasites with MROs, and yet have adopted different genomic strategies in the face of similar environmental pressures. Understanding the membrane trafficking system of this lineage will add another sampling point. Furthermore, like *Entamoeba* and *Giardia*, *Blastocystis* secretes cysteine proteases as a virulence mechanism to degrade immune molecules such as IgA.[461] There may be gene family expansions within the secretory system that support this, and a lineage-specific analysis will elucidate whether this is a feature of the parasitic *Blastocystis*, or whether it is pre-adaptive in *Proteromonas* and/or *Cafeteria*.

There are other reasons to study the membrane trafficking system in these organisms beyond parasitism; these organisms employ somewhat different trophic strategies – phagocytosis versus pinocytosis – and the effect of this may be seen in the complement of endocytic machinery. In regards to the morphological distinctiveness of *Blastocystis,* the central vacuole of *Blastocystis* sp. may impact trafficking in different ways. It is thought to be a lipid and carbohydrate storage organelle, the contents of which may be consumed during growth,[462] so there may be central vacuole-specific trafficking machinery. If this were the case, this machinery would likely be related to endolysosomal trafficking factors. Additionally, this lineage is the most basal branch of the Stramenopile clade,[463] so analysis of the membrane trafficking system will give a more complete picture of the origin and evolution of Stramenopile-specific aspects of membrane trafficking.

In order to answer these questions, the genome of *Blastocystis* ST1 was sequenced, and compared with the publicly available genomes of the ST4[464] and ST7[465] subtypes. Following an analysis of membrane trafficking machinery in these organisms (published in Gentekaki *et al.* 2017),[466] the genome of *Proteromonas lacertae* and transcriptome of *Cafeteria roenbergensis* were sequenced, and the *Blastocystis* sp. comparison was updated to include the membrane trafficking complements derived from these additional datasets (the *P. lacertae* and *C. roenbergensis* data are currently unpublished). Comparative genomic analyses of the membrane trafficking systems of these organisms were performed to gain insight into (i) the evolution of symbiosis and parasitism, and the related metabolic shift to obligate anaerobism, (ii) changes to the secretory system that underlie secretion of pathogenicity factors (i.e. cysteine proteases), (iii) general membrane trafficking changes in relation to changes feeding strategy and vacuolar storage, and (iv) the evolution of membrane trafficking at the base of the Stramenopile clade.

3.2 Specific methods

The genomes and predicted proteomes of three subtypes of *Blastocystis* sp., ST1, ST4, and ST7, were searched for vesicle formation machinery, vesicle fusion machinery, and autophagy machinery using BLAST, and in the case of predicted proteins, HMMer. The predicted proteome of *P. lacertae* and transcriptome of *C. roenbergensis* were searched using BLASTP or TBLASTN, respectively, as well as the genome of *P. lacertae* in cases where homologues could not be identified in the predicted proteome. If a homologue could not be identified in one or two of the taxa but was found in the third, this sequence was used to search the databases of the first two taxa, in order to reduce the chance of a false negative due to sequence divergence. Because the *C. roenbergensis* dataset is transcriptomic, it is not considered

to be complete, and therefore no inference of gene loss can be made based on absence in this organism. Phylogenetics was used to classify highly paralogous protein families (the Snf7 subfamily of ESCRT proteins, TBCs, SNAREs, ArfGAPs, adaptins, and TSET), as well as to determine whether *Blastocystis* and *Proteromonas* encode all three ancient paralogues of Sec24, a subunit of the COPII coat complex. BLAST and HMMer methodology and scoring cutoffs are defined in the Methods. Bayesian and Maximum-likelihood phylogenies were generated using the methodology outlined in the Methods.

The genome and predicted proteome of *Blastocystis* ST1 was produced by a genome project led by Dr. E. Gentekaki, while the *Proteromonas* and *Cafeteria* datasets were produced by R. Low. A. Schlacht identified the Blastocystis ST1 ArfGAPs, ArfGEFs, SNAREs, multi-subunit tethering complexes, and SM proteins, while E. Herman performed these analyses in *P. lacertae* and *C. roenbergensis*. E. Herman performed all comparative genomics analyses of the adaptins, TSET, the coat complexes, ESCRTs, endocytic proteins and autophagy machinery in the five datasets. E. Herman performed all phylogenetic analyses.

3.3 Membrane trafficking system evolution in three parasitic strains of *Blastocystis* sp., the related endobiont *Proteromonas lacerate* and free-living *Cafeteria roenbergensis* gives insight into the evolution of parasitism and endobiosis

3.3.1 Vesicle formation machinery

A comparative genomic analysis of membrane trafficking machinery was undertaken in *Cafeteria, Proteromonas*, and three subtypes of *Blastocystis* in order to understand how this system has evolved during transition to parasitism in this lineage.

Vesicle formation machinery is generally conserved in the three Stramenopiles, albeit with some losses and expansions. Gene presence and absence is shown in the Coulson plot in Figure 3.1, which is based on the data contained in Online Appendix Tables 3.1 and 3.2. There are duplications in several subunits of COPI, which functions in Golgi-ER and intra-Golgi retrograde transport. In addition to a single duplicated COPD subunit in *Blastocystis* sp. and *P. lacertae*, *Blastocystis*, sp. have also duplicated COPA and COPB' subunits. This result suggests the possibility of multiple species of COPI complexes in *Blastocystis* sp., and potentially subfunctionalization. There are several duplications of COPII subunits (ER-Golgi anterograde trafficking) across the taxa, including multiple paralogues of Sec24 in the three *Blastocystis* subtypes as well as *P. lacertae*. This finding suggests two possibilities, which are not mutually exclusive. First, Sec24 is the primary cargo adaptor protein in the complex,[467] and therefore some of the additional subunits may recognize different cargo. Second, that *Blastocystis* sp. encodes the ancient, third paralogue of Sec24.

As shown by Schlacht et al. 2015,[468] there were three Sec24 paralogues in the LECA. Most eukaryotes have retained Sec24I and Sec24II, but Sec24III has been patchily lost. These include several Amoebozoans, most plants, and very few SAR taxa. A phylogenetic analysis of the Sec24 proteins in the Stramenopiles analyzed here show that *Blastocystis* sp. encodes a Sec24III protein, adding another Stramenopile lineage to the few organisms that have retained it (Figure 3.2).

Figure 3.1. Comparative genomic survey of vesicle formation machinery in *Blastocystis* sp., *Proteromonas lacertae*, and *Cafeteria roenbergensis*.

The vesicle formation machinery includes the ESCRTs, adaptor proteins, COPI, COPII, clathrin and endocytic machinery, retromer, and TSET. In this and subsequent Coulson plots, grey circles indicate components present in the LECA. For each organism below, filled pie sectors indicate that an orthologue was identified, while unfilled sectors indicate that a homologue could not be found. Numbers in the sectors indicate the number of paralogues identified of that component. Asterisks (*) indicate putative Vps20 or Vps32 protein, as orthology could not be determined by phylogenetics. A hash (#) indicates that sortilin (Vps10)-like sequence was identified. A question mark (?) indicates that one of the two TTRAY sequences identified was excluded from the TTRAY1 clade (and COPI subunit clades), but did not group with a TTRAY2 clade with significant node support (posterior probability >0.80, bootstrap >50). Colour-coding is for visualization purposes only.

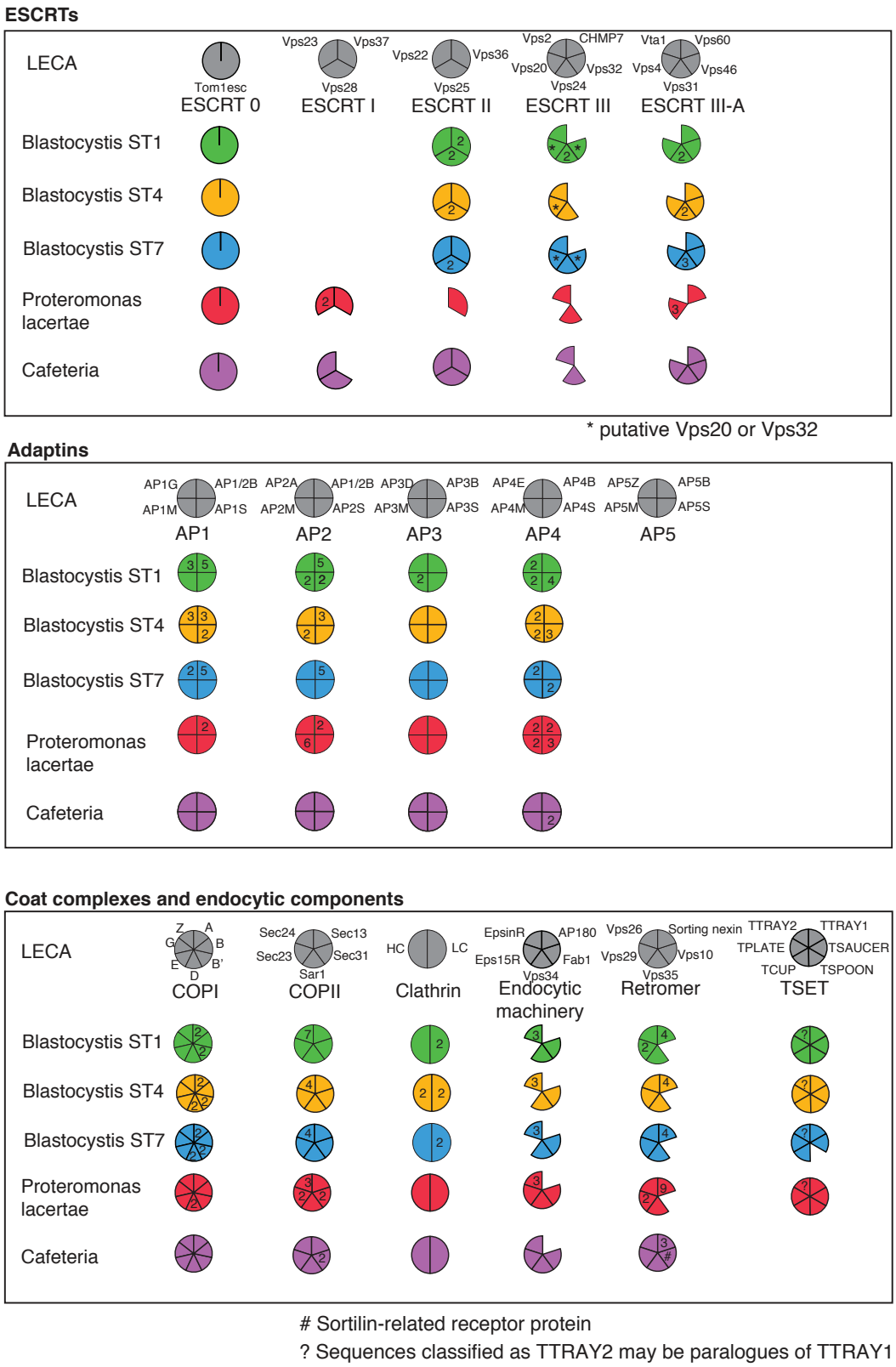

ESCRTs
LECA
Tom1esc
ESCRT 0
Vps23
Vps37
Vps28
ESCRT I
Vps22
Vps36
Vps25
ESCRT II
Vps2
CHMP7
Vps20
Vps32
Vps24
ESCRT III
Vta1
Vps60
Vps4
Vps46
Vps31
ESCRT III-A
Blastocystis ST1
Blastocystis ST4
Blastocystis ST7
Proteromonas lacertae
Cafeteria
* putative Vps20 or Vps32
Adaptins
LECA
AP1G
AP1/2B
AP1M
AP1S
AP1
AP2A
AP2M
AP2S
AP2
AP3D
AP3B
AP3M
AP3S
AP3
AP4E
AP4B
AP4M
AP4S
AP4
AP5Z
AP5B
AP5M
AP5S
AP5
Blastocystis ST1
Blastocystis ST4
Blastocystis ST7
Proteromonas lacertae
Cafeteria
Coat complexes and endocytic components
LECA
Z
A
G
B
E
B'
D
COPI
Sec24
Sec13
Sec23
Sec31
Sar1
COPII
HC
LC
Clathrin
EpsinR
AP180
Eps15R
Fab1
Vps34
Endocytic machinery
Vps26
Sorting nexin
Vps29
Vps10
Vps35
Retromer
TTRAY2
TTRAY1
TPLATE
TSAUCER
TCUP
TSPOON
TSET
Blastocystis ST1
Blastocystis ST4
Blastocystis ST7
Proteromonas lacertae
Cafeteria
Sortilin-related receptor protein
? Sequences classified as TTRAY2 may be paralogues of TTRAY1

Figure 3.1

Figure 3.2. Phylogenetic classification of Sec24 paralogues in *Blastocystis* sp. *P. lacertae*, and *C. roenbergensis*.

Node values indicating statistical support are listed as MrBAYES/Phylobayes/RAxML (posterior probability/posterior probability/bootstrap), or as circles indicating the minimum level of support for that node. Nodes without values do not have significant support in one or more phylogeny. The best Bayesian topology (MrBAYES) is shown with node values from all three methods. Numbers following the species name indicate the Sec24 paralogue classification from Schlacht *et al.* 2015.[468] *Blastocystis* sp. has retained the ancient but patchily lost Sec24-III paralogue.

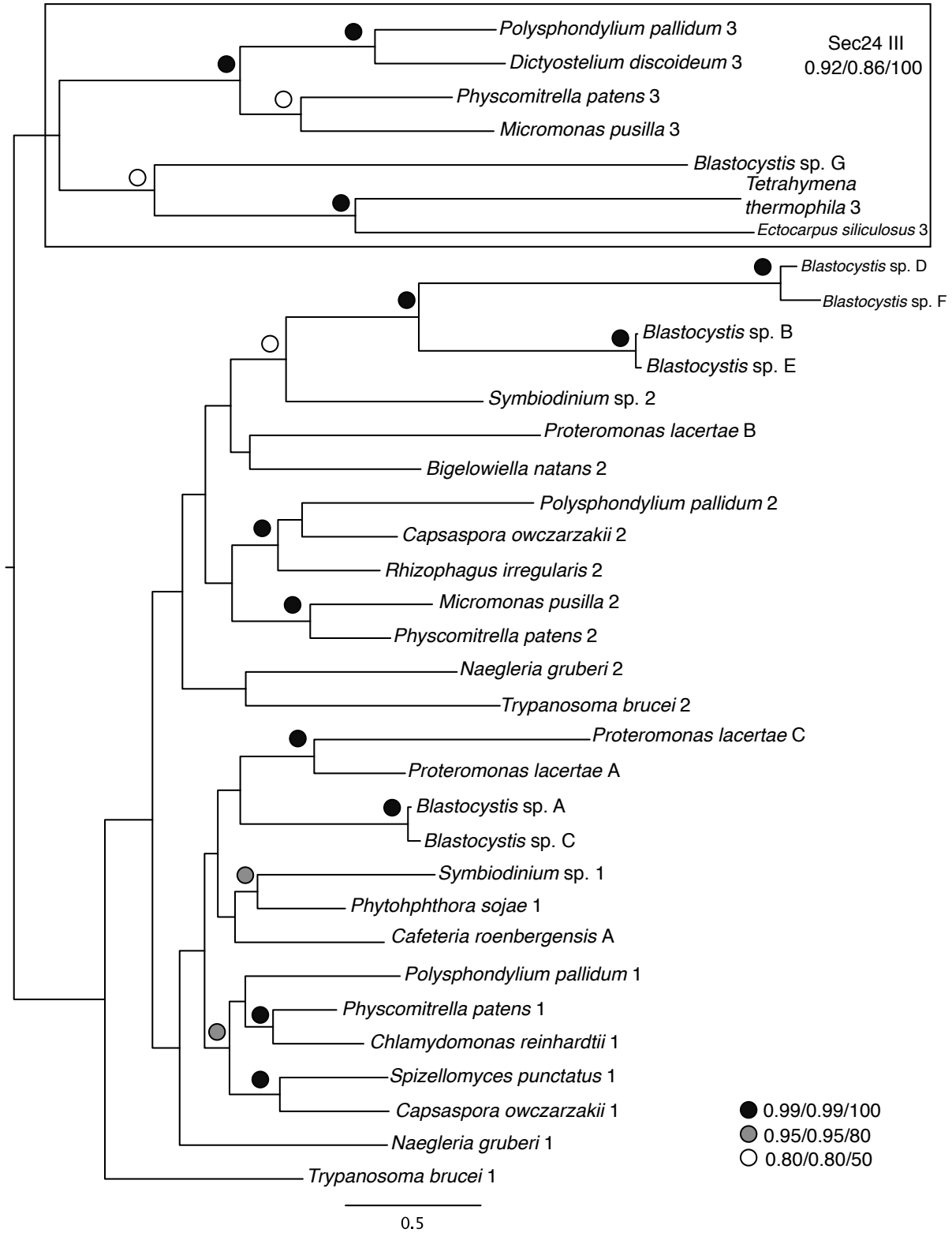

Sec24 III
0.92/0.86/100
Polysphondylium pallidum 3
Dictyostelium discoideum 3
Physcomitrella patens 3
Micromonas pusilla 3
Blastocystis sp. G
Tetrahymena thermophila 3
Ectocarpus siliculosus 3
Blastocystis sp. D
Blastocystis sp. F
Blastocystis sp. B
Blastocystis sp. E
Symbiodinium sp. 2
Proteromonas lacertae B
Bigelowiella natans 2
Polysphondylium pallidum 2
Capsaspora owczarzakii 2
Rhizophagus irregularis 2
Micromonas pusilla 2
Physcomitrella patens 2
Naegleria gruberi 2
Trypanosoma brucei 2
Proteromonas lacertae C
Proteromonas lacertae A
Blastocystis sp. A
Blastocystis sp. C
Symbiodinium sp. 1
Phytohphthora sojae 1
Cafeteria roenbergensis A
Polysphondylium pallidum 1
Physcomitrella patens 1
Chlamydomonas reinhardtii 1
Spizellomyces punctatus 1
Capsaspora owczarzakii 1
Naegleria gruberi 1
Trypanosoma brucei 1
0.99/0.99/100
0.95/0.95/80
0.80/0.80/50
0.5

Figure 3.2

Surprisingly, a complete TSET complex was identified in *Blastocystis* sp. and *P. lacertae*. Subunits were classified by phylogenetics, as shown in Supplementary Figures S3.1-S3.4 (Online Appendix Table 3.3). The TPLATE tree was unresolved, and therefore not included here. In several trees, the TSET sequences do not form a single clade, however, the new Stramenopile TSET subunits always group with *bona fide* TSET sequences with high node support values.

In other stramenopiles such as the giant brown kelp *Ectocarpus siliculosus*, TSET is incomplete, with subunits being patchily lost or otherwise unidentifiable due to low sequence similarity, and it is even less well conserved in Alveolates and Rhizaria. This is the first identification of organisms within the SAR clade that encode complete TSET complexes. Identification of a whole TSET complex in these taxa raises the possibility that it is functional, and potentially also trafficking material to the cell surface. The TSET complex was only identified in *P. lacertae* using HMMer, and hidden Markov models including the *P. lacertae* sequences were able to identify orthologues in *Blastocystis* sp. Therefore, the presence of the TSET complex was not published in the *Blastocystis* sp. genome paper; however, it will be published in the genome paper on *P. lacertae* and *C. roenbergensis*.

Phylogenetic trees classifying the adaptin subunits are shown in Supplementary Figures S3.7-S3.10 (Online Appendix Table 3.3). These trees were generated separately from the ones used to classify TSET sequences, as additional adaptins were identified following the TSET analysis, and were therefore not included in those trees.

Clathrin and AP2, involved in clathrin-mediated endocytosis, are present in all taxa, while AP180 was not identified, although previous work suggested it was lost at the base of the SAR clade, prior to the divergence of Stramenopiles.[469] In *Blastocystis* sp., the structure-

regulating light chain of clathrin has been duplicated. In vertebrates, the light chain influences cargo selection by influencing triskelion structure,[470] and may play a similar role in *Blastocystis* sp. There are also duplications of AP2 subunits. Six highly similar AP2M paralogues were identified in *P. lacertae*, differing by no more than 12 amino acids between them, suggesting that these duplications are recent. The shared beta subunit of AP1 and AP2 is also duplicated in *P. lacertae*, and to a greater extent in *Blastocystis* sp., particularly ST1 and ST7.

In addition to the AP1 and AP2 beta subunit mentioned above, the AP1G subunit is duplicated in all three *Blastocystis* subtypes. In human cells, the it is AP1G that interacts with EpsinR, and multiple copies of EpsinR are found in both *Blastocystis* sp. and *P. lacertae.*[180] That these interacting proteins have both been expanded in *Blastocystis* sp. could suggest that they have preferential interacting partners. Eps15R could not be identified in *Blastocystis* sp., but it is also lost in several eukaryotic lineages.[471] The AP4 complex also functions in TGN-endosome trafficking, and nearly all of its subunits have been duplicated in both *Blastocystis* sp. and *P. lacertae*. The retromer coat complex mediates receptor recycling from endosomes and is present in all taxa. *C. roenbergensis* encodes a sortilin-related receptor (Vps10-like), which is thought to be the universal cargo for retromer.[196] It has a patchy distribution in eukaryotes, and could not be identified in *P. lacertae* or *Blastocystis* sp.

Overall, *C. roenbergensis* has the most complete ESCRT complement, however, any losses is this organism cannot be reliably determined since the only available sequence data are transcriptomic. In *Blastocystis* sp. and *P. lacertae*, there is evidence for both shared and independent losses. Phylogenetics was used to classify the Snf7 and Vps24 families, shown in Supplementary Figures S3.5 and S3.6 (Online Appendix Table 3.3). While *P. lacertae* maintains two of three ESCRT I subunits (Vps23 and Vps37), ESCRT I appears to be completely lost in all

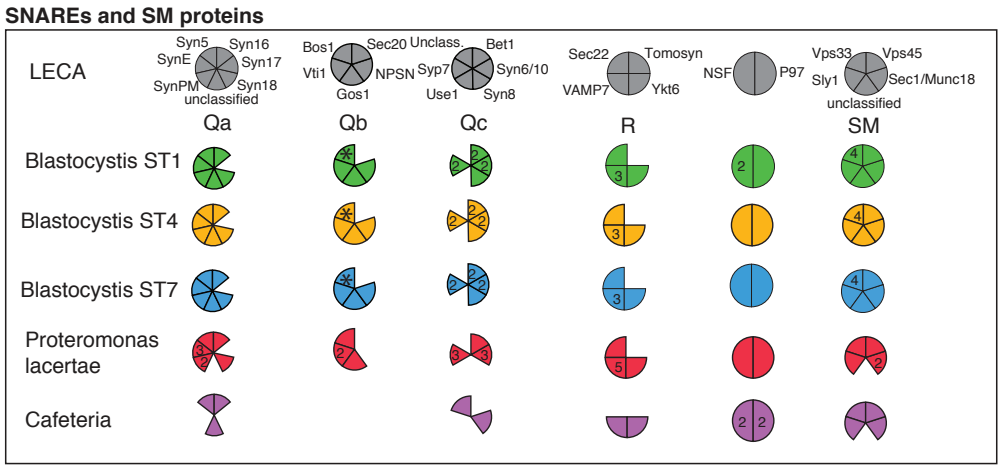
three *Blastocystis* subtypes. It is therefore surprising that the *Blastocystis* subtypes have retained more components of ESCRTs II, III, and III-associated than *P. lacertae*, and in some cases, have duplicated subunits, such as Vps25 and Vps31. Conversely *P. lacertae* appears to have independently lost at least five ESCRT subunits. This suggests that there is a pattern of loss or degradation of ESCRT components in both endobiotic stramenopiles, which is not seen to the same extent in the free-living *C. roenbergensis*. If the ESCRT subunits that could not be identified in *P. lacertae* are truly lost, it is difficult to envision MVB biogenesis, as these subunits are structural and scission proteins of ESCRTs II and III. The only AP complex with minimal subunit duplications is AP3, with the exception of a single duplication of the medium subunit in *Blastocystis* ST1. AP3 functions in trafficking material between the TGN and MVB/late endosome, so the relative paucity of AP3 paralogues is in line with the partial loss of the ESCRT complex.

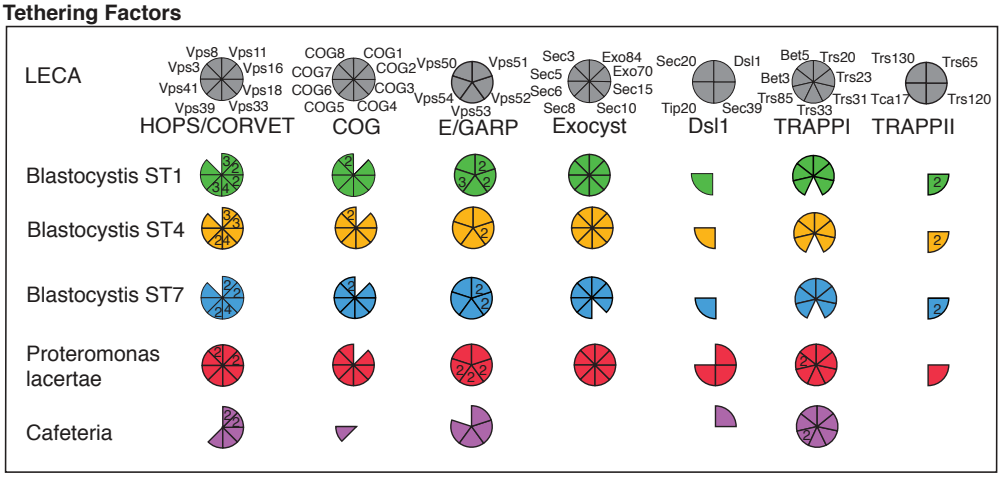
3.3.2 Vesicle fusion machinery

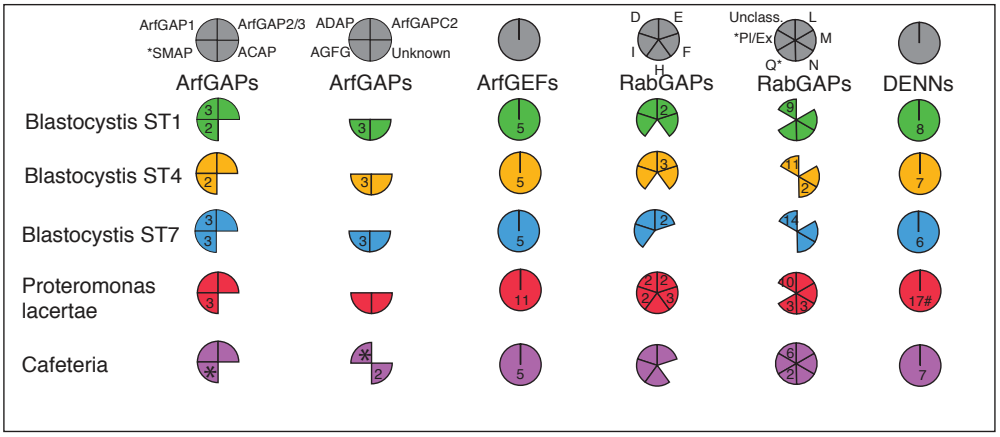
In general, vesicle fusion machinery is less conserved in the three taxa than vesicle formation machinery (Figure 3.3, Online Appendix Tables 3.1 and 3.2). SNAREs were classified using phylogenetics, shown in Supplementary Figures S3.11-S3.13 (Online Appendix Table 3.3). While ER-Golgi anterograde vesicle fusion occurs via the multisubunit tethering complex (MTC) TRAPPI, which is present, the retrograde trafficking step occurs via Dsl1, which is lost almost completely in *Blastocystis* sp.. This is not the case in *P. lacertae*, which encodes all subunits except Tip20. In both *Blastocystis* sp. and *P. lacertae*, the SNARE proteins Use1 and Sec20, which function in Golgi-ER trafficking, also could not be identified. Additionally, only one of the four TRAPPII MTC subunits, a tether for intra-Golgi retrograde transport, could be

Figure 3.3. Comparative genomic survey of vesicle fusion machinery and GTPase regulators in *Blastocystis* sp., *Proteromonas lacertae*, and *Cafeteria roenbergensis*.

The vesicle fusion machinery includes SNARE complexes and SM proteins, multisubunit tethering complexes (MTCs), while the GTPase regulators are the GTPase Activating Proteins (GAPs) and Guanine nucleotide Exchange Factors (GEFs) for Arf and Rab GTPases. An asterisk (*) indicates a putative homology assignment, which could not be reliably classified using phylogenetics. A hash (#) indicates that many of the identified sequences are fragments of DENN domains.







*putative classification

majority are DENN domain fragments

Figure 3.3

identified. Together, these findings suggest that while there is complexity in coat-based vesicle formation, tethering complexes are generally more dispensable in these taxa.

Plasma membrane vesicle fusion is mediated by the Exocyst MTC, plasma membrane SNAREs, subtypes of VAMP SNAREs, Qbc domain-containing SNAREs in Opisthokonts,[472] and NPSN and Syp7 SNAREs in plants.[91,293] *Blastocystis* encodes the Qb SNARE NPSN and Qc SNARE Syp7, while no Qbc proteins could be identified. However, NPSN and Syp7 orthologues could not be identified in *P. lacertae*, although *P. lacertae* encodes multiple SynPM and Vamp7-like paralogues, suggesting either that other proteins function in place of secretory Qb and Qc SNAREs in *P. lacertae*, or that SNARE complexes can be formed without them.

Endocytic machinery, in general, is present and often expanded in *Blastocystis* sp. and *P. lacertae*, with the exception of ESCRT subunit losses. As discussed above, there are duplications in AP1, AP2, and AP4 adaptor complexes. In terms of vesicle fusion machinery involved in TGN-endosome recycling, there are duplications in subunits of the GARP/EARP MTCs and the Qc SNARE Syntaxin 6 in *Blastocystis* sp. and *P. lacertae.* The endolysosomal HOPS and CORVET MTCs are complete in *P. lacertae* (with several duplicated subunits). In *Blastocystis* sp., the CORVET-specific Vps8 could not be identified; meanwhile, several VpsC core subunits have been duplicated, as has the HOPS-specific subunit Vps39. This expansion of endolysosomal tethering machinery contrasts with the reduction in the ESCRT complement, suggesting specialization of this arm of the membrane trafficking system. Furthermore, the apparent loss of Vps8 in *Blastocystis* sp. raises the possibility a lineage-specific factor can interact with VpsC core proteins.

Endo-lysosomal specialization via adaptor protein and MTC expansion is a curious observation against the backdrop of deteriorating MVB biogenesis machinery, a key step in degrading

transmembrane proteins. To understand more about other aspects of endo-lysosomal trafficking in this lineage, a comparative genomic analysis of the autophagy machinery was undertaken. Autophagosomes are dynamic endomembrane organelles that feed into the degradative pathway through HOPS-mediated fusion with lysosomes (Figure 3.4, Online Appendix Table 3.4). Based on comparison with *C. roenbergensis* and other stramenopiles,[473] both *Blastocystis* sp. and *P. lacertae* have both independent and lineage-specific losses. Both *P. lacertae* and *Blastocystis* sp. have lost ATG1 and ATG9, while *P. lacertae* has lost ATG20, ATG3, ATG7, and ATG5 independent of *Blastocystis* sp., and *Blastocystis* sp. has lost Vac8 independently of *P. lacertae*. Although the vesicle nucleation machinery appears to be present, many of these proteins have other cellular functions; these include TOR1 (intracellular signaling),[474] Vps15 and Vps34 (vacuolar protein sorting, among other functions),[475] Atg6/Beclin/Vps30 (endosomal recycling),[476] and PEP4 and PRB1 (proteases involved in acidification).[477] Surprisingly, *Blastocystis* sp. appears to be missing ATG3, ATG4, ATG7, and ATG8; parts of the vesicle expansion machinery that are conserved across eukaryotes.[473] Despite the transcriptomic nature of the *C. roenbergensis* dataset, these proteins, as well as other critical factors in this system, were identified, and therefore the losses are specific to the endobiotic *Blastocystis* sp. and *P. lacertae*.

The Arf and Rab small GTPase regulators (GAPs and GEFs) involved in regulating vesicle formation and vesicle fusion, respectively, were also queried (Figure 3.3). Arf and Rab GTPases themselves were not searched for, as these proteins were within the purview of another research group working on the *Blastocystis* sp. genome project.[466] Phylogenetic classification of TBC RabGAPs and ArfGAP proteins are shown in Supplementary Figures S3.14 and S3.15, respectively (Online Appendix Table 3.3).

Figure 3.4. Comparative genomic survey of autophagy machinery in *Blastocystis* sp., *Proteromonas lacertae*, and *Cafeteria roenbergensis*.

Autophagy machinery is classified by the following steps in autophagosome biogenesis: induction, cargo packaging, vesicle nucleation, vesicle expansion, retrieval, and vesicle breakdown, and factors associated with these processes. In general, machinery was only included if it is likely ancient and present across eukaryotes based on the analysis by Duszenko *et al.* (2011).[473] Dots indicate simple presence/absence, while numbers in dots indicate paralogues.

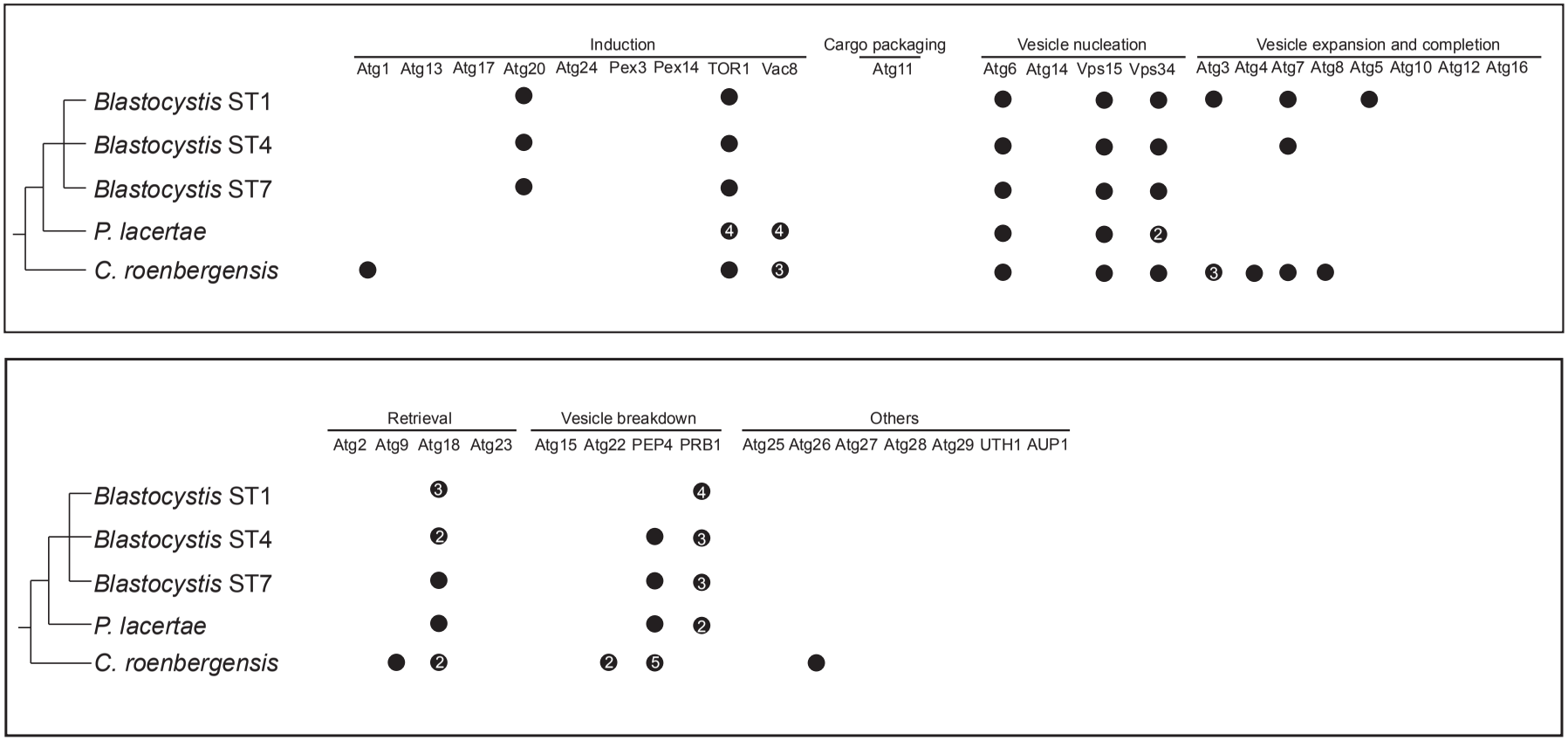

Induction
Cargo packaging
Vesicle nucleation
Vesicle expansion and completion
Atg1 Atg13 Atg17 Atg20 Atg24 Pex3 Pex14 TOR1 Vac8
Atg11
Atg6 Atg14 Vps15 Vps34
Atg3 Atg4 Atg7 Atg8 Atg5 Atg10 Atg12 Atg16
Blastocystis ST1
Blastocystis ST4
Blastocystis ST7
P. lacertae
C. roenbergensis
Retrieval
Vesicle breakdown
Others
Atg2 Atg9 Atg18 Atg23
Atg15 Atg22 PEP4 PRB1
Atg25 Atg26 Atg27 Atg28 Atg29 UTH1 AUP1

Figure 3.4

However, trees for ArfGEFs and DENN domain-containing RabGEFs were unresolved, and therefore not shown here. General features of these families can be uncovered by comparative genomics alone. Typically, *P. lacertae* has a more complete and expanded set of Arf and Rab regulators than *Blastocystis* sp., e.g. 11 versus 5 ArfGEFs, and 17 versus ~8 DENN domain RabGEFs. *P. lacertae* also has a more complete set of TBC domain RabGAPs than *Blastocystis* sp., including TBC-H and TBC-L (nomenclature from Gabernet-Castello *et al.* 2013).[266] However, the functions of these TBC proteins are unknown in both cases. TBC-E is one factor that has been duplicated in *Blastocystis* sp. and *P. lacertae*. It is specific to Rab35, which together function in a fast endosome-plasma membrane recycling pathway in mammalian cells.[260] Interestingly, a putative plant-specific TBC protein was identified in *C. roenbergensis*, but not in the other taxa. *Blastocystis* sp. has duplicated its ArfGAP1 and AGFG ArfGAPs, while *P. lacertae* has only one of each. ArfGAP1 functions in cargo sorting and biogenesis of COPI-coated vesicles at the Golgi,[154] and several subunits of the COPI complex have also been duplicated in *Blastocystis* sp.. AGFG, also known as Hrb, colocalizes with clathrin and AP2, and is involved in clathrin-mediated endocytosis.[205,478]

3.4 Discussion

Analysis of the membrane trafficking system complement in *Blastocystis* sp., *P. lacertae*, and *C. roenbergensis* gives insight into how this system has evolved over the transition from a free-living to an endobiotic/parasitic lifestyle. Unlike organisms such as *Giardia* and the Apicomplexa, there does not appear to be large-scale streamlining of the endomembrane system in the endobionts. However, there are a few key losses of trafficking machinery in both lineages.

ESCRT complex subunits appear to be differentially lost in these taxa. While *P. lacertae* has retained some ESCRT I components and *Blastocystis* sp. has not, it has lost other ESCRT II, III, and IIIA components that are necessary for vesicle budding. The most critical of these is Vps32, the ESCRT subunit that multimerises into spirals at the neck of the forming vesicle, generating the free energy needed to restructure the lipid bilayer during scission.[479] This result was confirmed using phylogenetics, as Vps32, Vps20 and Vps60 are related Snf7 domain-containing proteins, and therefore the *P. lacertae* Snf7 domain protein could potentially be a divergent Vps32. However, it is likely to be an orthologue of Vps60, based on robust phylogenetic support, placing it in a clade with other known Vps60 sequences. Based on this, and the absence of most of the ESCRT II machinery and Vps20, it is likely that *P. lacertae* is not capable of forming multivesicular bodies. Although *Blastocystis* sp. has retained more of the ESCRT III and IIIA components, its ability to form MVBs is also unclear, based on the complement of ESCRT machinery identified. The pattern of ESCRT subunit loss does not appear to be wholly progressive in *Blastocystis* sp. and *P. lacertae*. Only two subunits appear to have been lost in both taxa (Vps28 and Vta1); CHMP7 was also not identified, but it has been independently lost many times in eukaryotes.[226] All other losses are specific to either *Blastocystis* sp. or *P. lacertae*, suggesting that ESCRT complex degradation is occurring largely independently in these organisms.

ESCRT complex loss has been observed in other taxa, such as the Apicomplexa[50] and the haptophyte algae, as discussed in Chapter 5. In these lineages, ESCRTs I and II are lost first, with patchy losses of proteins in the ESCRT III and IIIA complexes. This suggests that the later-acting ESCRT subunits may have other cellular roles, and perhaps may retain their original role in cytokinesis. In the Apicomplexa and haptophytes, the degradation of the ESCRT machinery is

accompanied by the loss of the Exocyst MTC, but this is not the case in *Blastocystis* sp. and *P. lacertae*. ESCRT loss is not specific to parasites or commensal organisms. Its loss suggests that the normal degradation of transmembrane receptors may occur via targeting for lysosomal destruction through a currently unknown method, or are simply recycled to the TGN and plasma membrane. Perhaps the loss of ESCRT complexes is an indication that these organisms do not regulate the presence of transmembrane proteins in the plasma membrane specifically through uptake and degradation.

The near-complete loss of Dsl1 and the SNARE it specifically binds to, Use1, in *Blastocystis* sp. is typical of peroxisome-lacking organisms. These include the Apicomplexa, *Trichomonas*, *Entamoeba*, *Giardia*, *Thalassiosira*, and the microsporidian *Encephalitozoon*.[480] Trypanosomatids also lack complete Dsl1 complexes, and they have highly modified peroxisomes known as glycosomes.[481] In the *Blastocystis* genome paper published by Gentekaki *et al.* (2017),[466] searches for peroxisome biogenesis machinery in *Blastocystis* sp. revealed a minimal complement of pex proteins. However, the Dsl1 complex is nearly complete in *P. lacertae*, suggesting it may have more peroxisomal genes. Further comparative genomic searches for these will confirm whether the Dsl1-perixosome pattern holds true in this lineage. The minimal peroxin complement of *Blastocystis* sp. and the minimal ESCRT machinery of both *Blastocystis* sp. and *P. lacertae* suggest that peroxisomes and MVBs have either been lost, or are in the process of being lost. The loss of peroxisome biogenesis machinery may be related to obligate anaerobism of *Blastocystis* sp., although it is certainly not necessary, as several organisms with divergent or lost peroxisomes are neither anaerobic nor parasitic.[482] The trend of Dsl1 and peroxisome loss or modification in eukaryotes suggests that in general, Dsl1 is

expendable in terms of its role in fusing Golgi-derived vesicles at the ER, contrary to what is observed in yeast, where this function is essential.[483]

Gene family expansions of exocytic and endocytic machinery are extensive in *Blastocystis* sp. and *P. lacertae*. In general, these paralogues likely add complexity to the membrane trafficking system in these organisms. They may be swapped in and out of complexes, or form specific subcomplexes, generating multiple trafficking factors with unique cargo preferences, interacting partners, or activity kinetics, thus adding complexity to the trafficking system. One such example of how this might work is the COPI subunit gene duplications in *Blastocystis* sp., where there are multiple copies of the alpha, beta-prime, and delta subunits. In human cells, the alpha and beta prime subunits bind cargo with a KKXX motif,[484] found in ER resident proteins, while the delta subunit binds cargo with a $Wx_{n(1-6)}$[WF] motif.[485] Multiple paralogues of these cargo-binding subunits may generate COPI subpopulations with different specificity for different cargoes. Unfortunately, because of the quality of the *C. roenbergensis* transcriptome, it is not possible to assess the effect of endocytosis/phagocytosis on the membrane trafficking machinery in this lineage.

Overall, the endocytic and TGN-endosome recycling machinery has been duplicated in *Blastocystis* sp. and *P. lacertae*, suggesting complexity in these pathways in both organisms. However, while there are two duplicated subunits of the HOPS/CORVET complex in *P. lacertae*, the expansions are more extensive in *Blastocystis*. These MTCs promote fusion of endosomes and fusion of late endosomes and lysosomes. HOPS specifically is involved in autophagosome-lysosome fusion. At first glance, analysis of the autophagy machinery did not suggest that this process is even possible in *Blastocystis* sp. and *P. lacertae*, unless they use other lineage-specific machinery, or that the canonical machinery is highly divergent. Despite these

missing factors, Yin and colleagues (2010) observed autophagosomes in *Blastocystis* sp.[486] Curiously, these organelles were potentially within the central vacuole, raising the question of how the central vacuole contributes to autophagy, and indeed how movement of organelles destined for degradation into the central vacuole occurs. Novel autophagy biology may also implicate non-canonical machinery in this cellular process, especially in the absence of typical vesicle expansion proteins.

It is likely that some form of autophagy still occurs in these organisms, although similar extreme losses of machinery have been observed in *Giardia*, the red alga *Cyanidioschyzon merolae*, and the microsporidian *Encephalitozoon cuniculi*; the latter two organisms having highly reduced genomes.[473] Perhaps this is evidence of genomic streamlining in *Blastocystis* and *P. lacertae*, in contrast to the gene family expansions observed in the membrane trafficking system.

But then what could be the function of duplicated HOPS/CORVET MTCs in *Blastocystis* sp. alone? It is possible that it plays a role in trafficking to the central vacuole that is only present in *Blastocystis.* The large central vacuole is made up of accumulated carbohydrates and lipids,[462] and storage organelles of this type are often modified lysosomes termed 'lysosome-related organelles', or LROs (e.g. melanosomes in human cells, acidocalcisomes in trypanosomatids, and secretory granules in the Apicomplexa, to name a few). The HOPS complex has been implicated in LRO biogenesis in a diversity of organisms,[487–489] and it may share this function in *Blastocystis* sp. Curiously, the AP3 complex is often involved in LRO biogenesis, however, it is the only adaptin complex that has not undergone gene duplication events. However, this does not preclude its functioning in this process. Additionally, COPI and ArfGAP1 have undergone

expansions in *Blastocystis* sp., and are both involved in lipid droplet formation in human and yeast cells. It is possible that they may also play a role in lipid storage at the central vacuole.

Remarkably, there do not appear to be any clear hallmarks of parasitism in the membrane trafficking system of *Blastocystis* as distinct from *P. lacertae*. These would include reductions or expansions of membrane trafficking machinery in *Blastocystis* sp. alone that may point to genomic streamlining or system specialization, respectively. A singular potential example of this may be the presence of the secretory SNAREs NPSN and Syp7, which are not identified in *P. lacertae*. As a key factor in *Blastocystis* pathogenesis is the secretion of cysteine proteases,[461] a robust secretory system including these SNAREs may be relevant in their secretion. One interpretation of these results is that in general, endobiosis exerts a greater effect on membrane trafficking system sculpting in *Blastocystis* and *Proteromonas* than does parasitism. However, more comparative genomic analyses of lineages with host-associated and free-living organisms are necessary to determine whether this claim is more generally relevant.

Blastocystis, *Proteromonas* and *Cafeteria* are deep-branching Stramenopiles, and using their genomes, one can probe the evolution of the membrane trafficking system at the base of this clade. One case of retention in this lineage is the presence of the third ancient paralogue of Sec24 (COPII) in *Blastocystis* sp. and *P. lacertae*. Until now, the only SAR genome shown to contain this paralogue was that of *Ectocarpus siliculosus*, a Stramenopile, and *Tetrahymena thermophila*, a ciliate. Its identification in *Blastocystis* and *Proteromonas* provides further evidence that these are not false-positive results, or likely to be the result of lateral gene transfer from other eukaryotes. Secondly, the entire TSET complex is found in both *Blastocystis* sp. and *P. lacertae*; while subunits of the TSET complex are found in Stramenopiles,[144] this is the first example of the whole complex being retained in this lineage. Finally, a putative plant-specific

TBC RabGAP protein may have been identified in *C. roenbergensis*. Two clades of plant-specifc TBCs have been identified; PlA and PlB.[266] Low node support values make this sequence difficult to classify, but it may be evidence that plant-specific TBC proteins were found in members outside of the Archaeplastida, but have since been lost.

From this analysis, the effect of anaerobism and symbiosis/parasitism on the membrane trafficking system and associated cellular systems (autophagy and peroxisome biogenesis) can be reductive. However, gene family and complex expansions in the membrane trafficking system co-occur with these loss events. Why have the parasitic/symbiotic *Blastocystis* sp. and *P. lacertae* retained and expanded some membrane trafficking machinery while other systems are in the process of being lost? These patterns may be the result of adaptation to the niche of living in the gut; however, a full genome of the free-living *C. roenbergensis* is necessary to confirm this. There may be less pressure to streamline the genome than on other gut microbes like *Giardia*, but losses of whole systems appear not to have an overwhelmingly negative effect on these organisms. In this light, recent work by Eme and colleagues (2017) show that lateral gene transfer from bacteria is critical for *Blastocystis* sp. to colonize the gut environment.[466] It remains to be seen whether these are specific to *Blastocystis* sp., or are also found in *P. lacertae* and *C. roenbergensis*. The inherent incompleteness of the *C. roenbergensis* transcriptome may give the impression that these gene duplications are specific to the endobionts. The possibility remains that these events may predate the split of *C. roenbergensis*. Nonetheless, these comparative analyses have generated predictions about how the membrane trafficking functions in an extremely common and enigmatic human gut parasite.

4. Membrane trafficking system function during cyst formation in *Entamoebae invadens* and *Entamoeba histolytica*

4.1 Introduction

Comparative genomics can be used to make predictions about how the membrane trafficking system is involved in specialized cellular processes. One way to give functional weight to these predictions is to look at the gene expression changes associated with a specific process. This type of differential gene expression analysis was done to investigate how the membrane trafficking system in modulated in cyst formation in *Entamoeba* spp., a parasite of animal gastrointestinal tract.

Entamoeba is a member of the supergroup Amoebozoa, and all members of the genus are considered to be pathogenic. Most are gut parasites, the most famous being *E. histolytica*.[490] It is thought to infect 1 in 10 people; in one year approximately 40-50 million people develop amoebic colitis or abscesses, and ultimately, it can cause up to 100,000 deaths per year.[490,491] It is a major issue in tropical areas; within the Indian subcontinent, Africa and South and Central Americas.[492] The majority of *E. histolytica* infections are asymptomatic, but 10% of cases can become invasive[491], with symptoms including dysentery, colitis, and amoebic granulomas.[38] If during infection the intestinal wall is perforated, amoebae can spread to the liver and brain causing abscesses, to the heart causing pericarditis, and to the lungs causing pleuropulmonary disease.[38] Invasive infection and dysentery is commonly treated with nitroimidazole-derived drugs,[493] the aminoglycoside paromomycin, but surgical intervention may be necessary in cases of fulminant colitis.[494,495]

E. histolytica parasites are transmitted by ingestion of cyst-contaminated water. While the average infectious dose is >1000 organisms, a single infectious individual can pass up to 45 million cysts in their stool per day.[496] In the environment, cysts must survive the drying, nutrient poor, and thermo-variable conditions of the environment for several weeks to months, and then

the harsh acidity of the stomach and the majority of the small intestine, rich with digestive enzymes, as excystation occurs later in the terminal ileum.[497] After excystation, the *Entamoeba* trophozoite divides and colonises the colon. The amoebae encyst prior to excretion; however, little is known about what is occurring in the cell on a molecular level to facilitate this life stage conversion, and how the quiescent cyst is prepared for eventual excystation.

The process of forming a stable, resistant cyst relies on the specific, ordered secretion of glycoproteins and proteins. It is thought that the cyst wall is assembled in three phases, termed the "wattle-and-daub" model.[342] In the foundation stage, the lectin Jacob is trafficked to the cell surface and there binds constitutively expressed Gal/GalNAc lectins.[342,498] In the "wattle" stage, chitin is synthesized and secreted, likely crosslinked by Jacob's tandemly arranged chitin-binding domains. Jacob lectins and chitin are seen in separate vesicles in early encystation (12 hours post induction, hpi), and have begun to accumulate at the cyst wall at 24-36 hpi. Although the timing is not clear, the chitin-cleaving enzyme chitinase and deacetylases trim and deacetylate extracellular chitin.[343,499] Finally, the Jessie3 lectin, which binds chitin and may also self-aggregate, solidifies the cyst wall in the "daub" phase, making it impermeable to small molecules.[342] It is observed in vesicles beginning at 36 hpi, and is found in the cyst wall at 72 hpi. The heavy secretory load clearly implicates the membrane trafficking system in this process.

Several studies have considered Rab GTPases in the context of encystation. Fourteen Rab genes, which are involved in vesicle fusion dynamics, were found to be up-regulated during encystation by a microarray screen.[500] A targeted bioinformatic analysis has shown that *Entamoeba* has expanded its Rab protein family; it has nearly 100 Rab proteins, many of which lack homologues in human cells,[460] raising the possibility that some divergent Rab proteins may be potential targets for drugs abrogating cyst formation. However, a comprehensive assessment

of the full membrane trafficking gene complement of *Entamoeba* has not yet been done. To this end, a comparative genomic approach was taken to identify membrane trafficking genes in *Entamoeba* spp., in order to gain a more thorough understanding of how this system has evolved in another gut parasite, as well as to provide the background genomic information necessary to study membrane trafficking gene expression patterns during encystation. Encystation is obviously a secretory process, and therefore one would expect secretory genes to be up-regulated over the course of encystation, but it is not clear how the rest of the trafficking system is modulated to support heavy secretion.

As *E. histolytica* cysts do not form readily *in vitro*, the closely related model *Entamoeba invadens* is often used to study cyst formation. *E. invadens* is a reptile pathogen, and encystation is easily induced under axenic conditions by glucose starvation. To identify MTS gene expression patterns during this process, encystation was induced in *E. invadens* and RNA-Seq was performed on mRNA from the trophozoite (0 hpi early encystation stage (24 hpi), late encystation stage (48 hpi), and the mature cyst (72 hpi). Because there were no replicates of these samples, the expression of individual genes was not considered. Rather, membrane trafficking pathways were considered to be associated with encystation if several genes from that pathway had expression patterns showing up-regulation over the course of the four timepoints. This comparative genomic and transcriptomic analysis gives insight into: (i) how the membrane trafficking system of a second gut parasite has been sculpted as a consequence of its lifestyle, and (ii) how the process of encystation is supported by membrane trafficking pathways.

4.2 Specific Methods - Genomics and transcriptomics of encystation in Entamoeba

Comparative genomics of *E. invadens* and *E. histolytica* were performed using BLASTP. The *E. invadens* and *E. histolytica* proteomes (AmoebaDB, Release 28) were searched, as well as novel transcriptomic data generated in this analysis. Encystation of *E. invadens* IP-1 trophozoites was induced by transferring trophozoites in logarithmic growth phase to axenic glucose-free 50% LG-Y medium. RNA-Seq was performed on mRNA samples taken at 0 hours post-induction (hpi, trophozoite), 24 hpi (early encystation), 48 hpi (late encystation), and 72 hpi (mature cyst). In-depth methods can be found in Herman *et al.* 2017.[501]

The Trinity package[408,502] and edgeR[412] were used to determine the levels of gene expression using the mapped reads. Because only one mRNA sample from each condition was sequenced, edgeR was run with an assigned dispersion value of 0.4 (recommended for datasets with no biological replicates).[412] Transcripts with a false discovery rate < 0.01 and a minimum $\log_2$ fold change of 2-fold were classified as significantly differently expressed between two time points. As multiple transcripts were generated for a single gene locus, only the transcript with the highest average expression across conditions was considered for that locus in our analyses. Without biological replicates, one cannot make robust statistical conclusions about differentially expressed genes. To mitigate this, genes likely to be differentially expressed (exceeding the FDR and fold change cutoff) were clustered by expression pattern during encystation, and conclusions of membrane trafficking gene expression were only drawn in cases where several members of a pathway were identified in an expression pattern cluster. Significantly differentially expressed genes (FDR<0.01, fold change > 2) were clustered using *k*-means and MCL clustering (Online Appendix Table 4.1, 4.2).

E. Herman performed all comparative genomic analyses in *E. histolytica* and *E. invadens* and assessed membrane trafficking system divergence between these genomes. *E. invadens* IP-1

trophozoites were grown by M. Siegesmund. RNA-Seq was performed by L. Caler, and gene expression and clustering analyses were performed M. Bottery and R. van Aerle. E. Herman reconstructed membrane trafficking pathways associated with encystation based on trafficking gene inclusion in expression pattern clusters. As the work on *Entamoeba* encystation has been published in Herman *et al.* (2017),[501] sections of the introduction, methods, results, and discussion are reproduced here.

4.3 Membrane trafficking gene expression during encystation in *Entamoeba invadens:* a model for *Entamoeba histolytica* infection

4.3.1 Comparative genomics of the membrane trafficking system in *Entamoeba histolytica* and *Entamoeba invadens*

In addition to assessing the evolution of the membrane trafficking system in a eukaryotic gut parasite, the secondary purpose of comparative genomics was to ensure that *E. invadens* is similar enough in gene content to *E. histolytica* that it can be used as a model system to study encystation. As both *Entamoeba* spp. are pathogenic, the focus of this analysis is on shared gene presence, gains, or losses between the two species. Figures 4.1 and 4.2 show the vesicle formation and fusion machinery presence and paralogue number in the two *Entamoeba* genomes. Online Appendix Table 4.3 contains all orthologue accessions. Out of 367 membrane trafficking genes identified, 246 have 1:1 orthology between *E. histolytica* and *E. invadens*. Furthermore, if the Arf and Rab regulators are excluded, there are only 11 genes that vary in paralogue number.

Figure 4.1. Comparative genomic survey of vesicle formation machinery in *Entamoeba invadens* and *Entamoeba histolytica*.

The vesicle formation machinery includes the ESCRTs, adaptor proteins, COPI, COPII, clathrin and endocytic machinery, retromer, TSET, and the Arf GTPases and their GAP and GEF regulators. To highlight paralogue number differences between the two *Entamoeba* species, darker versus lighter segments indicate more or fewer paralogues, respectively.

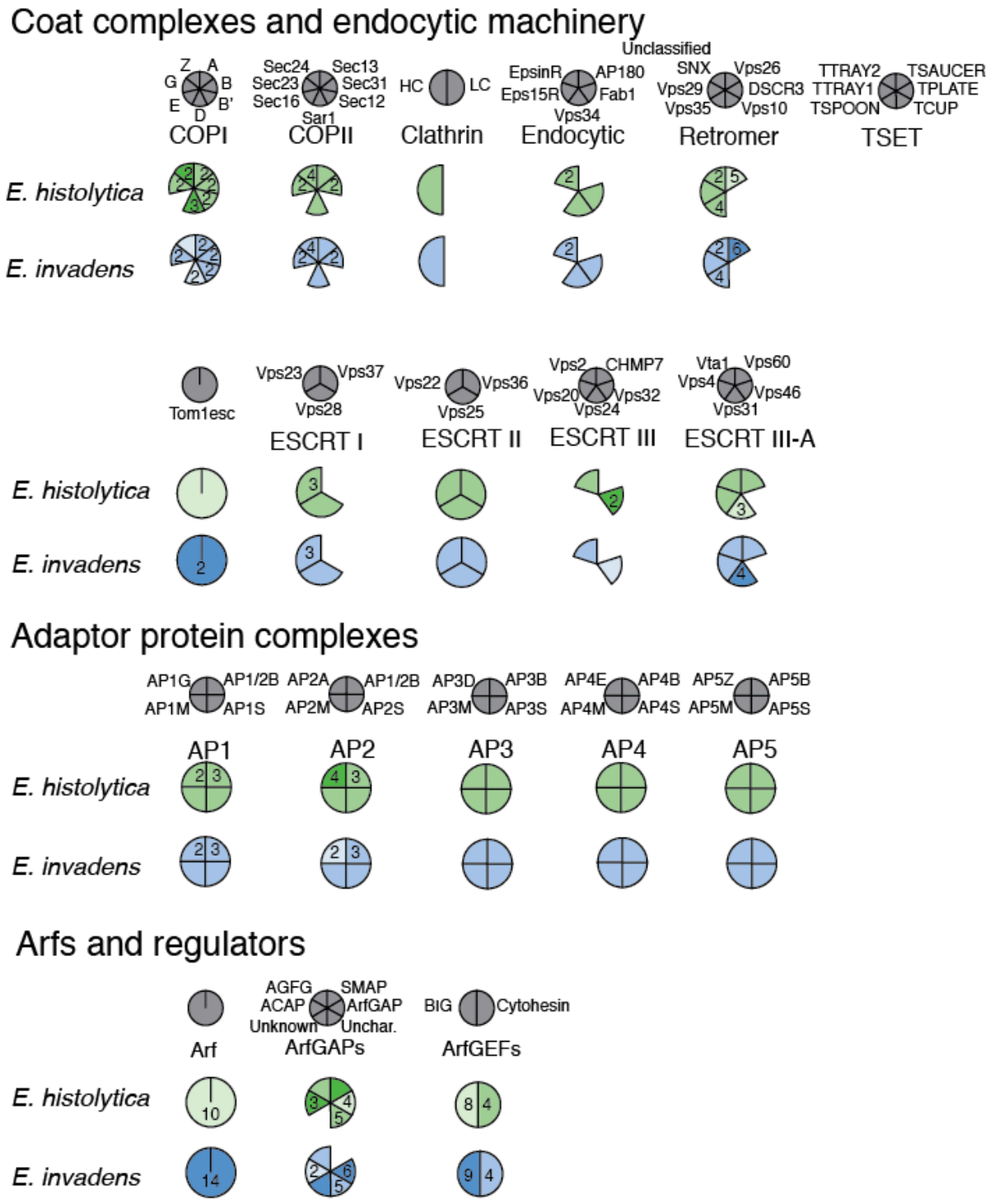

Coat complexes and endocytic machinery
Z A G B E B' D
COPI
Sec24 Sec13 Sec23 Sec31 Sec16 Sec12 Sar1
COPII
HC LC
Clathrin
EpsinR AP180 Eps15R Fab1 Vps34
Endocytic
Unclassified
SNX Vps26 Vps29 DSCR3 Vps35 Vps10
Retromer
TTRAY2 TSAUCER TTRAY1 TPLATE TSPOON TCUP
TSET
E. histolytica
E. invadens
Tom1esc
Vps23 Vps37 Vps28
ESCRT I
Vps22 Vps36 Vps25
ESCRT II
Vps2 CHMP7 Vps20 Vps32 Vps24
ESCRT III
Vta1 Vps60 Vps4 Vps46 Vps31
ESCRT III-A
E. histolytica
E. invadens
Adaptor protein complexes
AP1G AP1/2B AP1M AP1S
AP1
AP2A AP1/2B AP2M AP2S
AP2
AP3D AP3B AP3M AP3S
AP3
AP4E AP4B AP4M AP4S
AP4
AP5Z AP5B AP5M AP5S
AP5
E. histolytica
E. invadens
Arfs and regulators
Arf
AGFG SMAP ACAP ArfGAP Unknown Unchar.
ArfGAPs
BIG Cytohesin
ArfGEFs
E. histolytica
E. invadens

Figure 4.1

Figure 4.2. Comparative genomic survey of vesicle fusion machinery in *Entamoeba invadens* and *Entamoeba histolytica*.

Vesicle fusion machinery includes the SNAREs and SM proteins, the multisubunit tethering complexes, and Rab GTPases and their GAP and GEF regulators. As in Figure 4.1, paralogue differences are highlighted by darker versus lighter segments, indicating more or fewer paralogues in one taxon versus the other. For Rab proteins, the total number of rab paralogues is shown above the number of species-specific Rab paralogues in brackets.

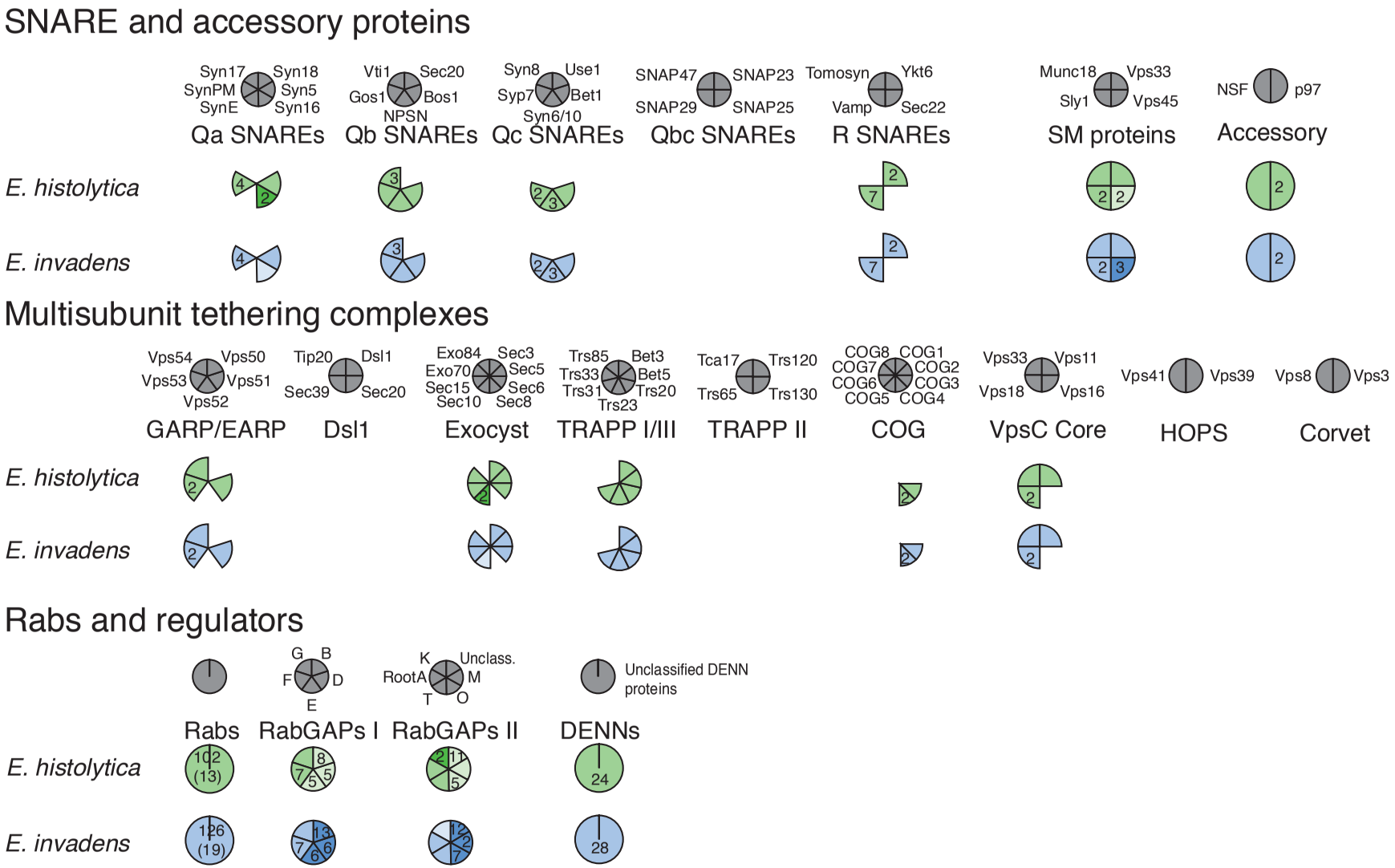

SNARE and accessory proteins
Syn17
Syn18
SynPM
Syn5
SynE
Syn16
Qa SNAREs
Vti1
Sec20
Gos1
Bos1
NPSN
Qb SNAREs
Syn8
Use1
Syp7
Bet1
Syn6/10
Qc SNAREs
SNAP47
SNAP23
SNAP29
SNAP25
Qbc SNAREs
Tomosyn
Ykt6
Vamp
Sec22
R SNAREs
Munc18
Vps33
Sly1
Vps45
SM proteins
NSF
p97
Accessory
E. histolytica
E. invadens
Multisubunit tethering complexes
Vps54
Vps50
Vps53
Vps51
Vps52
GARP/EARP
Tip20
Dsl1
Sec39
Sec20
Dsl1
Exo84
Sec3
Exo70
Sec5
Sec15
Sec6
Sec10
Sec8
Exocyst
Trs85
Bet3
Trs33
Bet5
Trs31
Trs20
Trs23
TRAPP I/III
Tca17
Trs120
Trs65
Trs130
TRAPP II
COG8
COG1
COG7
COG2
COG6
COG3
COG5
COG4
COG
Vps33
Vps11
Vps18
Vps16
VpsC Core
Vps41
Vps39
HOPS
Vps8
Vps3
Corvet
E. histolytica
E. invadens
Rabs and regulators
G
B
F
D
E
RabGAPs I
K
Unclass.
RootA
M
T
O
RabGAPs II
Unclassified DENN proteins
Rabs
DENNs
E. histolytica
E. invadens

Figure 4.2

As the Arf and Rab GTPase regulators are highly paralogous families, phylogenetic trees were generated to determine 1:1 orthologues and the nature of any expansions in these families. These are summarized in Table 4.1, supported by work shown in Online Appendix Figures 4.1-4.10 (Online Appendix Table 4.4). With the exception of the Arfs, more than half of these proteins have direct 1:1 orthology between *E. invadens* and *E. histolytica*, while less than 1/3 of the proteins are 'singletons' (i.e. one or more gene family members in one *Entamoeba* species with no clear orthologues in the other). The rest of the proteins make up orthologous groups with some expansion or loss events (i.e. 1+:1 or 1:1+ orthology). Overall, the vast majority of GTPases and their regulators share orthology between the two species (>75%), although there have been multiple independent expansions in both species. While there is divergence in this one sub-system of the membrane-trafficking machinery between the two organisms, the overall similarity of the remaining MTS complement in *E. histolytica* and *E. invadens* is high. 130 of 155 genes have 1:1 orthology, and there are 61 cases where a factor deduced as present in the LECA was not identified in either genome. This suggests that *E. invadens* is a good proxy for studying membrane trafficking gene expression during encystation in the human pathogen *E. histolytica*.

The vesicle formation machinery complement in *Entamoeba* spp. is largely complete (Figure 4.1). There are some partial losses of complexes, although these may be false negatives due to high sequence divergence. In both *E. invadens* and *E. histolytica*, the COPI and COPII complexes have multiple duplicated subunits. These coat complexes are involved in intra-Golgi and Golgi-ER trafficking in the case of COPI, ER-Golgi trafficking in the case of COPII. Both species encode multiple ArfGAP paralogues (4 in *E. histolytica* and 6 in *E. invadens*), which is

Table 4.1. Summary of orthology analysis of Arfs, Rabs, and their GAP and GEF regulators in *E. invadens* and *E. histolytica*

	ArfGAPs		Arfs		ArfGEFs		Rabs		TBCs		DENNs	
INVADENS : HISTOLYTICA	**#**	**%**	**#**	**%**	**#**	**%**	**#**	**%**	**#**	**%**	**#**	**%**
1 : 1	8	50.00	2	22.22	6	54.55	53	48.62	31	65.96	16	69.57
1 : 2+	1	6.25	0	0.00	1	9.09	7	6.42	2	4.26	0	0.00
2+ : 1	1	6.25	3	33.33	2	18.18	13	11.93	7	14.89	1	4.35
2+ : 2+	2	12.50	2	22.22	2	18.18	2	1.83	2	4.26	2	8.70
0 : 1+	1	6.25	1	11.11	0	0.00	10	9.17	1	2.13	1	4.35
1+ : 0	3	18.75	1	11.11	0	0.00	24	22.02	4	8.51	3	13.04
sum	16	100.00	9	100.00	11	100.00	109	100.00	47	100.00	23	100.00

	ArfGAPs		Arfs		ArfGEFs		Rabs		TBCs		DENNs	
INVADENS : HISTOLYTICA	**#**	**%**	**#**	**%**	**#**	**%**	**#**	**%**	**#**	**%**	**#**	**%**
1+ : 1+ orthology	12	75	7	77.78	11	100	75	72.82	42	89.36	19	82.61
0 : 1+ or 1+ : 0	4	25	2	22.22	0	0	31	27.18	5	10.64	4	17.39
sum	16	100	9	100	11	100	106	100	47	100	23	100

the GTPase activating protein for Arf, which functions in COP1-mediated vesicle formation. ArfGAPs are classified in Supplementary Figure S4.1 (Online Appendix Table 4.4). This subunit expansion suggests potential subfunctionalization in the early secretory system. The TSET complex was not identified, although it is patchily retained in eukaryotes, so this loss is not surprising.

Clathrin-mediated endocytosis proteins such as clathrin heavy chain AP2 are present, and there are duplications in both large subunits of AP2. Because of the duplications in this and other GADEZ subunits, phylogenetic studies were performed to classify them (Supplementary Figures S4.2, Online Appendix Table 4.4). The clathrin light chain was not identified, although this may be a false negative due to the short light chain sequence and high divergence of *Entamoeba*, it may also be truly missing, as the light chain is regulatory rather than structural, and may not be necessary for triskelion formation.[470] Clathrin can also function at the TGN with AP1, EpsinR, and Eps15R. Like AP2, the large subunits of AP1 (including the shared AP1/2B) are duplicated in both organisms, as is EpsinR. While there does not appear to be a clear Eps15R homologue, other ENTH domain proteins have been identified in *Entamoeba* spp.

The ESCRT complexes are present, although several subunits are missing in *Entamoeba* spp. These include Vps37, Vps20, Vps24, CHMP7, and Vps46. Vps37 is part of a structural ESCRT I stalk that recruits ESCRT II components. Vps20 and Vps24 initiate and complete vesicle scission, respectively, while Vps46 regulates the ATPase Vps4 that removes ESCRT components from membranes after scission. One Snf7 protein was identified that could potentially be a Vps20 orthologue (Supplementary Figure S4.3, Online Appendix Table 4.4). However, phylogenetic analysis shows that this sequence is excluded from a clade with other Vps20 sequences. Although it appears to group with Vps32 homologues, it does not form a well-

supported clade with these sequences, and thus its identity is unclear. Despite these losses, it is likely that they are functional and able to form multivesicular bodies. The AP5 complex, which generally has a patchy distribution in eukaryotes, is present and complete; it is thought to transport protein cargo to lysosomes. AP3 has a similar role in trafficking cargo, and it is also present and complete in both *Entamoeba* spp.

Both species also have multiple paralogues of the Vps26 and Vps35 cargo-binding subunits of the retromer coat complex. While generally in eukaryotes, retromer functions in recycling internalized receptors to the cell surface via the TGN, in *Entamoeba*, retromer has been shown to be targeted to pre-phagosomal vacuoles;[503] which are a modified endo-lysosomal organelle specific to *Entamoeba* that is important for pathogenesis. It may be that these multiple paralogues allow for retromer complex subfunctionalization, which could support the generation of this divergent organelle. The two TGN-associated adaptor protein complexes, AP1 and AP4 are both present, and as mentioned above, two of the four AP1 subunits have undergone duplications in this lineage.

In general, the vesicle fusion machinery complement is less complete than that of the vesicle formation machinery in *Entamoeba* spp. (Figure 4.2). There are several examples of losses in the vesicle fusion machinery that are common to both *Entamoeba* spp. Neither the SNARE proteins associated with retrograde Golgi-ER trafficking (Syntaxin 18, Sec20, Use1, Sec22), nor the MTC Dsl1, appear to be present. The loss of Dsl1 is corroborated by the complete absence of peroxisomes in *Entamoeba*.[504] For the MTC responsible for intra-Golgi trafficking, COG, only two of the eight subunits could be found. These potential losses raise questions about the nature of vesicle fusion in the ER-Golgi trafficking pathway, particularly

since vesicle formation machinery is present and in some cases expanded in both taxa (the COPI coat and the SNARE fusion protein Sly1).

There is evidence of gene duplication in the endosome-to-TGN trafficking pathway, including Syntaxin 16, Vti1, Syntaxin6/10, VAMP-like proteins, the SM protein Vps45, and subunits of the GARP MTC. These duplications echo the expansion of the retromer coat subunits in both organisms. Because of the duplications in Syntaxin 6/10 proteins, and apparent loss of other Qc SNAREs such as Use1, the Qc subfamily was classified using phylogenetics (Supplementary Figure S4.4, Online Appendix Table 4.4). The TBC RabGAP subfamily TBC-E is expanded in *Entamoeba*, which interacts with Rab35 in human cells, and may play a role in fast endosome-plasma membrane recycling. TBC family proteins are classified by phylogenetics in Online Appendix Figures 4.1-4.2 (Online Appendix Table 4.4). Curiously, fusion machinery of the endosomal maturation pathway is somewhat incomplete, as both syntaxin 7 and syntaxin 8 could not be found, nor were the HOPS or CORVET accessory proteins (although some subunits of the VpsC core are retained). However, the TBC RabGAP proteins that function with endolysosomal Rab proteins have undergone extensive expansion. TBC-B and TBC-F subfamilies, which are GAPs of late endosomal Rab7, have between 7-13 members. TBC-O, which interacts with Rab5 at early endsomes[505] has 5-7 copies in *Entamoeba* spp. The expansion of RabGAP subfamilies that function in endosomal maturation is at odds with the loss of the HOPS/CORVET-specific interacting subunits, and it raises the question of whether there are novel, lineage-specific factors that interact with the VpsC core proteins.

VAMP-like proteins and plasma membrane syntaxins have both been duplicated; these proteins are SNAREs that promote vesicle fusion at the PM. The exocytic SM protein Sec1/Munc18 is present in both organisms, as is a partial but potentially functional exocyst

MTC. The plasma membrane SNARE expansion raises the possibility that this extra machinery may support the secretion of virulence factors and/or cyst components.

4.3.2 Gene expression in *E. invadens* during encystation

To gain insight into membrane trafficking system modulation during encystation, RNA-Seq was performed on different stages of encysting *E. invadens*: in the trophozoite (0 hours post induction, hpi), the early cyst (24 hpi), the late cyst (48 hpi), and the mature cyst (72 hpi). Transcriptome sequencing resulted in 40 to 160 million reads per time point. These reads were mapped to the *E. invadens* IP-1 genome with PASA identifying 116 potential new transcripts. The *E. invadens* genome contains 11,553 predicted proteins, and only 473 have expression values of 0 FPKM at all time points. Therefore, approximately 96% of the *E. invadens* genome is transcribed at least at one time point during encystation. Of the 116 novel transcripts, only 26 retrieved any other sequence when used to search the NCBI non-redundant nucleotide and protein databases, and 11 of these are related to transposable element sequences (Online Appendix Table 4.1).

Statistical pair-wise comparisons of each time point were conducted to identify significantly differently expressed genes using edgeR.[412] This resulted in a set of 9,073 genes identified as differentially expressed between at least two timepoints, of which 4,987 were considered to be significant (defined as FDR <0.01, fold change >2, when using a dispersion of 0.4). As the dispersion between genes was set *a priori* rather than estimated from the data, expression values of individual genes are not necessarily reliable. Therefore, genes were clustered based on their expression patterns, and only membrane trafficking pathways with clearly co-regulated genes were considered to be meaningful. *k*-means and MCL clustering identified 10 clusters with

elevated expression at one or more timepoint (Figure 4.3). Three over-arching patterns of subclusters stood out: genes whose expression increases during early and/or late encystation, but decreases in the mature cyst; genes whose expression decreases during encystation; and genes whose expression increases in the mature cyst. The first pattern – genes up-regulated during encystation – includes subclusters 1, 2, 5, and 9. Within this group, it is possible to further distinguish genes whose expression remains constant in both early (24 hpi) and late encystation (48 hpi, subcluster 2) from those whose expression peaks in late encystation (subclusters 1, 5, and 9). Subclusters 3 and 8 are characterised by a general decrease in expression during the encystation process, including in the mature cyst (72 hpi). Finally, genes whose expression peaks in the mature cyst make up subclusters 4, 7, and 10. Subcluster 7 is particularly intriguing, as its members are generally down-regulated during early and late encystation, but return to expression levels equivalent to or slightly higher than in the trophozoite (0 hpi). Subcluster 10, on the other hand, shows a pattern of steadily increasing expression throughout encystation and in the mature cyst. Subcluster 6 shows a pattern of downregulation between the trophozoite stage and early encystation (24 hpi); however, the magnitude of this change is low relative to the other expression patterns. As it is not a robust pattern, it will not be considered in the following analyses.

To test how well this clustering method represents known encystation biology, encystation-specific proteins with verified gene expression patterns were searched for in the clusters. Jacob lectins are found in subcluster 9, except EiJacob6 and EiJacob7, which are found in subclusters 4 and 5, respectively. Jessie lectins are found in subcluster 5. This is congruent with the notion that the Jacob proteins act at an earlier stage, as subcluster 9 describes genes that have high expression during early and late encystation, and Jessie proteins act later, as subcluster

Figure 4.3 Clusters of gene expression patterns during *E. invadens* encystation.

Gene expression patterns for genes that are differentially expressed (FDR <0.01, fold change >2) between at least two timepoints during encystation were clustered using *k*-means and MCL clustering. 10 subclusters were generated, which fall into three categories: those whose expression increases during encystation, those whose expression decreases during encystation, and those whose expression increases in the mature cyst. Each grey line represents an individual expression pattern of one gene in that cluster. Red lines indicate the cluster's mean expression profile. Online Appendix Table 4.2 shows the cluster assignments of all *E. invadens* genes.

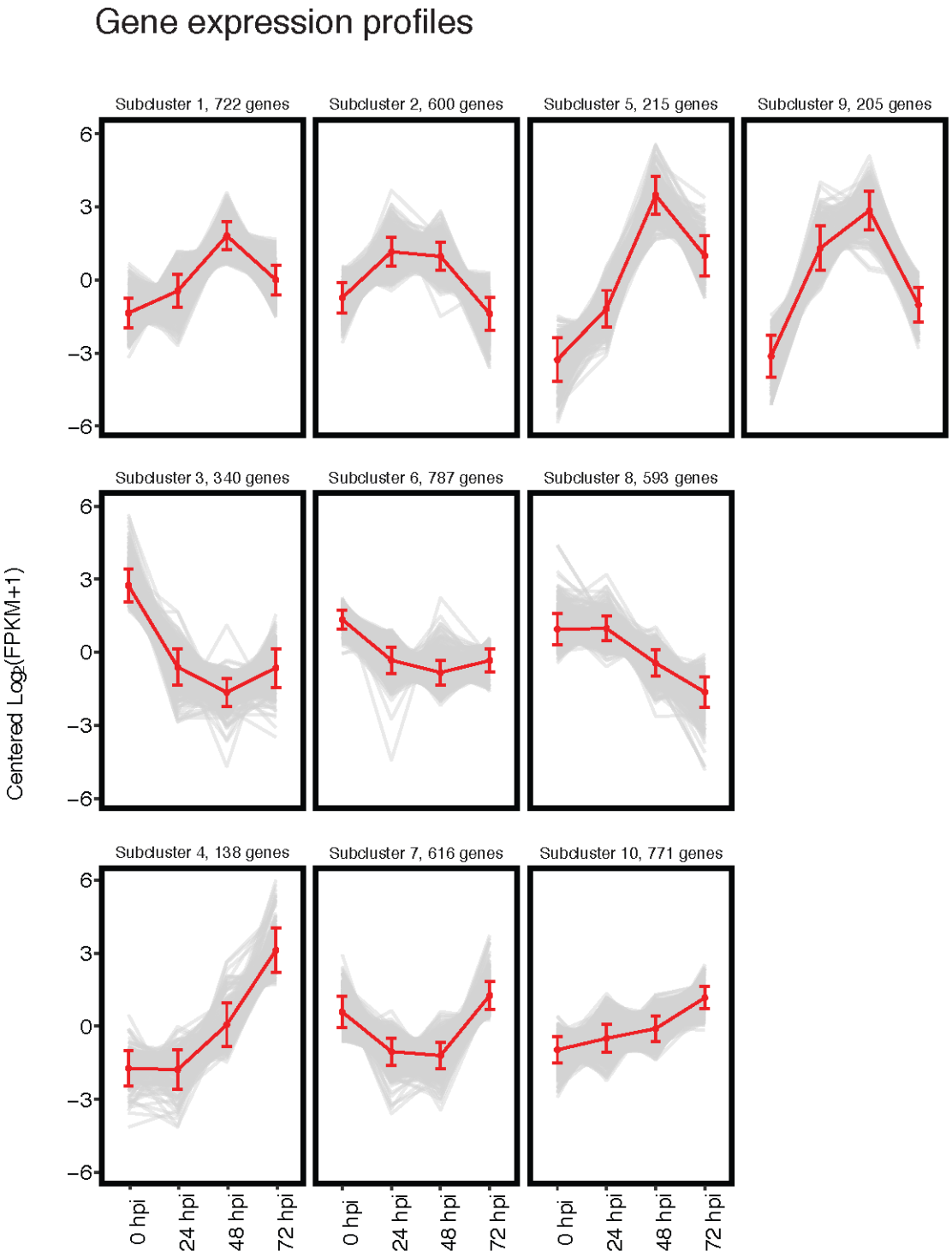

Gene expression profiles
Subcluster 1, 722 genes
Subcluster 2, 600 genes
Subcluster 5, 215 genes
Subcluster 9, 205 genes
Subcluster 3, 340 genes
Subcluster 6, 787 genes
Subcluster 8, 593 genes
Subcluster 4, 138 genes
Subcluster 7, 616 genes
Subcluster 10, 771 genes
Centered Log2(FPKM+1)
6
3
0
−3
−6
0 hpi
24 hpi
48 hpi
72 hpi

Figure 4.3

5 describes genes specifically up-regulated in late encystation. Furthermore, all four chitinases have expression patterns in agreement with previous RT-PCR work[506] and are also members of subclusters 5 and 9.

4.3.2.1 Membrane trafficking gene expression in *E. invadens*

Of approximately 400 MTS genes, 223 were identified as differentially expressed and could be clustered into the groups described above during encystation (Online Appendix Table 4.3). Out of ~400 trafficking genes identified in *E. invadens*, approximately ¼ encode multiple paralogues where two or more are found in different subclusters. This suggests that the paralogues may be functionally distinct, and may be swapped into protein complexes to modify trafficking during encystation. It suggests MTS modulation, rather than a blanket increase or decrease in trafficking pathway function.

Eighteen genes were identified in subclusters 1, 2, 5, or 9 that are involved in secretion show a pattern of up-regulation during encystation. These do not have paralogues with opposing expression patterns (Online Appendix Table 4.3). Several members of the COPII vesicle coat complex, responsible for the initial ER-to-Golgi step of trafficking secreted material, are up-regulated during encystation (Sar1[507] and two paralogues of Sec24). Intra-Golgi and Golgi-ER retrograde trafficking is also represented in these subclusters, which include two of the six subunits of the COPI vesicle coat complex required for intra-Golgi and Golgi-ER trafficking, four BIG-like ArfGEF paralogues, as well as several SNARE proteins that function in this pathway (Syntaxin 5, Gos1, Bos1, Bet1) and the SM protein Sly1, which helps in SNARE complex formation. Post-Golgi secretory SNAREs are members of these subclusters: Syntaxin

PM (two paralogues), and Syp7 as well as the cognate SM protein Munc18. Syp7 is absent in human cells, and therefore it may represent a therapeutic target to disrupt encystation.

Because of the large contingent of Rab proteins and the role of *Entamoeba*-specific Rabs in pathogenesis,[508] they were included in differential expression analyses. Rab1 and EhRabA, thought to function in early secretion,[508,509] appear to be down-regulated during encystation. However, they may play multiple roles in trafficking: while the *Entamoeba* Rab1 homologue has not been functionally characterized, Rab1 proteins in *Dictyostelium* cells are recruited to phagosomes,[510] and EhRabA has been shown to be involved in cell motility.[511]

The TGN-endosome recycling pathway is more modestly represented in subclusters 1, 2, 5, and 9, with five members. These include the SNARE protein Syntaxin 16, its cognate SM protein Vps45, EpsinR, and Fab1 (PIKFYVE in mammals), which as well as a subunit of the GARP multi-subunit tethering complex, Vps51, is involved in the tethering of vesicles during fusion.

Overall, there is a weak endocytic/phagocytic signal in subclusters 3 and 8 (five members), whose expression decreases during encystation. These include both paralogues of Rab8, two cytohesin-like ArfGEFs, and Rab5. Contrary to this, the phagocytic Rab EhRabB appears to be up-regulated during encystation. However, this gene is known to be activated by heat shock,[512] and may be activated as part of the stress response cascade that has been proposed to regulate encystation,[513] rather than as part of an increase in phagocytic machinery. The *E. invadens* genome has many genes with multiple paralogues, and it is often the case that while one paralogue of a gene appears to be down-regulated during encystation (in subcluster 3 or 8), the other paralogue does not change in expression. As such, there are few genes and pathways

that clearly exemplify a pattern of downregulation. Despite the high modulation and differential expression of MTS components, it is relatively rare for two paralogues of any given gene to have 'opposing' expression patterns, i.e. one paralogue is a member of a subcluster whose expression increases during encystation, and the other is part of a subcluster whose expression decreases under those same conditions. The majority of these cases occur in highly paralogous gene families, such as the TBC-B family of RabGAPs, or the exocytic SNARE protein Vamp7. Given the extensive duplications and diversification in these gene families, it would be inappropriate to use their subcluster membership as evidence of an emphasized MTS pathway.

Curiously, there are 57 MTS genes whose expression increases from late encystation (48 hpi) to the mature cyst (72 hpi; subclusters 4, 7, 10). The most obvious example is the ESCRT complex proteins, where 7 of 17 are found in these subclusters. The ESCRT complexes are responsible for multivesicular body biogenesis, a mechanism by which plasma membrane and cytosolic proteins can be targeted to the lysosome for degradation. Other members of these subclusters that are involved in endosome recycling and endo-lysosomal function include the adaptor protein complex AP1γ subunit (two paralogues), EpsinR, Syntaxin 6 or 10, two putative DENN2 paralogues, Rab7, TBC-B, and TBC-F. Clathrin heavy chain is also highly expressed in the mature cyst. There are 7 secretory pathway genes that make up these subclusters, including two members of the COPII coat complex (Sec23 and Sec31), two non-human SNARE proteins thought to function in ER-to-plasma membrane trafficking in plants (NPSN11 and Syp7), a plasma membrane syntaxin, the COPI subunit COPB', and two BIG-like ArfGEFs. Finally, there are three ArfGAP1, one SMAP, and three uncharacterized ArfGAP paralogues in these subclusters, although their interacting Arf proteins and therefore functions are unknown. Other small G proteins and their regulators that could not be classified, or whose function is unknown,

include five RabGEFs, three RabGAPs, two Arf proteins, and nine Rabs. These mRNAs are highly abundant in the mature cyst, although it is not clear whether they are translated into protein products at this stage. As the mature cyst is quiescent, it is possible that this is done to prepare for excystation, so that the cell can quickly perform early endocytic, exocytic and recycling functions prior to beginning the gene expression program seen in trophozoites. This may be analogous to increases in mRNA longevity seen in late stage trophozoites of the malaria parasite *Plasmodium falciparum*, which has been suggested to allow nascent merozoites to rapidly activate their development cycle upon invasion of a new erythrocyte.[514] However, two proteomic analyses of *E. histolytica* have identified several membrane trafficking proteins (among others) present in the cyst,[340,515] suggesting that at least some MTS proteins are retained.

4.4 Discussion

E. invadens is a commonly used model system in which to study cyst formation in the human parasite *E. histolytica*. The process of encystation involves secretion of various lectins and chitin through the membrane trafficking system. As the cyst form is required for pathogenesis, an analysis of membrane trafficking gene expression during encystation was undertaken to better understand how it supports this process, and if there are *Entamoeba*-specific factors that can be targeted to prevent encystation and therefore infection.

While *E. invadens* is a well-established proxy for studying encystation, the similarity in membrane trafficking gene complement between the two organisms was unknown. A comparative genomic analysis of the membrane trafficking system in the two species was done, showing that this system is highly similar in both organisms. Relative to the model systems of

humans, and yeast most vesicle formation machinery was identified. The exception to this was that several ESCRT III and IIIA components could not be identified. However, it is possible that these proteins are too divergent to identify using BLAST alone. In their absence, the partial ESCRTs may still be capable of generating multivesicular bodies. The AP5 complex, which is thought to function at MVBs, was found in both taxa, and several ESCRT components also show a pattern of up-regulation in the mature cyst, suggesting that ESCRTs are functional. While MVBs typically fuse with lysosomes to promote degradation, they can also be secreted via an unconventional pathway involving autophagy machinery.[516] Furthermore, the ESCRT III and IIIA complexes have non-trafficking membrane scission functions in other organisms (e.g. cytokinesis in plants, pre-peroxisome membrane budding in yeast), and may well have other functions during encystation or excystation.

As foreshadowed by the identification of over 100 Rab proteins in *E. histolytica*, the membrane trafficking system in *Entamoeba* is not streamlined as a result of a parasitic lifestyle. To the contrary, three arms of the trafficking system have undergone expansions: ER-Golgi transport, endosome-TGN transport, and to a lesser extent, post-Golgi secretion. Gene family expansions are a mechanism by which complexity can be generated in membrane trafficking. Particularly, cases where whole complexes have multiple paralogues of each subunit, such as COPI, raise the possibility of two distinct complexes, or multiple complexes with shared subunits that can be swapped in and out.

Complexity in the membrane trafficking system of *Entamoeba* may be related to pathogenesis. Cysteine proteases are a key pathogenicity factor in *Entamoeba* that are secreted to degrade mucin, destroying the barrier protecting the gut epithelium.[517] The secretion of cysteine proteases by *E. histolytica* as pathogenicity factors is regulated by Rab11[518] which in animal

cells is involved in endosomal recycling[519] and both constitutive and regulated secretion.[520] It is possible that the additional trafficking factors in the late secretory pathway and the endosome-TGN recycling pathway are involved in regulating the secretion of pathogenesis factors such as the cysteine proteases, as separate from other secretory processes. Another explanation may be that they are involved in trafficking to the pre-phagosomal vacuole, an *Entamoeba*-specific vacuolar compartment that stores digestive proteins, and is regulated by Rab5, Rab7 and the retromer coat complex (all of which have been expanded to some extent in both genomes).[503] Further experimental work is required to test these bioinformatics-generated hypotheses regarding paralogue function in *Entamoeba*.

Overall, vesicle fusion machinery is less well-conserved than vesicle formation machinery in *E. invadens* and *E. histolytica*. Particularly, Golgi-ER retrograde SNAREs and the MTC Dsl1 could not be identified, and the intra-Golgi MTC COG is highly reduced. As has been observed in works by others as well as in previous chapters, the loss of Dsl1 is often found in parallel with a loss of peroxisomes.[480] However, the loss of multiple Golgi-related vesicle-fusion factors is intriguing. Despite these losses, the COPI coat complex is nearly complete, and nearly all subunits have been duplicated in *Entamoeba* spp. *Entamoeba* has an unstacked Golgi, visualized as large vesicles near the nucleus by confocal microscopy of amoebae stained with anti-Arf antibodies.[345] While Golgi-to-ER and intra-Golgi trafficking pathways are likely functional, perhaps the alternative Golgi morphology of *Entamoeba* is related to the use of novel SNARE-like or tethering factors functionally analogous to those found in this pathway in other eukaryotes. These results suggest that Golgi-ER trafficking in *Entamoeba* differs from what is required in mammalian and yeast cells,[483,521] in that it can be accomplished without the function of certain SNARE complexes and MTCs that tether incoming vesicles to target membranes.

Despite this, paralogous expansions of the Rabs and Rab regulator proteins that are involved in vesicle fusion are encoded in *E. invadens* and *E. histolytica*.

The MTC complexes of the endocytic system are somewhat reduced, with subunit losses in GARP, which supports endosome-TGN trafficking,[522] and the TRAPPII-specific subunits of TRAPP (a late Golgi/TGN tethering complex),[523] and both CORVET and HOPS (early and late endosome/lysosome trafficking, respectively).[311] However, other MTS proteins that support these functions are present, such as the retromer coat complex, the heterotetrameric adaptor protein complexes 1-5, and clathrin. These pathways are likely functional, but may not require the canonical MTCs for vesicle fusion. In general, MTCs in *Entamoeba* spp. are either wholly lost or reduced; there is not a single complete complex that could be identified, raising questions about the necessity of MTC-based tethering in *Entamoeba*. If these complexes or subunits are truly missing (rather than highly divergent) with no unknown functional analogues, it could be an example of genomic reduction common to parasites, and it indicates that these complexes are not essential in *Entamoeba* spp.

With the exception of Arf and Rab GTPases and their regulators, the presence and number of paralogues of MTS proteins is relatively similar between the two *Entamoeba* spp. Because encystation is a feature of both organisms, it is not likely that gene paralogue innovation in either *Entamoeba* spp. is especially relevant to encystation. It is not unusual for the number of Ras family GTPases to vary in number in closely related taxa[524] and may instead be an adaptation to their respective hosts, or other lifestyle factors. Table 4.1 shows that one-to-one orthology was observed in at least 50% of the Rab GTPases, and Arf and Rab regulators, suggesting a balance of functional retention and innovation. Differences in the number of cognate GAP and GEF proteins may be related to differential Rab and Arf subfamily expansion.

However, there are six *E. invadens*-specific Rab genes that are members of subclusters 1, 2, 5, and 9 – which are up-regulated during encystation – suggesting they may be involved in the process, just as there are a number of *E. histolytica*-specific Rabs, whose expression pattern during encystation is unknown. Therefore, the possibility exists that there may be some differences in the trafficking events underlying encystation in the two *Entamoeba* species.

The secretory pathway and TGN-endosome recycling are the most well represented trafficking pathways represented in the subclusters whose expression increases during encystation. The secretory pathway supports Jacob and Jessie lectin secretion, and therefore cyst formation. This correlates with the increase in expression of chitin-cleaving chitinases during encystation (Herman et al. 2017). The role of the TGN-endosome recycling pathway in encystation is less clear. Encystation requires vesicles containing cyst-forming material to be transported to the cell surface. To achieve membrane homeostasis in the cell, extra membrane from this exocytic process must be retrieved. It may be that endosome-to-TGN trafficking functions to retrieve surface proteins that have been internalized with this membrane, thus preventing them from being trafficked to the lysosome for degradation.

Small monomeric G proteins and their regulators, i.e. Arf and Rabs, make up many of the differentially expressed genes that have a pattern of up-regulation during encystation. It suggests that these proteins are the 'gatekeepers' of cyst formation, as they control vesicle formation and fusion kinetics. Rabs in particular have been shown to be differentially expressed during cyst formation.[500] Other than Rabs 5, 7, 11, RabA, and RabB, little is known about Rab function in *Entamoeba*. Therefore, these unstudied Rabs represent potential therapeutic targets, in particular, the proteins that are highly expressed during encystation.

E. histolytica encodes four Rab11 paralogues. Rab11A is known to be involved in encystation,[525] and Rab11 paralogues are involved in protein recycling in both *Entamoeba* and human cells.[518,525,526] In *E. invadens*, Rab11A is highly but consistently expressed at time points during encystation. Rab11B is a member of subcluster 6, and appears to decrease in expression during encystation, and Rab11C and Rab11D expression increases in early encystation, peaks at late encystation, and drops again in the mature cyst (subcluster 2). This expression pattern implicates Rab11C and Rab11D in cyst formation.

Much of the work on *Entamoeba* Rabs has examined their involvement in phagocytosis and pathogenesis. Rab5 and Rab7, particularly, have been shown to be involved in phagosome and pre-phagosomal vacuole maintenance.[132,503,527,528] Three Rab7 paralogues are members of subclusters 1 and 2, suggesting that they may be involved in encystation, in contrast to other paralogues that have different expression patterns. This again lends support to the idea of a highly specialized MTS in *Entamoeba* where gene families encode functionally divergent paralogues. Curiously, the CORVET and HOPS complexes that specifically interact with these Rabs are not found in either *E. invadens* or *E. histolytica*. In the case of mammalian cells, CORVET is a Rab5 effector, and HOPS is a Rab7 GEF. This raises the question of what other factors may take the place of these MTCs as tethers and RabGEFs. The presence of a nearly complete VpsC core suggests that there may be other Rab-specific interacting partners similar to Vps3/8 and Vps39/41. However, other, distinct complexes or factors may be at play as Rab regulators or tethering factors.

The only clear pattern of downregulation during cyst formation is seen in some endocytic and phagocytic trafficking components, which is congruent with the fact that the amoeba is preparing to become a quiescent cyst. However, there are also many membrane trafficking genes

found in the up-regulated dataset. It is possible that mRNAs for endocytic and degradative trafficking genes are kept untranslated so that during excystation, they can be translated rapidly to aid in that process. An analogous system is the mRNA localization that occurs in *Drosophila* embryogenesis.[529] However, the mechanism by which mRNA is temporarily silenced in *Entamoeba* is unclear. Proteomics analyses of *E. histolytica* cysts have shown a handful of MTS proteins, among many others, present in mature cysts,[340,515] raising the possibility that these mRNAs may indeed be translated. Two proteins identified in other analyses were clathrin heavy chain and a TBC-B homologue (EHI_094140), which are both members of subcluster 10, whose expression peaks in the mature cyst. However, the difference between long-lived proteins generated during encystation and untranslated mRNA pools kept in the cyst for excystation cannot be determined by RNA-Seq alone. Regardless, it raises the question of a mechanism to prevent mRNA translation, degradation, or protein functioning in the quiescent cyst. Cysts can remain infective for weeks under optimal conditions,[530] suggesting that there is a mechanism by which mRNA (or protein) is stored in an inactive manner during this time. Interestingly, there is evidence for the accumulation of mRNA in cysts of the amoeba *Acanthamoeba*,[531] and the ciliate *Colpoda inflata*,[532] suggesting that mRNA storage in cysts occurs in other organisms, although the mechanisms have not yet been deduced.

A similar study was performed by Ehrenkaufer *et al.* (2013),[533] which assessed genome-wide transcriptional changes in *E. invadens* using RNA-Seq. They generated gene expression profiles, and identified Gene Ontology (GO) terms enriched in each profile or profile group. Genes annotated with the GO term 'vesicle-mediated transport' were found to be significantly enriched in profiles with a pattern of increased gene expression during encystation. A portion of the membrane trafficking genes analysed here were cross-checked with those of the Ehrenkaufer

and colleagues, and found the assigned expression patterns to be largely consistent with their profile assignments. Furthermore, they identified genes involved in RNA metabolism expressed in the mature cyst, potentially preconfiguring the cell for rapid excystation.

Several proteins were identified with no human orthologue. One of these, Syp7, is a SNARE protein specifically up-regulated during encystation. While further work is required to determine whether disruption of Syp7 expression affects cyst formation, it represents a class of potential drug targets. Limiting cyst formation would not treat a current infection, but it could reduce parasite spread, which is a critical issue in areas with poor sanitation.

5. Membrane trafficking evolution in the haptophytes and the biology of scale formation and secretion

5.1 Introduction

The previous two chapters explored the effect of the parasitism on membrane trafficking system evolution. However, unique environments or lifestyles can also generate diversity in the membrane trafficking system. In this chapter, the haptophyte algae are studied. Haptophytes are a monophyletic group of algae found in both marine and fresh waters. They are a major eukaryotic lineage, which, prior to 2013, had no sequenced representatives. The most up-to-date phylogenomic analysis suggest that haptophytes branch basal to the SAR clade, following the split of this lineage with the Archaeplastida.[463] In order to fill this taxonomic sampling gap, the genome of the haptophyte *Emiliania huxleyi* was sequenced[62] and the comparative genomics was used to annotate the membrane trafficking machinery. Following the genome project, this lineage offered the opportunity to study membrane trafficking dynamics in organisms with the unique cell biology of scale production. This latter project involved sequencing the genomes of several related haptophytes, and performing transcriptomic analysis of genes that are differentially expressed under scale-forming conditions.

Coccolithophores are members of the haptophyte clade. They cover their entire surface with scales (coccoliths) of calcium carbonate, and are thus a critical part of carbon cycling. In some systems, they are responsible for as much as 20% of total carbon fixation.[534] Their carbon regulatory role is complex, as they serve as both sequester inorganic carbon in $CaCO_3$-bound scales, and release CO_2 to the ocean surface layer.[347] Coccolithophore blooms can span over 100,000 kilometers, and due to sunlight reflecting off the coccoliths, they are visible from space.[535] As such, coccolithophores can have a dramatic impact on their local environment. In addition to playing a major role in carbon cycling, they produce dimethylsulfoniopropionate, the precursor to dimethyl sulfide, which causes cloud condensation,[536,537] exerting a general cooling

effect. This raises the question of how increased atmospheric CO_2 (and therefore ocean acidification) due to climate change will affect the ocean ecosystem, and particularly the coccolithophores. One 12-year study of *E. huxleyi* in the Mediterranean Sea show a long-term decrease in coccolith weight and therefore less calcification, most likely as a result of ocean acidification.[538] However, the extent of the environmental impact of coccolithophore thinning is not clear at this time.

The secretion of calcium carbonate scales begins with the production of baseplate organic scales generated in what appears to be a Golgi-related organelle, based on ultrastructural studies.[348] The organic scale – or body scale – is primarily composed of complex acidic polysaccharides.[539] Calcification occurs via a 'reticular body', which is closely associated with the coccolith vesicle, and in one haptophyte, is thought to be contiguous with the ER.[540] Recent work in *E. huxleyi* has shown that calcium is stored in a separate vacuole-like compartment in a disordered calcium-polyphosphate phase, and the authors suggest that calcium ions are released/transferred to the coccolith vesicle/reticular body system, in which the calcium precipitates as calcite.[541] Following calcification, the scales are exocytosed; scale formation and exocytosis occurs one at a time, at a rate of up to one scale per hour.[542] However, some haptophytes produce only organic body scales, either because they diverged prior to the advent of biomineralization in this lineage, or because they lost the ability to biomineralize. Additionally, calcification and scale formation can be dependent on cell type; some haptophytes including *E. huxleyi* have multiple cell types, which differentially secrete calcified scales or body scales.

These novel organelles associated with coccolith formation have been observed by microscopy, but because the haptophytes are not a system with established molecular tools, these

organelles cannot be experimentally characterized. However, multiple aspects of coccolith formation rely on membrane trafficking, and so this lineage may have a modified membrane trafficking gene complement to accommodate its unique cell biology. To add more depth to the initial analysis of membrane trafficking machinery in *E. huxleyi*, the genomes of two related haptophytes were sequenced: *Gephyrocapsa oceanica* and *Isochrysis galbana*.

E. huxleyi, G. oceanica, and *I. galbana* are members of the Order Isochrysidales, and are suggested to form a monophyletic lineage that is separate from other coccolithophorids, based on 18S rDNA and RubisCO gene phylogenies,[543,544] and because they share morphological, biochemical, and ultrastructural features.[545] They have complex life cycles with different cell types, and their production of coccolith and body scales is cell-type dependent. In addition to naked, non-motile cells (N cells), *E. huxleyi* can also be coccolith-bearing and non-motile (C cells), or motile and covered only in body scales (S cells).[546] Unlike most coccolithophores, the C cell stage of *E. huxleyi* does not have a body scale under-layer. The life stages and scale types in *E. huxleyi* are identical to those of *G. oceanica*; both can form non-motile, coccolith-bearing cells, and motile, haploid cells covered in body scales. Although coccolith morphology does differ between these two taxa, they share structural similarities that suggest a similar mechanism for biomineralization. In *E. huxleyi* and *G. oceanica,* scale formation occurs in small Golgi-derived vesicles that fuse, and remain directly apposed to the nucleus.[346] As the scale starts to form, a tubular, membranous reticular body forms near the coccolith-forming vesicle, and calcification occurs. The fully formed scale is then released from its nuclear-proximal position and secreted from the cell (reviewed in Paasche 2002).[547] *I. galbana*, on the other hand, has secondarily lost the ability to calcify, and contains only body scales, identical to those in *E. huxleyi* and *G. oceanica*.[346,548]

A fourth haptophyte outside of the Isochrysidales is *Chrysochromulina tobin*, a member of the Prymnesiales, and its genome is currently publicly available.[549] The Prymnesiales may have diverged prior to the introduction of biomineralization, although it has been suggested that they did have the ability to calcify, but lost it secondarily.[548] Nonetheless, *C. tobin* does produce organic body scales like other haptophytes. They appear to also be made in Golgi-like compartments;[550] however, multiple scales are produced at one time, as opposed to single scales in *E. huxleyi* and *G. oceanica*. Furthermore, *C. tobin* has retained the haptonema, a cytoskeletal organelle found only in the haptophytes. The haptonema is found near the flagella, and in Chyrsochromulina, the haptonema rapidly coils and uncoils to support attachment and potentially phagotrophic feeding.[551] *E. huxleyi* and *G. oceanica* lack a haptonema, while *I. galbana* has shorter, vestigial haptonema that may instead be used as an obstacle-sensing device.[346]

Scale formation in all four haptophytes relies on the membrane trafficking system, as is likely also the case for calcification in *E. huxleyi* and *G. oceanica*. This raised the question of whether the membrane trafficking system complement in the haptophytes has been modified to support these processes. With the genomes of two biomineralizing haptophytes (*E. huxleyi* and *G. oceanica*), and genomes of two haptophytes that either lost or potentially never had the ability to calcify (*I. galbana* and *C. tobin*), comparative genomics was performed to identify differences in membrane trafficking gene complement in these organisms.

However, differences in gene complement can only go so far in predicting functional association with biomineralization. In order to identify the membrane trafficking genes whose expression is regulated by the induction of scale formation, a comparative transcriptomic experiment was performed on the calcifying *E. huxleyi* and *G. oceanica*. These haptophytes were grown under the following four conditions, followed by mRNA extraction: with the addition of

calcium, with the addition of a spike of bicarbonate, with both calcium and bicarbonate, and with neither. Genes that are differentially expressed in *E. huxleyi* and *G. oceanica* grown in the presence of calcium and bicarbonate, compared to those grown in artificial seawater media, are likely to be involved in biomineralization. By also looking at gene expression with the addition of calcium alone or bicarbonate alone, the individual effects of these chemicals on gene expression can be determined.

The comparative genomic and transcriptomic analysis of membrane trafficking machinery into four haptophyte organisms will provide insight into the membrane trafficking gene complement in a major eukaryotic clade that was previously unsampled, how the membrane trafficking system has been modified to support scale formation and secretion, and how growth under scale-forming conditions modulates membrane trafficking system gene expression.

5.2 Specific methods

The haptophyte genomic and transcriptomic sequencing, assembly, and differential expression analysis was performed by collaborators B. Read and X. Zhang at California State University San Marcos. As part of the initial *E. huxleyi* genome project headed by B. Read, the *E. huxleyi* CCMP 1516 strain was sequenced. Following the genome project, the genomes of *G. oceanica*, and *I. galbana* were sequenced. B. Read and X. Zhang also performed differential expression analyses of *E. huxleyi* (biomineralizing strain 217) and *G. oceanica* grown under four combinations of conditions: 0mM calcium, 9mM calcium, 20mM $NaHCO_3^-$ (bicarbonate spike), and no bicarbonate spike.

Concatenated gene phylogenies of the haptophytes were performed by E. Herman. Sixteen genes that share between 70-95% identity in all seven genomes were identified. A further eight genes met this criteria in all taxa except *C. tobin*. Individual gene trees were generated using Phylobayes with the GTR model of sequence evolution. Sequences that generated trees with a strongly discordant signal were discarded, as were those with too few informative positions to generate a tree with any supported clades. This generated a list of nine genes found in all taxa, as well as two genes missing only in *C. tobin*, which produced trees with clear and non-discordant phylogenetic signal. Genes were concatenated and aligned, and trees were generated as described in the Methods. The final alignment contained 7 taxa and 4829 positions. Phylobayes, MrBAYES, and RAxML were run with GTR model with gamma-distributed rate variation across sites.

Vesicle formation and fusion genes were searched for in the reference genome *E. huxleyi* CCMP1516, *G. oceanica*, *I. galbana*, and *C. tobin* using the BLAST methodology described in Methods. Phylogenetic trees were generated to classify genes in highly paralogous gene families such as SNAREs and TBC RabGAPs. Trees were also generated for the adaptins in order to classify several short adaptin-like sequences.

The analysis of the vesicle formation machinery of *E. huxleyi* CCMP1516 was performed by E. Herman, while the analysis of vesicle fusion machinery was performed by M. Klute. Identification of trafficking proteins in *G. oceanica*, and *I. galbana* were done by L. Lee, who was supervised by E. Herman. *C. tobin* gene searches were performed by E. Herman and B. Richardson. E. Herman performed all phylogenetic analysis with the exception of the adaptin trees, which were generated by L. Lee, an undergraduate student supervised by E. Herman. E.

Herman also analyzed differential expression data to identify membrane trafficking genes and pathways involved in biomineralization.

Work from this chapter has been published in Read *et al.* 2013 (supplementary information),[62] and in Lee *et al.* 2015 (on which E. Herman is the corresponding author).[552]

5.3 Evolution of the haptophyte clade and the membrane trafficking system

5.3.1 Phylogenomics of the Haptophyta

Before examining the membrane trafficking gene complement in these taxa, the question of their evolutionary relatedness was addressed. The high similarity between *E. huxleyi* and *G. oceanica* has raised the question of whether they truly represent separate genera.[553] A concatenated, multi-gene phylogeny was generated to assess the organismal relationships between the haptophytes *Chrysochromulina tobin*, *Isochrysis galbana*, *Gephyrocapsa oceanica*, and four strains of *E. huxleyi*; CCMP1516, Van556, EH2, and 92A. Phylogenies were rooted on *C. tobin*, as it is the sister taxon to the Isochrysidales (*E. huxleyi*, *G. oceanica*, and *I. galbana*) based on 18S ribosomal DNA maximum-likelihood trees.[543] Both Bayesian and Maximum-Likelihood methods generated phylogenies with the same well-supported topology: *Gephyrocapsa* grouping well within the clade of *E. huxleyi* strains, with CCMP1516 as a basal member (Figure 5.1). This suggests that *Gephyrocapsa* and *Emiliania* are not separate species. The removal of certain genes from this alignment generated trees with different internal organization in the *E. huxleyi* + *G. oceanica* clade, however, *G. oceanica* is never an "outgroup" to

Figure 5.1. Concatenated phylogeny of eleven genes to infer haptophyte evolution.

Nine genes with orthologues found in the *E. huxleyi* strains CCMP1516, EH2, Van556, and 92A, and the haptophytes *G. oceanica*, *I. galbana*, and *C. tobin* were concatenated, in addition to two genes with orthologues in all haptophytes except *C. tobin*. The haptophyte orthologues share between 70-95% sequence identity, and individual gene trees have clades with good statistical support and not strongly discordant phylogenetic signals. Node values indicating statistical support are listed as MrBAYES/Phylobayes/RAxML (posterior probability/posterior probability/bootstrap), and the tree is rooted on the *C. tobin* sequence, as it is the known outgroup to the Isochrysidales. Node values are shown on the best Bayesian topology. *G. oceanica* groups within a larger clade of *E. huxleyi* strains.

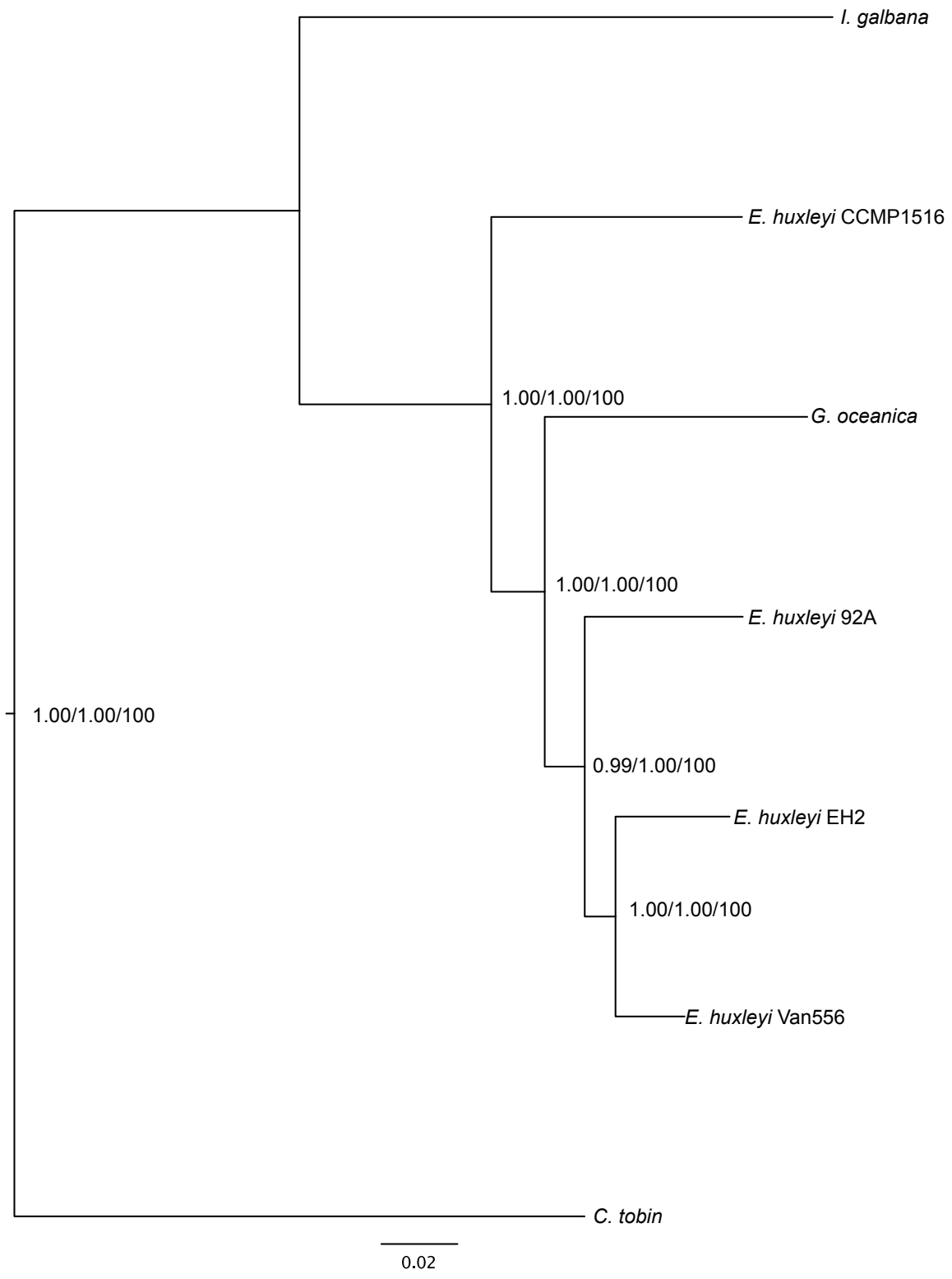

I. galbana
E. huxleyi CCMP1516
1.00/1.00/100
G. oceanica
1.00/1.00/100
E. huxleyi 92A
1.00/1.00/100
0.99/1.00/100
E. huxleyi EH2
1.00/1.00/100
E. huxleyi Van556
C. tobin
0.02

Figure 5.1

the *E. huxleyi* strains. In all cases, CCMP1516 is the first-branching strain in the clade (Supplementary Figure S5.1, S5.2). This variability in internal organization is not surprising as *E. huxleyi* has a pan genome,[62] and therefore these genes may have a different evolutionary history due to loss/transfer events between strains. However, since *E. huxleyi* CCMP1516 is consistently an outgroup to the clade containing *G. oceanica*, it is highly likely that *G. oceanica* is not distinct from *E. huxleyi.* Although it is not clear whether *E. huxleyi* and *G. oceanica* are morphotypes of the same species, it can at least be said that *E. huxleyi* and *G. oceanica* are more closely related than previously thought.

5.3.2 Comparative genomics of the membrane trafficking system of *Emiliania huxleyi*, *Gephyrocapsa oceanica*, *Isochrysis galbana*, and *Chrysochromulina tobin*

Much of the vesicle formation machinery is conserved in the haptophytes, *albeit* with some conspicuous losses. These include machinery that is highly conserved; the AP3 complex and ESCRTs I and II, discussed further below. As well, many orthologues of trafficking machinery could not be identified in *C. tobin*, however, these are likely false negative results due to genome or gene prediction quality. Therefore, non-identified genes in *C. tobin* will be treated as false negatives unless there is additional information to suggest that they are truly missing. Figure 5.2 (Online Appendix Tables 5.1-5.4) shows the presence of coat complexes and adaptor proteins in the haptophytes. The COPII coat complex is present with several duplicated subunits in the haptophytes. COPII functions in ER-Golgi anterograde transport. The Sec23, Sec24, and Sar1 subunits are duplicated; in this pre-budding complex, Sec23 and Sec24 form a bowtie structure that binds cargo (Sec24)[176,554] and the Sar1 GTPase (Sec23)[555]. This raises the possibility of COPII complexes with different combinatorial possibilities of subunits, with

Figure 5.2. Comparative genomic survey of vesicle coat, adaptor proteins, and endocytic machinery in the haptophytes.

Identification of vesicle formation machinery in *E. huxleyi* CCMP1516, *G. oceanica*, *I. galbana*, and *C. tobin.* In addition to canonical adaptor proteins, two novel alpha- and gamma-adaptor protein ear proteins were identified in the haptophytes.

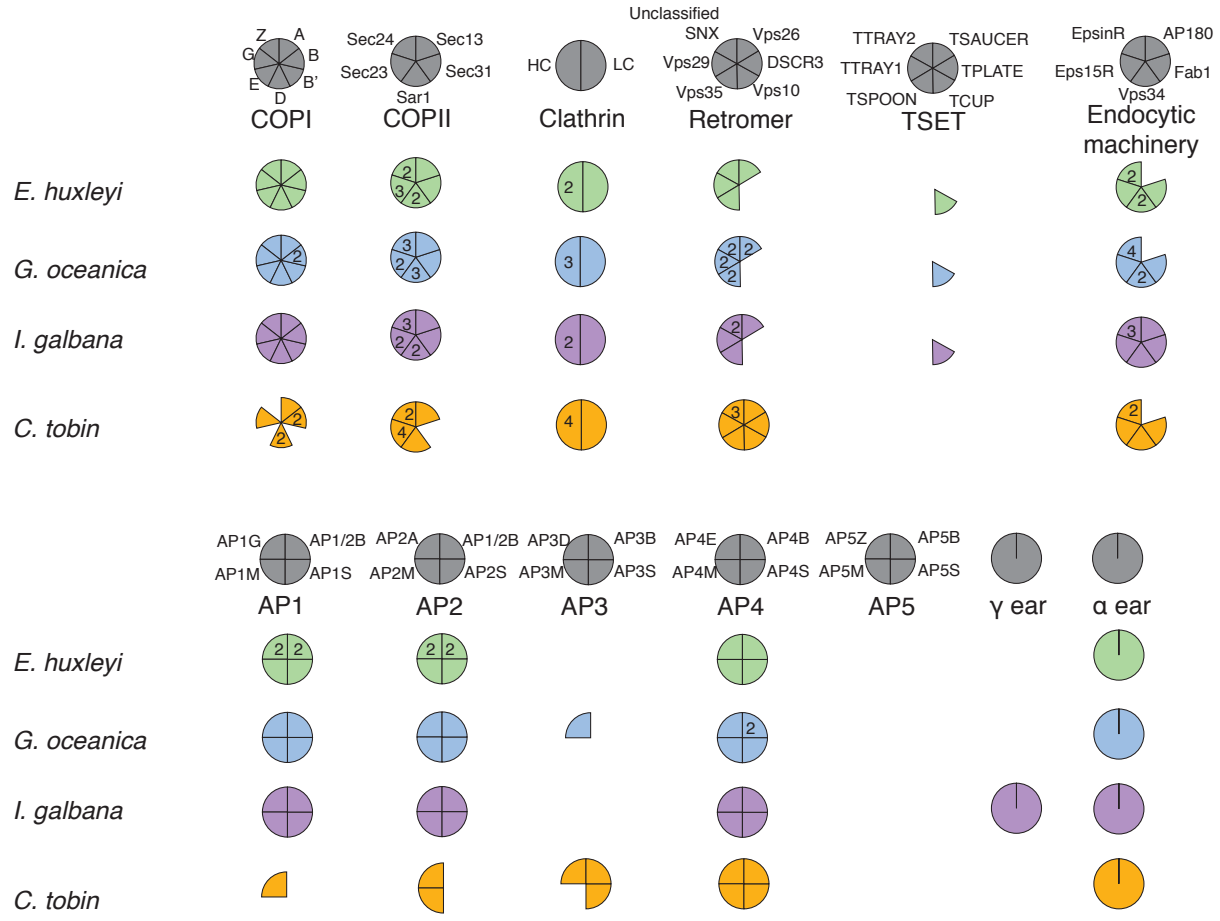

Z
A
G
B
E
B'
D
COPI
Sec24
Sec13
Sec23
Sec31
Sar1
COPII
Unclassified
HC
LC
Clathrin
SNX
Vps26
Vps29
DSCR3
Vps35
Vps10
Retromer
TTRAY2
TSAUCER
TTRAY1
TPLATE
TSPOON
TCUP
TSET
EpsinR
AP180
Eps15R
Fab1
Vps34
Endocytic machinery
E. huxleyi
G. oceanica
I. galbana
C. tobin
AP1G
AP1/2B
AP1M
AP1S
AP1
AP2A
AP1/2B
AP2M
AP2S
AP2
AP3D
AP3B
AP3M
AP3S
AP3
AP4E
AP4B
AP4M
AP4S
AP4
AP5Z
AP5B
AP5M
AP5S
AP5
γ ear
α ear
E. huxleyi
G. oceanica
I. galbana
C. tobin

Figure 5.2

different cargo preferences and vesicle budding kinetics. The reverse pathway, from Golgi to ER, occurs by COPI-mediated vesicle budding. COPI is largely complete in the haptophytes with some gene duplications in *I. galbana* and *C. tobin*.

TSET is localized at the cell plate in *Arabidopsis*,[150] and may have a similar role in plasma membrane-directed trafficking in the haptophytes. A single subunit of the TSET complex, TCUP, was identified in the Isochrysidales. TCUP is the medium subunit that is likely able to bind cargo. Its retention despite the loss of all other subunits raises the question of whether it is able to still act as a cargo binding protein and potentially interact with other evolutionarily related components (e.g. the adaptins). Opisthokonts have also retained the TCUP subunit while losing all other TSET components;[144] the tendency to keep the TCUP subunit in both the haptophytes and all of opisthokont evolution may be evidence of another role for this protein outside of the TSET complex.

In terms of endocytic function, clathrin is complete, with duplications of the heavy chain of clathrin in all four haptophytes. AP2 is present, and there are duplications in the large gamma subunit of AP1, and the shared AP1 and AP2 beta subunit in *E. huxleyi.* AP1 and clathrin function at the TGN,[556] together with EpsinR and Eps15R.[180] EpsinR has also been expanded in all taxa.

The ESCRTs, which function in multivesicular body biogenesis, are drastically reduced in the haptophytes. The cargo-binding protein Tom1esc, and ESCRTS I and II are completely absent from the haptophytes, as well as the Vps20 subunit of ESCRT III (Figure 5.3, Online Appendix Tables 5.1, 5.4, and 5.5). CHMP7 is patchily distributed in eukaryotes,[226] so its loss is not surprising. ESCRT IIIA components are not only present, but appear to have undergone lineage-specific expansions in the Isochrysidales, particularly Vps4, Vps46, and Vps31. There

Figure 5.3. Comparative genomic survey of ESCRT machinery and Arf and Rab regulators in the haptophytes.

Identification of ESCRT vesicle formation machinery, and Arf and Rab GTPase GAP and GEF proteins in *E. huxleyi* CCMP1516, *G. oceanica*, *I. galbana*, and *C. tobin.* The haptophytes do not appear to encode ESCRT complexes I and II.

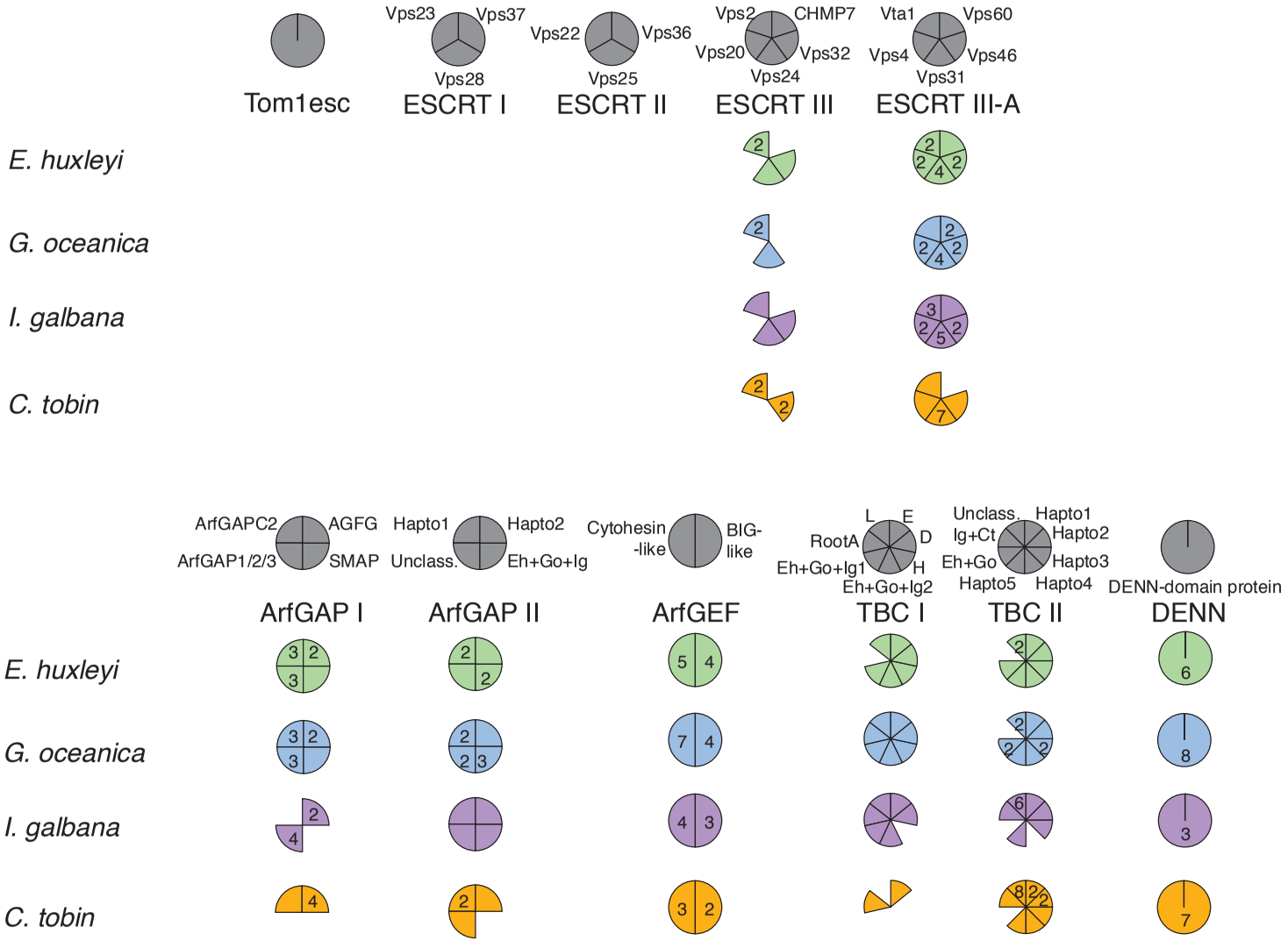

Tom1esc
Vps23
Vps37
Vps28
ESCRT I
Vps22
Vps36
Vps25
ESCRT II
Vps2
CHMP7
Vps20
Vps32
Vps24
ESCRT III
Vta1
Vps60
Vps4
Vps46
Vps31
ESCRT III-A
E. huxleyi
G. oceanica
I. galbana
C. tobin
ArfGAPC2
AGFG
ArfGAP1/2/3
SMAP
ArfGAP I
Hapto1
Hapto2
Unclass.
Eh+Go+Ig
ArfGAP II
Cytohesin
-like
BIG-
like
ArfGEF
L
E
RootA
D
Eh+Go+Ig1
H
Eh+Go+Ig2
TBC I
Unclass.
Hapto1
Ig+Ct
Hapto2
Eh+Go
Hapto3
Hapto5
Hapto4
TBC II
DENN-domain protein
DENN

Figure 5.3

are additionally seven Vps31 paralogues in *C. tobin*. Vps4 is the MVB-associated ATPase that removes ESCRT components from membranes following vesicle scission, and Vps46 regulates its interaction with the membrane.[557] The role of Vps31 is less clear, but it may potentially act as a scaffold; it interacts with both the ubiquitinated cargo-binding Vps23 protein of ESCRT I, as well as the ESCRT III-associated factor AMSH,[558,559] which is a deubiquitinating enzyme. The loss of ESCRT complexes I and II, and patchy loss of ESCRT III, suggest that the ESCRT-mediated MVB biogenesis pathway is undergoing degradation. However, this raises the question of the ESCRT III-A subunit duplications. In addition to a role in endocytic sorting and degradation, Vps23/Tsg101, ESCRT III, and ESCRT III-A machinery is involved in cytokinesis in mammalian cells[219,221] and likely also in plant cells.[223,560] Congruent with MVB pathway degradation is the near-complete loss of AP3, which appears to have been progressively lost in this lineage. AP3 transports cargo from tubular endosomes to late endosomes, lysosomes, and lysosome-related organelles. The loss of AP3 has been seen together with ESCRT degradation in other taxa such as the Apicomplexa.[50]

The retromer complex, which functions in receptor recycling from endosomes to the TGN, is complete in *C. tobin*, including the Vps26 paralogue DSCR3 and the ancestral cargo protein Vps10. However, while the subunits of the coat complex are retained in the Isochrysidales, DSCR3 and Vps10 could not be identified. Both are considered to have a patchy distribution in eukaryotes; this is another example of their lineage-specific losses.[196] All retromer coat subunits have been duplicated in *G. oceanica*, although the functional significance of this is unclear.

In searching for adaptin proteins, several short adaptin-like sequences were identified that resembled truncated “ear” appendages of alpha and gamma large subunits. Often, these short

adaptin-like regions were part of much longer genes, however, this appeared to be the result of exon fusion with another gene of unrelated function. No additional adaptin sequence was found upstream or downstream of the putative ear region, suggesting that they are not part of a typical large subunit. The sequences were then classified using phylogenetics, and were shown to belong to clades containing alpha or gamma adaptin subunits (Figure 5.4, Online Appendix Table 5.6). Gamma ear proteins were found in all haptophytes, while an alpha ear protein was found only in *I. galbana*. qPCR showed that these were not sequencing artifacts, and are expressed in *E. huxleyi*, *I. galbana*, and *G. oceanica*.[552] These appear to be analogous to the Golgi-localized, γ ear-containing, ADP-ribosylation factor-binding proteins (GGAs) in animals. The GGAs are Arf-dependent clathrin adaptors that have been shown to function with clathrin and AP1 in TGN-to-endosome anterograde trafficking.[561–564] However, the GGAs have additional VHS, GAT, and clathrin interaction domains to support interaction with cargo and trafficking factors, while there are no predicted domains other that the gamma/alpha adaptin ear domains in the haptophyte ear proteins, leaving the question of their function open.

In terms of vesicle fusion machinery, there are both patterns of duplications and losses, particularly in the multi-subunit tethering complexes (Figure 5.5, Supplementary Figures S5.3-S5.6, Online Appendix Tables 5.1, 5.4, 5.7-5.10). Syntaxin 5, which functions in ER-to-Golgi transport,[565] is duplicated in all Isochrysidales. As coccoliths are found in Golgi-associated vesicles, this duplication may have functional consequence in directing coccolith versus 'normal' vesicular traffic. Syntaxin 5 can also function in retrograde Golgi trafficking in yeast,[566] and strikingly, retrograde trafficking to the TGN in complex with Syntaxin 16.[567] Multiple paralogues of Syntaxin 5 may therefore have roles in different pathways. The retrograde Golgi-

Figure 5.4. Phylogenetic classification of the large adaptor protein subunit ear domains and ear homology domain proteins.

Gamma and alpha adaptins and related ear proteins from *E. huxleyi* strains CCMP1516, 92A, Van556, and EH2, as well as *G. oceanica* and *I. galbana* were included in the phylogeny, with previously classified *H. sapiens*, *A. thaliana*, and *Yarrowia lipolytica* AP1G and AP2A sequences. Ear domain proteins are indicated by an asterisk (*). Node values indicating statistical support are listed as MrBAYES/PhyML/RAxML (posterior probability/bootstrap/bootstrap) as shown in the inset on the best Bayesian topology. A dash indicates that the clade was not reconstructed.

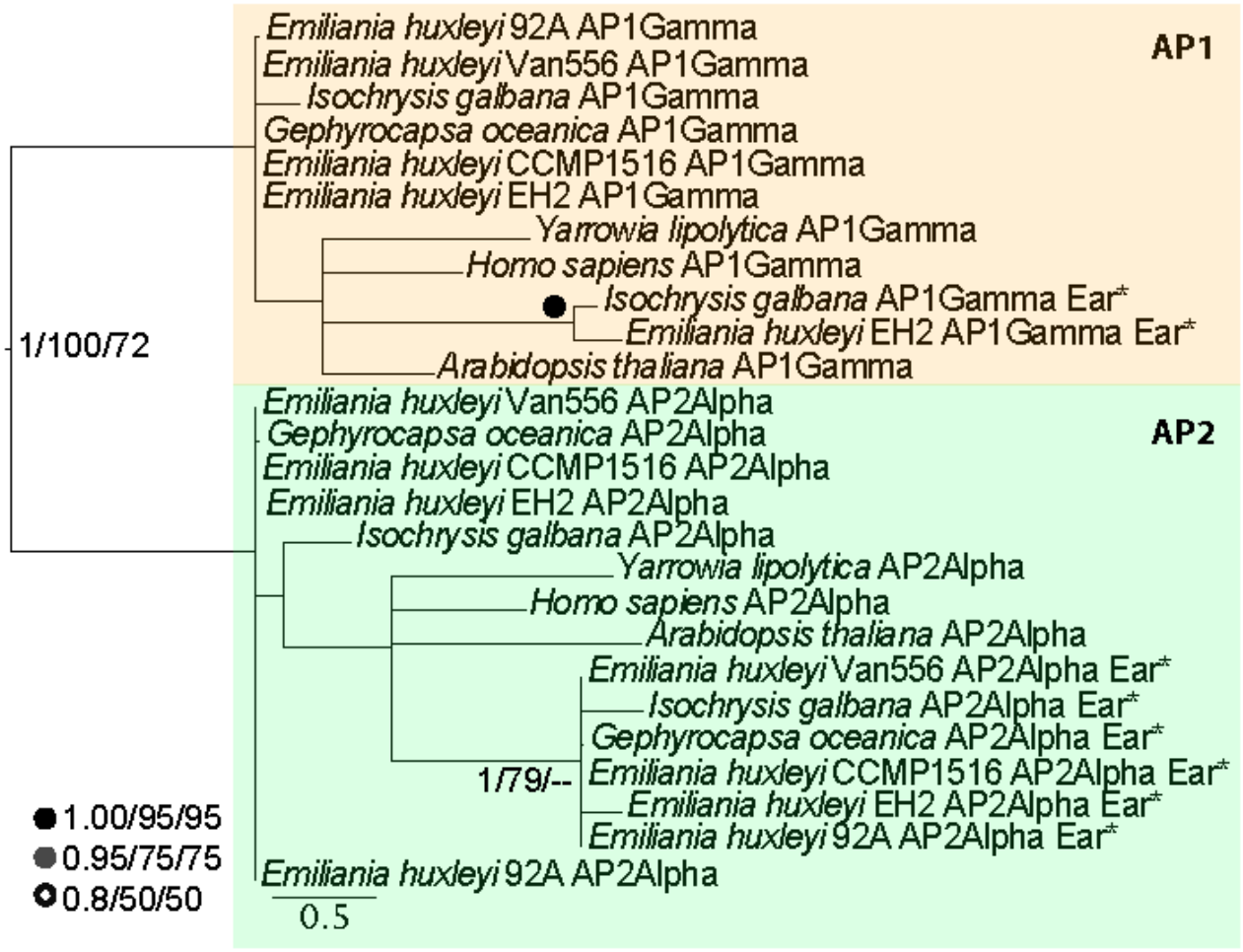

Emiliania huxleyi 92A AP1Gamma
Emiliania huxleyi Van556 AP1Gamma
Isochrysis galbana AP1Gamma
Gephyrocapsa oceanica AP1Gamma
Emiliania huxleyi CCMP1516 AP1Gamma
Emiliania huxleyi EH2 AP1Gamma
Yarrowia lipolytica AP1Gamma
Homo sapiens AP1Gamma
Isochrysis galbana AP1Gamma Ear*
Emiliania huxleyi EH2 AP1Gamma Ear*
Arabidopsis thaliana AP1Gamma
AP1
1/100/72
Emiliania huxleyi Van556 AP2Alpha
Gephyrocapsa oceanica AP2Alpha
Emiliania huxleyi CCMP1516 AP2Alpha
Emiliania huxleyi EH2 AP2Alpha
Isochrysis galbana AP2Alpha
Yarrowia lipolytica AP2Alpha
Homo sapiens AP2Alpha
Arabidopsis thaliana AP2Alpha
Emiliania huxleyi Van556 AP2Alpha Ear*
Isochrysis galbana AP2Alpha Ear*
Gephyrocapsa oceanica AP2Alpha Ear*
1/79/--
Emiliania huxleyi CCMP1516 AP2Alpha Ear*
Emiliania huxleyi EH2 AP2Alpha Ear*
Emiliania huxleyi 92A AP2Alpha Ear*
AP2
1.00/95/95
0.95/75/75
0.8/50/50
Emiliania huxleyi 92A AP2Alpha
0.5

Figure 5.4

Figure 5.5. Comparative genomic survey of vesicle fusion machinery in the haptophytes.

Identification of SNAREs, SM proteins, and multisubunit tethering complexes in *E. huxleyi* CCMP1516, *G. oceanica*, *I. galbana*, and *C. tobin.* The haptophytes do not appear to encode Exocyst, TRAPPII, or CORVET-specific subunits.

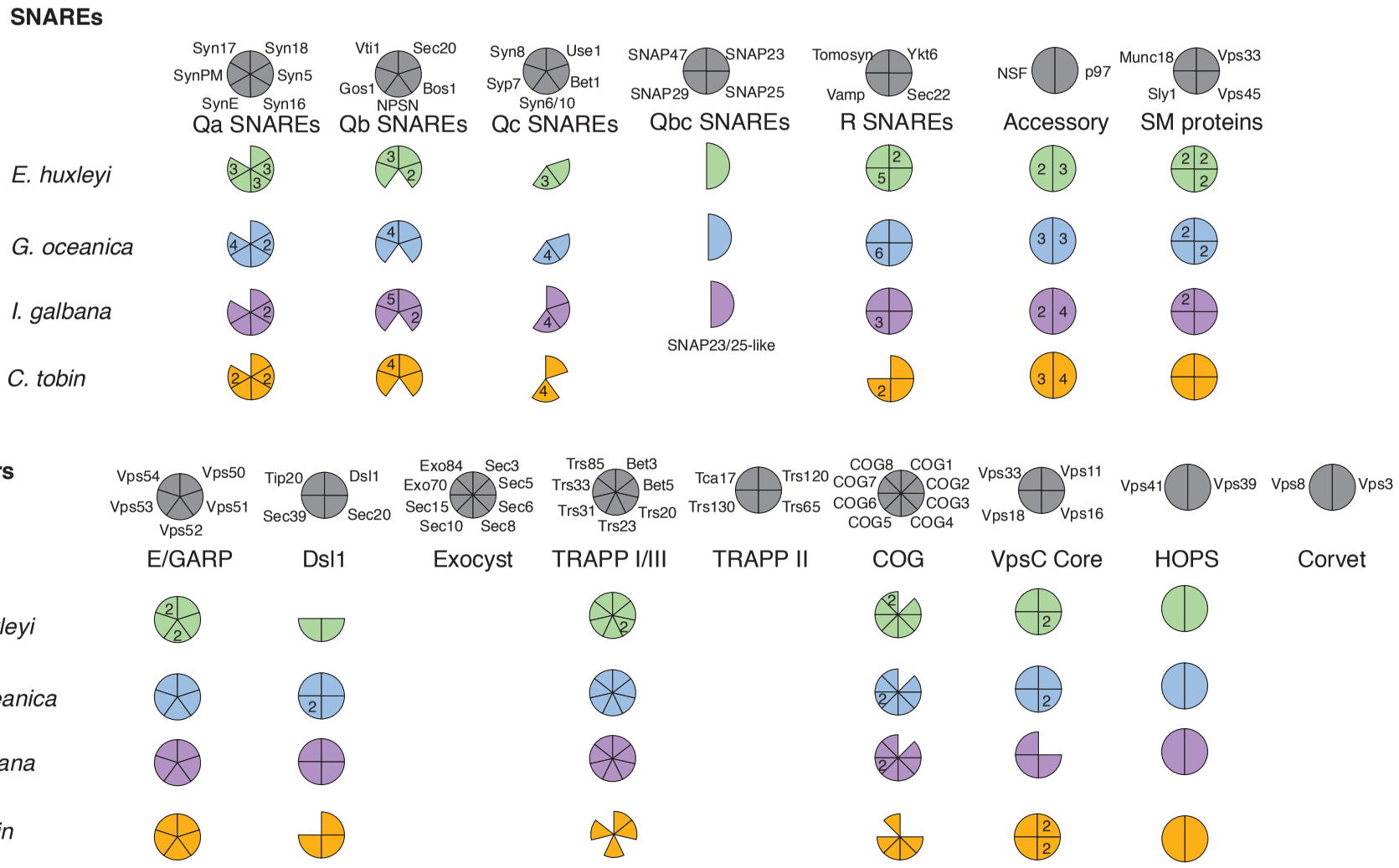

SNAREs
Syn17
Syn18
SynPM
Syn5
SynE
Syn16
Qa SNAREs
Vti1
Sec20
Gos1
Bos1
NPSN
Qb SNAREs
Syn8
Use1
Syp7
Bet1
Syn6/10
Qc SNAREs
SNAP47
SNAP23
SNAP29
SNAP25
Qbc SNAREs
Tomosyn
Ykt6
Vamp
Sec22
R SNAREs
NSF
p97
Accessory
Munc18
Vps33
Sly1
Vps45
SM proteins
E. huxleyi
G. oceanica
I. galbana
C. tobin
SNAP23/25-like
Tethers
Vps54
Vps50
Vps53
Vps51
Vps52
E/GARP
Tip20
Dsl1
Sec39
Sec20
Dsl1
Exo84
Sec3
Exo70
Sec5
Sec15
Sec6
Sec10
Sec8
Exocyst
Trs85
Bet3
Trs33
Bet5
Trs31
Trs20
Trs23
TRAPP I/III
Tca17
Trs120
Trs130
Trs65
TRAPP II
COG8
COG1
COG7
COG2
COG6
COG3
COG5
COG4
COG
Vps33
Vps11
Vps18
Vps16
VpsC Core
Vps41
Vps39
HOPS
Vps8
Vps3
Corvet
E. huxleyi
G. oceanica
I. galbana
C. tobin

Figure 5.5

to-ER MTC Dsl1 is incomplete in *E. huxleyi*, but not the other haptophytes. The absence of Dsl1 components has previously been shown here and in others' works to be associated with divergent or lost peroxisomes.[480] Peroxisomal function has yet to be studied in *E. huxleyi*, but this result suggests that it may have modified peroxisomes. The haptophytes have relatively complete ER-Golgi and intra-Golgi TRAPPI and COG tethering complexes. There is evidence for multiple exocytic pathways in the Isochrysidales, which have a SNAP23/SNAP25-like Qbc SNARE; in mammalian cells, SNAP-23 and SNAP-25 are involved in regulated exocytosis of synaptic vessels and secretory granules.[255] The haptophytes also have between 3 to 6 paralogues of a VAMP-like protein, which can function in both endocytic and exocytic pathways.[568] The Isochrysidales have two copies of Sec1/Munc18, the exocytic SM protein. Together, these suggest a diversified exocytic system. Strikingly, however, all four haptophytes have lost the entire Exocyst MTC, which tethers exocytic vesicles for fusion at the plasma membrane.[295,569] Given the haptophytes routinely exocytose large scales as plasma membrane-directed vesicular cargo, and there is retention and expansion of other exocytic machinery, the loss of Exocyst MTC apparently does not impair scale extrusion or 'normal' vesicle fusion at the plasma membrane.

As mentioned above, machinery involved in endocytosis (e.g. clathrin, AP-2) and TGN-endosome trafficking (EpsinR, AP-1, AP-4) is present and in some cases expanded in the haptophytes. This pattern is echoed in the corresponding vesicle fusion machinery. SNAREs Syntaxin 6 and Vti1, and the interacting SM protein Vps45, are duplicated in one or more of the haptophytes, and they are responsible for endosome-to-TGN transport.[300] The HOPS and CORVET MTCs share a core set of VpsC components, which are present in the haptophytes, with duplications of the Vps16 subunit. However, while HOPS-specific subunits are present, the

CORVET-specific Vps3 and Vps8 subunits appear to be lost. CORVET functions in early endosome fusion, and may be related to the loss of ESCRTs and AP-3. On the other hand, HOPS is a lysosomal tether that is involved in LE-lysosome and autophagosome-lysosome fusion.[311] Curiously, this pattern is reversed in the Apicomplexa, where the absence of ESCRTs I and II, and AP3 are associated with the loss of the HOPS complex, and occasional loss of the CORVET Vps8 subunit.[50,480]

The Arf and Rab GTPase regulators have undergone some lineage-specific expansion in the haptophytes (Figure 5.3, Online Appendix Table 5.1, 5.4, 5.11). Again, Arf and Rab proteins were within the purview of another lab working on these haptophyte genomes. ArfGAP sequences are classified in Supplementary Figure S5.7 (Online Appendix Table 5.10). All taxa have multiple copies of the AGFG protein, which functions in clathrin-mediated endocytosis,[205] and ArfGAP1 or ArfGAP2/3, which function at the Golgi. ArfGAPC2 is also present in the haptophytes with the exception of *Isochrysis*, and has been duplicated in *E. huxleyi* and *G. oceanica*. ArfGAP C2 is a recently discovered to be an ancient, patchily distributed ArfGAP protein lost in animals and fungi.[199] Although its function is unknown, it contains a lipid-binding C2 domain in addition to the ArfGAP domain, suggesting it may be able to interact with membranes. Additionally, there are multiple paralogues of ArfGAPs in haptophyte-specific and Isochrysidales-specific clades, particularly in *E. huxleyi* and *G. oceanica*. In all haptophytes, there are multiple cytohesin-like ArfGEFs and BIG-like ArfGEFs; in mammals cytohesins function at the plasma membrane, while BIG/GBF ArfGEFS are associated with the Golgi (Supplementary Figure S5.8, Online Appendix Table 5.10).[160]

Five ancient TBC subfamilies were identified in at least one member of the haptophytes, as well as five additional haptophyte-specific clades, two Isochrysidales-specific clades, an *E.*

huxleyi+*G. oceanica*-specific clade, and an *I. galbana*+*C. tobin*-specific clade, where presumably the calcifying haptophytes lost this gene (Supplementary Figure S5.9, Online Appendix Table 5.10). These lineage-specific duplications in the GTPase regulator families suggest that haptophytes have a complex network of membrane trafficking regulation that may be important for scale formation and possibly biomineralization.

5.4 Transcriptomic analysis of membrane trafficking gene expression during coccolith formation

While there are many trafficking gene duplication events that are common to the four haptophytes, these data only allow for speculation about the role of specific genes in relation to scale formation. While all four organisms produce organic scales, only *E. huxleyi* and *G. oceanica* extrude calcium carbonate scales. In order to understand the role of the membrane trafficking system in this process, RNA-Seq was performed on cultures of *E. huxleyi* and *G. oceanica* grown under scale-forming and non-scale-forming conditions. Membrane trafficking genes that are differentially expressed may therefore play a role in scale formation.

E. huxleyi, and *G. oceanica* cultures were grown with 0 vs. 9mM calcium, and with or without a bicarbonate spike. The effect of calcium or bicarbonate alone may induce gene expression changes, however, those that are significantly differentially expressed when grown in both calcium + bicarbonate spike are most likely involved in biomineralization. The addition of calcium and bicarbonate spike likely has pleiotropic effects, some of which may be unrelated to scale production (e.g. membrane trafficking associated with lipid storage upon the addition of bicarbonate). Although it may not be possible to tease apart gene expression differences due to scale production versus other cellular processes using these transcriptomic data alone, they nonetheless provide insight into the cell biology of *E. huxleyi* and *G. oceanica*.

Overall, the addition of calcium has little effect on gene expression, regardless of the bicarbonate spike. The effect of the bicarbonate spike on membrane trafficking gene expression in the presence or absence of calcium is shown in Table 5.1 (Online Appendix Table 5.12). In *E. huxleyi*, the bicarbonate spike alone greatly affects membrane trafficking gene expression, while in *G. oceanica*, the bicarbonate spike also induces gene expression changes, but only in the presence of calcium. Because of the minimal effect of calcium on gene expression, for both organisms, consideration is given to genes that are differentially expressed with the addition of bicarbonate spike regardless of the presence of calcium.

The early secretory system appears to be up-regulated in both *E. huxleyi* and *G. oceanica* (Table 5.1). In *E. huxleyi*, the up-regulated genes in this system include COPII coat subunits (two Sec23 genes and the GTPase Sar1), as well as COG4 of the intra-Golgi MTC COG, and the Golgi-associated Arf1. In *G. oceanica*, early secretory proteins Sec23 and Rab1 (ER-to-Golgi vesicle fusion) are up-regulated by spike and calcium, as are the COPI subunits COPB and COPG. Conversely, in *E. huxleyi*, the Golgi-to-ER retrograde trafficking SNARE Syntaxin 18 is down-regulated. There appears to be a clear signal of up-regulated ER-Golgi anterograde trafficking genes in both coccolith-forming organisms.

In both *E. huxleyi* and *G. oceanica*, there is some evidence for regulation of exocytosis as a function of biomineralization. The phospholipid transfer protein Sec14 is up-regulated in the bicarbonate spike+calcium condition in both organisms. In yeast, Sec14 regulates specific TGN-to-plasma membrane export pathways.[570] One plasma membrane syntaxin (SyntaxinPM) is down-regulated under these conditions in *E. huxleyi*, although there are two additional syntaxin PM paralogues in *E. huxleyi* that are not differentially expressed. This raises the possibility of complex regulation of trafficking factors, rather than a blanket increase

Table 5.1. Membrane trafficking pathways represented in differential expression data from *E. huxleyi* and *G. oceanica* when grown with the addition of a bicarbonate spike, regardless of the presence of calcium

Effect of bicarbonate spike		Organism	
		Emiliania huxleyi	*Gephyrocapsa oceanica*
Pathway	early secretory (ER anterograde)	UP: Sec23, Sar1	UP: Rab1, Sec23
	Golgi-ER retrograde	DOWN: Syn18	none
	intra-Golgi transport	UP: COG4, Arf1	UP: 2xBIG-like, COPB, COPG
	post-TGN secretion	UP: Sec14, DOWN: SynPM	UP:Sec14
	Endocytosis	UP: synaptojanin, drebrin, clathrin light chain, Eps15R	UP: AGFG
	endosomal recycling	UP: Vps13, Vps53, Vps50	UP: EpsinR, Syn6/10, Rab9
	endolysosome maturation	DOWN: Vps4, Vps41	DOWN: Vps16, Vps39, Vps34

or decrease in pathway function.

Several endocytic factors are up-regulated in *E. huxleyi* under scale-forming conditions, including clathrin light chain, two paralogues of Eps15R (clathrin-mediated endocytosis),[571] drebrin (adaptor for F-actin and dynamin),[572] and synaptojanin (involved in fluid phase endocytosis).[573] Only one endocytic gene is differentially expressed in *G. oceanica*; the ArfGAP protein AGFG.[205]

In both *E. huxleyi* and *G. oceanica*, endocytic recycling proteins are up-regulated. In *E. huxleyi*, these are Vps50 and Vps53, part of EARP complex at recycling endosomes,[303] and Vps13, which functions in endosome-to-TGN protein sorting. In *G. oceanica*, syntaxin 6/10, EpsinR, and Rab9[574] are up-regulated in the spike+calcium condition compared to calcium alone.

Late endosome/lysosome trafficking is conversely down-regulated under these conditions in *E. huxleyi* (the HOPS subunit Vps41 and the ESCRT protein Vps4). In *G. oceanica*, HOPS tethering complex members are also down-regulated (two Vps16 paralogues and Vps39), as is the PI(3)P kinase Vps34, which functions at late endosomes and lysosomes. While the transcriptomic evidence is limited, it hints that the cell may not be taking up exogenous material for degradation, but instead bringing membrane back from the plasma membrane (while recycling resident PM proteins) to maintain an endomembrane balance within the cell.

Regulation of vesicle formation and fusion dynamics by Arf and Rab GAP and GEF proteins may be important in scale trafficking kinetics. It appears that these GTPase regulators are generally down-regulated in both *E. huxleyi* and *G. oceanica* under biomineralizing conditions. There is also a case of potentially orthologous proteins being inversely regulated:

there is an ArfGAP protein in *E. huxleyi* and *G. oceanica* from the same Isochrysidales-specific clade that is up-regulated in the former and down-regulated in the latter.

Additional membrane trafficking factors that function throughout the cell are differentially expressed under biomineralizing conditions. Those that are up-regulated in one or both calcifying haptophytes are p97, and dynein and kinesin proteins (retrograde- and anterograde-directed motor proteins, respectively, for vesicle transport along microtubules).

While there are several clear patterns, such as up-regulation of early secretory machinery and recycling machinery, there are cases where a differentially expressed gene is one of several paralogues. The clearest example is transcriptional changes of GTPase regulators, in both gene family expansions and expression under calcifying conditions. This suggests that, like the process of encystation studied in Chapter 4, biomineralization involves complex regulation of membrane trafficking, which cannot be well-understood without functional characterization of trafficking genes.

5.5 Discussion

In this Chapter, the gene complement of the membrane trafficking system in *E. huxleyi*, a major taxonomic sampling point as the first sequenced haptophyte, was revealed. Further comparison with the genomes of related haptophytes helped identify several novel proteins and changes to the trafficking system in these algae. Multiple concatenated gene phylogenies suggest that *E. huxleyi* and *G. oceanica* are not separate species. Recently, Bendif *et al.* 2014 have shown using nuclear 18S and 28S rDNA, plastidial 16S rDNA, and plastidial *tufA* phylogenies that *E. huxleyi* and *Gephyrocapsa muellerae* are sister taxa, with *G. oceanica* as an outgroup.[553] The

authors suggest that this is the result of introgressive hybridization within this clade, with *E. huxleyi* arising from an unspecified *Gephyrocapsa* lineage. The data presented here are congruent with the idea that *E. huxleyi* and *G. oceanica* represent a single species complex, although genome data from more *Gephyrocapsa* species will be necessary to make any specific claims about the relationships within this clade.

Several major losses of complexes have occurred in the haptophytes. These include the loss of ESCRT complexes I and II (and the ESCRT 0 analogue Tom1esc), the AP3 complex, the Exocyst MTC, and the CORVET MTC. The TRAPPII subunits of the Golgi-plasma membrane MTC complex also could not be identified, however, the loss of TRAPPII is not uncommon in eukaryotes.[254] Loss of whole ESCRT complexes has been observed in several eukaryotic taxa, including some members of the Apicomplexa,[50,480] *Giardia intestinalis*,[226] and as described in this work, in *Blastocystis* sp. and *Proteromonas lacertae*. However, the uniting feature of these organisms is parasitism (or symbiosis), whereas the haptophytes are undoubtedly free-living. The purpose of the ESCRT machinery within the membrane trafficking system is the degradation of transmembrane receptors and other cargo. One potential explanation for its loss in the haptophytes is that the coccoliths or body scales minimize the turnover of underlying plasma membrane-bound receptors, so a system to regulate them via internalization and degradation is not necessary. However, plants have analogous cell walls, and yet have relatively complete ESCRT complements.[226] While it is unlikely that the haptophytes can generate functional MVBs, the ESCRT III-A subunits that are present are often duplicated and may function in abscission during cytokinesis.[221,223] Further work is required to understand the function of ESCRT components in the haptophytes.

In addition to the ESCRTs, AP3, Exocyst, and CORVET are lost, a pattern that is also observed in members of the Apicomplexa. It may seem that a free-living algae and a human intracellular parasite have vastly different lifestyles, yet they both have unique secretory requirements; haptophytes secrete large scales, while apicomplexan organisms secrete contents of specialized secretory organelles to modify the parasitophorous vacuole that they inhabit.[575] Perhaps, in both lineages, the Exocyst can be lost because targeted secretion is no longer necessary. On the other hand, the haptophytes have also expanded their secretory SNARE subfamilies and the SM protein Sec1/Munc18, which suggests that these proteins are involved in regulation of secretion, even with the loss of the Exocyst MTC. This suggests that vesicle fusion at the plasma membrane via SNARE function occurs, although the upstream tethering step is either unnecessary in the haptophytes or occurs via a novel, unknown complex. In addition to the loss of AP3 and CORVET, subunits of the HOPS complex are down-regulated during scale-forming conditions in all Isochrysidales. The loss of AP3 and CORVET functionality, and down-regulation of HOPS under biomineralizing conditions, may generally indicate less emphasis on lysosomal degradation.

The early secretory, late secretory, and endocytic recycling pathways have undergone expansion in the haptophytes and Isochrysidales, and in some cases, their constituents members are differentially expressed under scale-forming conditions. In the early secretory pathway (ER-to-Golgi), there is a duplication of syntaxin 5 and duplications in the COPII coat complex in all four haptophytes. Members of the COPII complex are up-regulated in the coccolith-forming taxa under scale-forming conditions. In *E. huxleyi* and/or *G. oceanica*, Rab1, Sar1, Arf, BIG-like ArfGEFs, and members of the COPI complex, which function in intra-Golgi transport are up-regulated under biomineralizing conditions. One potential explanation for the expansion and up-

regulation of this system is that early secretory trafficking machinery plays a specific role in trafficking components to the growing coccolith vesicle. However, another explanation is that of increased lipid production at the ER. One of the unique features of the haptophytes is their ability to produce polyunsaturated long-chain alkenone lipids, which are located in the coccolith-producing compartment,[576] as well as lipid bodies.[577] Several of the early secretory genes that are duplicated or up-regulated in these taxa are also involved in lipid droplet storage, such as COPI components and Arf1.[151] Given the duplications in the ER anterograde trafficking machinery, and the up-regulation of both ER anterograde and intra-Golgi trafficking, both processes are likely relevant for coccolith production. As the coccolith vesicle is thought to be derived from the Golgi, one potential explanation is that alternate COPII subunits traffic coccolith-specific cargo to this organelle, and the COPI and Arf1 proteins may be involved in the transfer of alkenone lipids between the lipid body and the coccolith vesicle.

As mentioned above, despite the absence of the exocyst complex, other secretory protein families are present or expanded in the haptophytes, such as VAMP-like proteins and the SM protein Sec1/Munc18. Interestingly, a Qbc SNAP23/25-like SNARE is found in all four taxa. In mammalian cells, Qbc SNAREs are associated with specialized, regulated secretion, particularly in neurons.[578] However, a similar protein was identified in the ciliate *Paramecium*, and rather than being involved in exocytosis, RNAi silencing of the SNAP protein increased the number of food vacuoles due to increased uptake.[290] It is therefore possible that the haptophytes use this Qbc SNARE in a similar manner.

The TGN-endosome recycling pathway has many duplicated components in the haptophytes, as well as up-regulated members under scale-forming conditions. Duplicated subfamilies include proteins in nearly every aspect of vesicle trafficking, including AP1,

retromer, EARP, EpsinR, the SM protein Vps45, and SNAREs Syntaxin 6 and Vti1. Furthermore, transcriptomics analysis showed that GARP subunits, some EpsinR paralogues, Syntaxin 6, Rab9, and Vps15 are up-regulated in *E. huxleyi* and *G. oceanica* under scale-forming conditions. One potential explanation for the role of TGN-endosome recycling in scale production is the recycling of membrane from scale-containing vesicles that fuse with the plasma membrane. Unless the scale-containing vesicle interacts with the membrane in a 'kiss-and-run' scenario, which seems unlikely given the size of the cargo, membrane must be taken back into the cell to preserve homeostasis (Figure 5.6). Presumably, endocytosed membrane will include cell surface proteins, which then must be recycled back to the plasma membrane, lest they be degraded. It is notable, then, that much of the endolysosomal degradation machinery is missing (e.g. ESCRT, AP3), and when present, is down-regulated under biomineralizing conditions (e.g. HOPS).

Another possible function of this machinery is generation of the reticular body during calcification. The reticular body forms on the distal face of the coccolith vesicle, and is a tubular organelle distinct from the Golgi.[348] While its cellular origin is unclear, it may be derived from TGN or endosomal membrane, and thus the expansions in TGN-endosome trafficking machinery may represent a trafficking pathway to this organelle. The reticular body has only been observed in calcifying taxa, although TGN-endosome trafficking gene duplications are present in all four haptophytes. Because only *E. huxleyi* and *G. oceanica* were studied with regards to scale formation, it is not clear whether the differential expression of these genes is related to biomineralization or scale formation in general. Functional work is therefore necessary to understand the role of this TGN-endosomal machinery in scale formation in the haptophytes.

Figure 5.6. Diagram of membrane trafficking system gene expression regulation during biomineralization in *E. huxleyi* and *G. oceanica*.

Pathways with genes that are co-regulated in *E. huxleyi* and *G. oceanica* are shown with the specific DE genes listed. The cells on the left show membrane trafficking processes associated with scale secretion, while those on the right show compensatory processes to retrieve excess membrane following the fusion of coccolith vesicles with the plasma membrane.

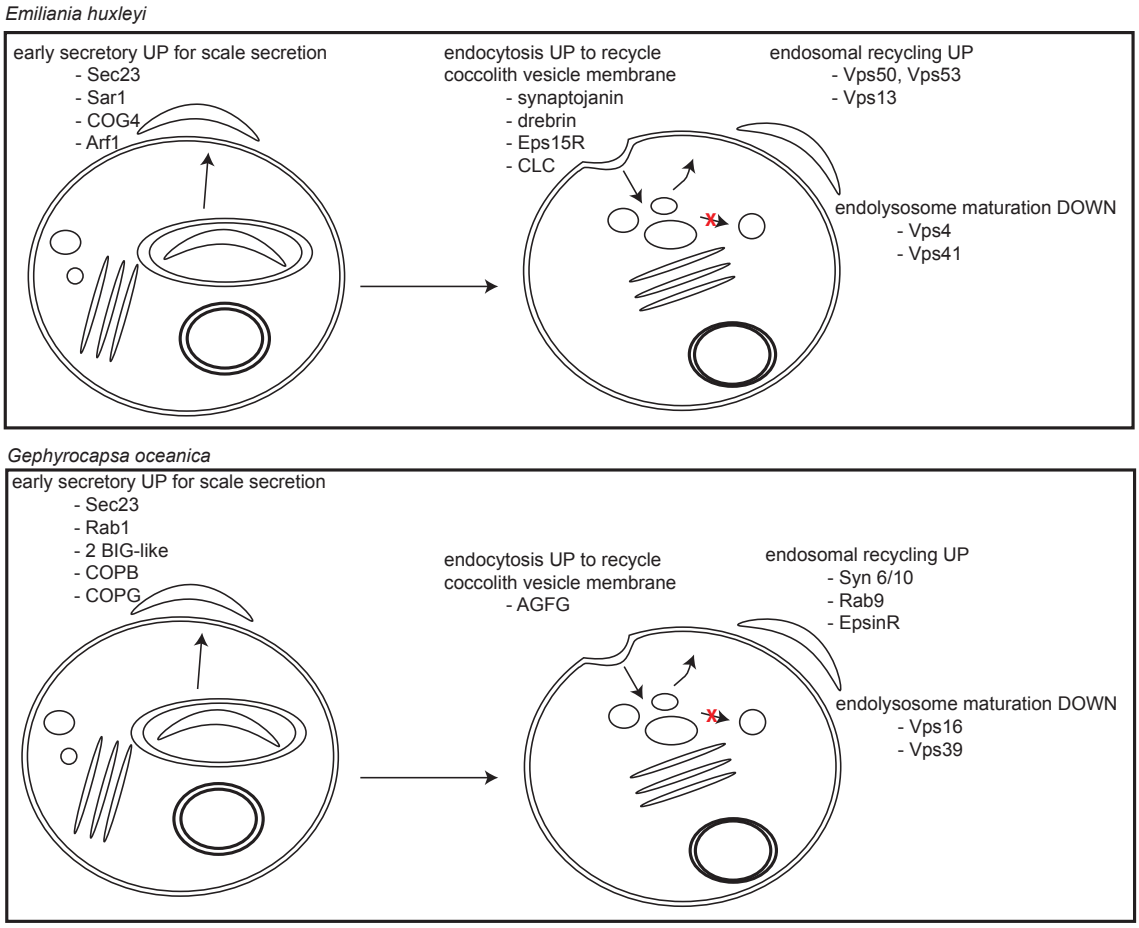

Emiliania huxleyi
early secretory UP for scale secretion
- Sec23
- Sar1
- COG4
- Arf1
endocytosis UP to recycle
coccolith vesicle membrane
- synaptojanin
- drebrin
- Eps15R
- CLC
endosomal recycling UP
- Vps50, Vps53
- Vps13
endolysosome maturation DOWN
- Vps4
- Vps41
Gephyrocapsa oceanica
early secretory UP for scale secretion
- Sec23
- Rab1
- 2 BIG-like
- COPB
- COPG
endocytosis UP to recycle
coccolith vesicle membrane
- AGFG
endosomal recycling UP
- Syn 6/10
- Rab9
- EpsinR
endolysosome maturation DOWN
- Vps16
- Vps39

Figure 5.6

The Rab and Arf GAP and GEF regulators have undergone lineage-specific expansions related to scale production. Unfortunately, little can be said of their actual purpose without functional analysis in the haptophytes. They do raise questions, however, of the importance of gene expression regulation during biomineralization. In several cases, multiple paralogues of the same gene had different expression patterns during scale formation, which suggests that membrane trafficking regulation during coccolith formation is more complex than 'more' or 'less' trafficking pathway function. This echoes what was observed in Chapter 4; rather than simple presence or absence of genes, it is sculpting of the membrane trafficking system complement and finely tuned gene regulation that is the choreography of a biological process.

Overall, the effect of calcium and bicarbonate on the calcifying haptophytes is up-regulation of genes in the secretory, endocytic, and recycling systems. However, the bicarbonate spike had a much greater effect than calcium on gene regulation. This may be because calcium is involved in many cellular functions, such as intracellular signaling, cytoskeletal function, and as an enzyme cofactor. The intracellular concentration is likely to be tightly regulated, potentially by calcium efflux or sequestration, which is consistent with its non-effect on transcription. On the other hand, bicarbonate may regulate membrane trafficking gene expression as a function of scale formation, and/or lipid production. There is no genetic system to do functional analysis in haptophyte algae; however, characterizing the genes and pathways suggested here to be relevant represents the next step in understanding how the membrane trafficking systems underpins scale formation in the haptophytes.

6. Genome and pathogenicity-associated transcriptome of the neuropathogenic amoeba *Naegleria fowleri*

6.1 Introduction

The first Results chapter focused on membrane trafficking gene presence and absence in understanding the evolution of a lineage. In the next two chapters, this approach was extended to assess membrane trafficking system function in relation to specific cellular behaviours. In this final chapter, the scope further expands. Comparative genomics and transcriptomics were used to investigate pathogenesis in the neuropathogenic amoeba *Naegleria fowleri*, but rather than concentrating specifically on membrane trafficking, multiple cellular systems are explored. In this way, we can understand pathogenesis in the context of the whole organism, and the large-scale genomic evolution that separates the deadly *N. fowleri* from its harmless relative, *N. gruberi*.

Naegleria fowleri is an opportunistic pathogen of humans and animals, and is a member of the supergroup *Excavata*. *N. fowleri* causes primary amoebic meningoencephalitis, killing more than 95% of those infected, usually within two weeks.[579] Symptoms include changes in the senses of taste or smell, nausea, fever, nosebleeds, severe headache, finally leading to coma and death.[580] Infection occurs when contaminated water enters the nose (e.g. when swimming),[581] and *N. fowleri* passes through the cribriform plate to the olfactory bulb in the brain.[582–584] In addition to swimming, sinus irrigation and ritual ablution using contaminated water are other means of infection.[585]

The worldwide reported cases number in the dozens each year, but infection is likely to be more prevalent particularly in developing countries with warm climates.[586] As infection is relatively rare and resembles other types of meningitis, the first cases of *N. fowleri* infection were described only ~50 years ago, in Texas, Florida, and Virginia (US), South Australia, and the former Czechoslovakia.[587–591] However, since then, there have been over 300 cases

worldwide, with this number growing annually.[592] It seems likely that we are still not aware of the global scale of *N. fowleri* infection.

N. fowleri prefers to live at 32°C, but it is thermotolerant, growing at temperatures up to 45°C (and up to 50°C for a short time).[593] Although trophozoites are killed by low temperatures, cysts can survive for weeks to months at temperatures above freezing.[594] In addition to tropical and sub-tropical locations, *N. fowleri* can also survive in the temperate zone when in association with thermal waters; it has been found in hot springs in Japan and in Yellowstone National Park in the United States.[592,595–597] Unfavorable conditions such as temperature extremes, drying, and lack of nutrients induce cyst formation.[598] Unlike *Entamoeba*, the cyst is not infective. The infective stage is amoeboid, and it is found in soil and water, particularly following heavy precipitation.[599] Changes in temperature or nutrient availability can induce *N. fowleri* to become a flagellate, growing a basal body and flagella *de novo*.[600] In the environment, *N. fowleri* is a bacterivore, and it does not require a host for any stage of its life cycle.

N. fowleri is the only species of *Naegleria* that regularly infects humans, suggesting that pathogenicity is a gain-of-function. *N. fowleri* is also a threat to livestock such as cattle,[601] and can infect other mammals,[602] although the incidence of *N. fowleri* infection in wild animals is unknown. *Naegleria australiensis* and *Naegleria italica* are thermotolerant and can infect mice, although they are considered weakly pathogenic as they are only able to kill mice following intracranial, but not intranasal, inoculation.[360,603,604] While phylogenies showing *Naegleria* species evolution are generally not well-resolved, there is no evidence of these species forming a clade with *N. fowleri.*[605,606] The *Naegleria* species most closely related to *N. fowleri* is *Naegleria lovaniensis*, which is also thermotolerant, but not pathogenic.[607,608] This suggests that pathogenesis and thermotolerance have arisen independently multiple times in the *Naegleria*

genus, and likely go hand-in-hand to allow *Naegleria* spp. to infect mammals with body temperatures between 37°C and 40°C. This raises the question of whether the mammalian host is a dead-end that *N. fowleri* "accidentally" becomes trapped in, or whether pathogenesis is the result of positive selection.

Several groups have identified potential pathogenicity factors in *N. fowleri*, including proteases, pore-forming proteins, and various surface proteins that may be involved in resisting complement-mediated lysis.[350,354,356,358–360,609] However, secretion of pathogenicity factors is not the only aspect of host invasion. Cell motility, immune system evasion, and underlying cellular processes such as membrane trafficking must also play a role. For example, NfActin has been identified as a pathogenicity factor due to its role in food cup formation and potentially also membrane vesiculation or blebbing as an anti-complement measure.[351,355] A whole-genome approach in studying *N. fowleri* in relation to a non-pathogenic relative is the first step in understanding pathogenesis. In addition to gene family sculpting in the systems mentioned above as well as others, this approach could also identify even more potential pathogenicity factors that are specific to *N. fowleri*.

The genome of *Naegleria gruberi* – a harmless relative of *N. fowleri* – was sequenced in 2010.[610] In addition to being related to *N. fowleri*, interest in sequencing *N. gruberi* grew because of its ability to generate a basal body *de novo*, and to perform both anaerobic and aerobic metabolism with a hydrogenosome. The publicly available genome of the harmless *N. gruberi* is therefore an attractive resource for a comparative genomic analysis with *N. fowleri*. To this end, the genome of *N. fowleri* CDC:V212 was sequenced by the *N. fowleri* Genome Consortium. Collaboration with other groups has given access to two additional strains, *N. fowleri* ATCC 30863 and 986. Strains V212 and 30863 were isolated from patients, while 986 is an

environmental isolate. Using these four genomes, the question of the large-scale genomic differences between *N. fowleri* and *N. gruberi* could be addressed. Importantly, an analysis of multiple *N. fowleri* genomes generates a more complete picture of pathogenesis that decreases the chance of identifying strain-specific differences as false positives. In addition to general genomic features, several systems were investigated more closely that were thought *a priori* to be involved in host invasion, including the membrane trafficking system.

The diversity within the *Naegleria* genus is high; based on estimates of the SSU rDNA, the most distantly related *Naegleria* species share the same level of diversity as the entire clade of tetrapods.[611] Furthermore, it is likely that *N. fowleri* and *N. gruberi* are not closely related within the *Naegleria* genus, and many of the differences observed between the two species may be unrelated to pathogenesis. There also might be pathogenesis factors that are found in both organisms, but whose expression is differently regulated such that the factor is involved in pathogenesis in only *N. fowleri*. In order to home in on the most likely pathogenesis factors, as well as to get a sense of the cellular dynamics that underpin human infection, a comparative transcriptomic approach of high-pathogenicity versus normal-pathogenicity *N. fowleri* was undertaken. In 1987, Whiteman and Marciano-Cabral showed that the *N. fowleri* LEE strain – another patient isolate – was more pathogenic after repeated passage through mice (50 mouse passages), compared to *N. fowleri* LEE grown in axenic culture.[607] Not only did mouse-passaged *N. fowleri* have a lower LD50 in guinea pigs ($3x10^4$ LEE-MP50 cells versus $5x10^6$ unpassaged LEE-Ax cells), it was also resistant to complement-mediated killing, while axenically cultured *N. fowleri* LEE and *N. gruberi* were not. To exploit the mouse-passaging enhancement of pathogenesis, RNA-Seq was performed using highly pathogenic (mouse-passaged) and regularly pathogenic (axenically grown) *N. fowleri* LEE. Genes that are differentially expressed in highly

virulent mouse passaged versus axenically grown *N. fowleri* are likely aiding in or responding to host infection. With these data, one can piece together the cellular systems and gene families that are involved in this process, and also build on the comparative genomic work. If, for example, one specific protease is up-regulated in highly pathogenic *N. fowleri*, comparative genomics allows for the exploration of all other members of this gene family in the three other *N. fowleri* strains, as well as *N. gruberi* (e.g. presence/absence, domain composition, conserved active site residues, etc.). The results of the comparative genomic and transcriptomic analyses are synergistic in that they give a clearer and more complete picture of pathogenesis than would be possible taking either approach alone.

Finally, the last part of this chapter goes beyond predicting and assessing membrane trafficking function. The *Naegleria* Golgi does not resemble the typical 'stack of pancakes' found in many cells. This unstacking of Golgi cisternae has occurred several times independently over the course of eukaryotic evolution.[612] In order to visualize the Golgi body for the first time in *N. gruberi*, antibodies to three endomembrane marker proteins were generated (including the Golgi marker NgCOPB), and immunofluorescence microscopy and immunoelectron microscopy were performed, showing the presence of a Golgi body for the first time in this organism. This work relied on the previous comparative genomic and transcriptomic analyses, as they both strongly suggested the presence of a Golgi organelle in *Naegleria*; an assertion that could be tested using basic cell biological techniques.

6.2 Specific methods

An overview of the methods specific to Chapter 6 is found in Methods section 2.4. Collaborators in the *N. fowleri* Genome Sequencing Consortium sequenced and assembled the *N.*

fowleri V212, ATCC 30863, and 986 genomes, described in more detail in the Methods. E. Herman performed gene prediction analyses, comparisons of genome statistics, mitochondrial genome annotation, orthogroup comparison, an analysis of lateral gene transfer, annotation of the membrane trafficking system, the autophagy system, the ER-associated degradation machinery and unfolded protein response, and the adhesion and cell-cell interaction machinery. BLASTP was used in comparative genomics searches, using the queries listed in the appropriate Online Appendix Tables. E. Herman also assembled the HiSeq-generated transcriptome used for gene prediction and transcriptomes of high- and normal-pathogenicity *N. fowleri*, as well as read mapping and differential expression analyses. E. Herman and M. I. Ramirez-Macias performed comparative genomics of the proteases in *N. fowleri*. Work done to confirm the presence of a Golgi body in *N. fowleri* was initially performed by E. Herman, but was then followed up by work from L. Yiangou, D. Cantoni, and A. Tsaousis. The genome of *N. fowleri* ATCC 30863 has already been made public.[613] Additionally, sections related to the *N. fowleri* mitochondrial genome have been published in Herman *et al.* 2013,[417] and are reproduced from that work here. Specifically, nested PCR to confirm correct assembly of a test 60kb genomic region was performed by A. Greninger.

For comparative genomic analysis of the membrane trafficking system machinery, the *N. fowleri* V212 strain was the main genome searched to identify orthologues, which were than used to find corresponding orthologues in the other two strains. A second pass using reference queries was then made to identify any paralogues in 30863 and 986 that were missing from the V212 strain.

6.3 Genomic comparison of three strains of *Naegleria fowleri*: CDC:V212, ATCC 30863,

and 986

6.3.1 Genome statistics

A comparative analysis of genome statistics was performed of the haploid genomes of three *Naegleria fowleri* strains: V212, 986, and ATCC 30863. For the *N. fowleri* V212 genome, 1,875 scaffolds were generated with an average size of 14,806 and an N50 of 92,316. In comparison to the values for the 986 strain (990 contigs, N50 of 101,682) and ATCC 30863 (1,124 scaffolds, N50 of 136,406), there are many more contigs or scaffolds, however, the N50s are relatively similar. 50% of the assembly is contained in contigs of equal or greater lengths than the N50 value. Therefore, the V212 assembly simply has more small contigs than assemblies of the other strains, but the bulk of sequence data is found in longer contigs. The genome data for the *N. fowleri* strains V212, 30863, and 986 are found in Online Appendix Files 1, 2, and 3, respectively.

The genome sizes range between 27.5 and 28 million base pairs, as shown in Table 6.1. Repetitive regions were detected by RepeatScout and RepeatMasker, and these make up 2.28% of the V212 genome (1.41% for strain 986, 1.32% for 30863), compared to 5.1% in *N. gruberi*.[610] While GC content and other coding content statistics were similar between *N. fowleri* strains, they were remarkably different from those values for the related species *N. gruberi* (Table 6.1). The *N. fowleri* genome is, on average, 2/3 of the size of that of *N. gruberi*, with 80% of the coding content.

The mitochondrial genome and extrachromosomal plasmid containing the 18S, 5.8S, and 28S rRNA genes of *N. fowleri* V212 were also recovered from sequence data (Online Appendix file 4). The plasmid sequence was originally published by Maruyama and Nozaki (2007).[614]

Table 6.1. Genome statistics of three strains of *Naegleria fowleri* and *Naegleria gruberi* NEG-M

Species	Genome size	Gene number	Average gene length	Exons/gene
***N. fowleri* V212**	27.7 Mbp	12,677	1,785 bp	2
***N. fowleri* 986**	27.5 Mbp	11,599	1,955 bp	2
***N. fowleri* 30863**	28.0 Mbp	11,499	1,984 bp	2
***N. gruberi* NEG-M**	41.0 Mbp	15,708	1,677 bp	1.7

Species	Average exon length	% Coding	Introns/gene	Average intron length
***N. fowleri* V212**	777 bp	71.35	1	126 bp
***N. fowleri* 986**	849 bp	73.01	1	138 bp
***N. fowleri* 30863**	825 bp	70.79	1	144 bp
***N. gruberi* NEG-M**	894 bp	57.8	0.7	203 bp

Data from *N. gruberi* is taken from Fritz-Laylin *et al.* 2010[610]

6.3.1.1 The mitochondrial genome of Naegleria fowleri V212

The mitochondrial genome of *N. gruberi* was published by Fritz-Laylin *et al*. (2010) along with the nuclear genome.[610] As the *N. fowleri* mitochondrial genome was retrieved in the sequence data, it provided a small, tractable test case for initial genome comparisons. Therefore, the coding content and organization of the mitochondrial genome of *N. fowleri* was compared with that of *N. gruberi*, in order to get a first glimpse into the potential differences between these two organisms.

The mitochondrial genome of *N. fowleri* V212 was sequenced, generating 393,244 reads that were assembled into a circular consensus mitochondrial genome with an average coverage of 2,732X (range of 766-5,317X). The sequence was deposited at GenBank (accession JX174181). The genome is 49,519bp in length and is AT-rich, having a GC content of 25.2% (Table 6.2). Coding sequence comprised 90% of the genome, in which introns were not identified. The *N. gruberi* mitochondrial genome is only slightly larger than the *N. fowleri* mitochondrial genome (49,842bp versus 49,519bp), while their GC contents, coding contents, and median exon lengths are very similar (Table 6.2).

The gene complement encodes products involved in reduction and oxidative phosphorylation in the mitochondrion, such as the NADH dehydrogenases, ATP synthase subunits, cytochrome c oxidase subunits, apocytochrome b, and succinate:cytochrome c oxidoreductase (Supplementary Table ST6.1). It encodes three proteins involved in protein import and maturation: haem biosynthesis, haem lyase, and an ABC transporter. It also encodes ribosomal proteins, rRNAs, tRNAs, a SecY-independent transporter protein (Ymf16), and four ORFs of unknown function. The majority of the genes are ribosomal proteins, tRNAs, or are involved in reduction and oxidative phosphorylation. In comparison with the mitochondria of

Table 6.2. Mitochondrial genome statistics from *Naegleria gruberi* and *Naegleria fowleri.*

Species	**Size (bp)**	**%GC**	**%Coding**	**Exons per gene**	**Median exon length**
N. gruberi mitochondria[a]	49,842	22	92	1	795 bp
N. fowleri mitochondria	49,519	25.2	90	1	766.5 bp

[a]*N. gruberi* nuclear genome data are from Fritz-Laylin et al. 2010.[610]

other eukaryotes, shown in Figure 6.1, the proportions of genes in each category are standard among the diversity observed in other species.

Figure 6.2 shows a diagram of genes encoded in the circular *N. fowleri* mitochondrial genome. The genes are tightly packed, as only 10% of the genome is non-coding sequence. Regarding gene order, the mitochondrial genomes of *N. fowleri* and *N. gruberi* are entirely syntenic. ORF prediction software failed to identify any protein-encoding regions in the *N. fowleri* genome corresponding to genes not present in *N. gruberi,* and all annotated *N. gruberi* genes were found in *N. fowleri*.

The *Naegleria* sp. mitochondrial genomes have a conserved bacteria-like order of genes encoding small and large ribosomal proteins. The similarity extends over a contiguous alignment of the *str*, S10, *spc*, and α operons (Figure 6.2). Although there are some losses, rearrangements, and insertions of genes without a bacterial origin, the gene order remains widely similar. Related organisms share this bacteria-like organisation, including the jakobids *Jakoba libera* and *Reclinomonas americana*, and the malawimonad *Malawimonas jakobiformis*.[616]

In addition to an analysis of the *N. fowleri* mitochondrial genome, a 60-kilobase contig was manually annotated. This was done to get a sense of the quality of the assembly, and served as a first-pass at whole-genome analysis. Thirty-three genes were identified on that contig, but surprisingly, their *N. gruberi* orthologues did not share a similar gene order. To confirm that the *N. fowleri* V212 genome was not misassembled, nested PCR was performed of this region (Figure 6.3), and showed that it is indeed contiguous.[417]

Figure 6.1. Gene content of the *Naegleria fowleri* and *Naegleria gruberi* mitochondrial genomes in comparison with other eukaryotes.

Genes are included in the classes in the following way, as in Burger *et al.* (2003).[615] Reduction and oxidative phosphorylation (horizontal bars): *atp1, atp3, atp4, atp6, atp8, atp9*; *cob*; *cox1-3*; *nad1-4, nad4L, nad6-11, sdh2-4*. rRNAs (solid black): *rnl*, *rns*, *rrn5*. tRNAs (diagonal dots): *trnA-Y*. Ribosomal proteins and EF-Tu (solid dark grey): *rps1-4, rps7, rps8, rps10-14, rps19; rpl1, rpl5, rpl6, rpl10, rpl11, rpl14, rpl16, rpl18-20, rpl27, rpl31, rpl32, rpl34, rpl36*; *tufA*. Protein import and maturation (solid white): *secY, ymf16, tatC, yejR (ccmF), yejU (ccmC), yejV (ccmB), yejW (ccmA)*; *cox11*. RNA maturation (solid light grey): *rnpB*. Transcription (black squares): *rpoA-D*. Other (hatched): ORFs of unknown function. Data for all organisms except *N. fowleri* was retrieved from NCBI. In *Dictyostelium discoideum*, *cox1* and *cox2* are encoded as a single ORF, but are counted as two genes here. Some other species of *Plasmodium* also encode ribosomal RNAs.

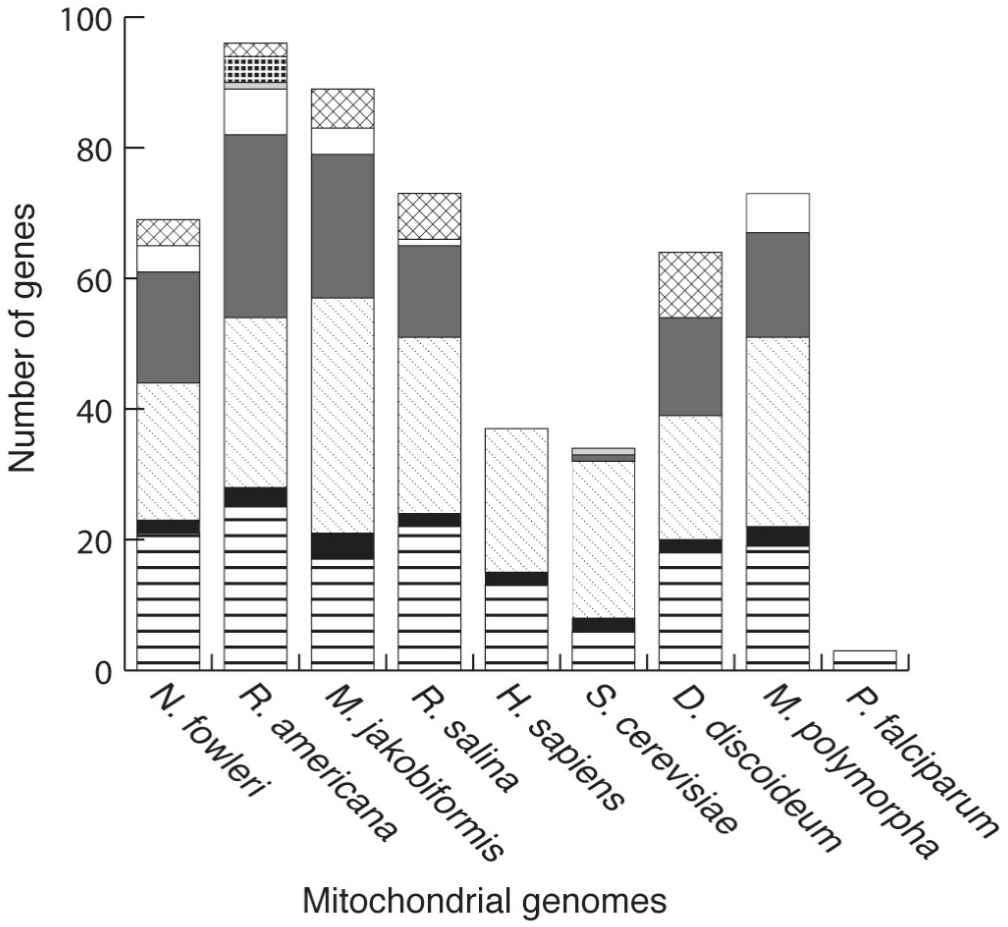

100
80
60
40
20
0
Number of genes
N. fowleri
R. americana
M. jakobiformis
R. salina
H. sapiens
S. cerevisiae
D. discoideum
M. polymorpha
P. falciparum
Mitochondrial genomes

Figure 6.1

Figure 6.2. Circular map of the *Naegleria fowleri* mitochondrial genome.

Genes encoding proteins are annotated based on BLAST results, and genes encoding RNA are annotated based on tRNA and rRNA scanning software predictions. For the full name of each gene, see Supplementary Table ST6.1. Black arrows with gene names on the outside of the map represent genes on one strand, and white arrows with gene names on the inside of the map represent genes on the alternate strand. The *N. fowleri* mitochondrial genome is identical in gene content and organization to the mitochondrial genome of *Naegleria gruberi*.

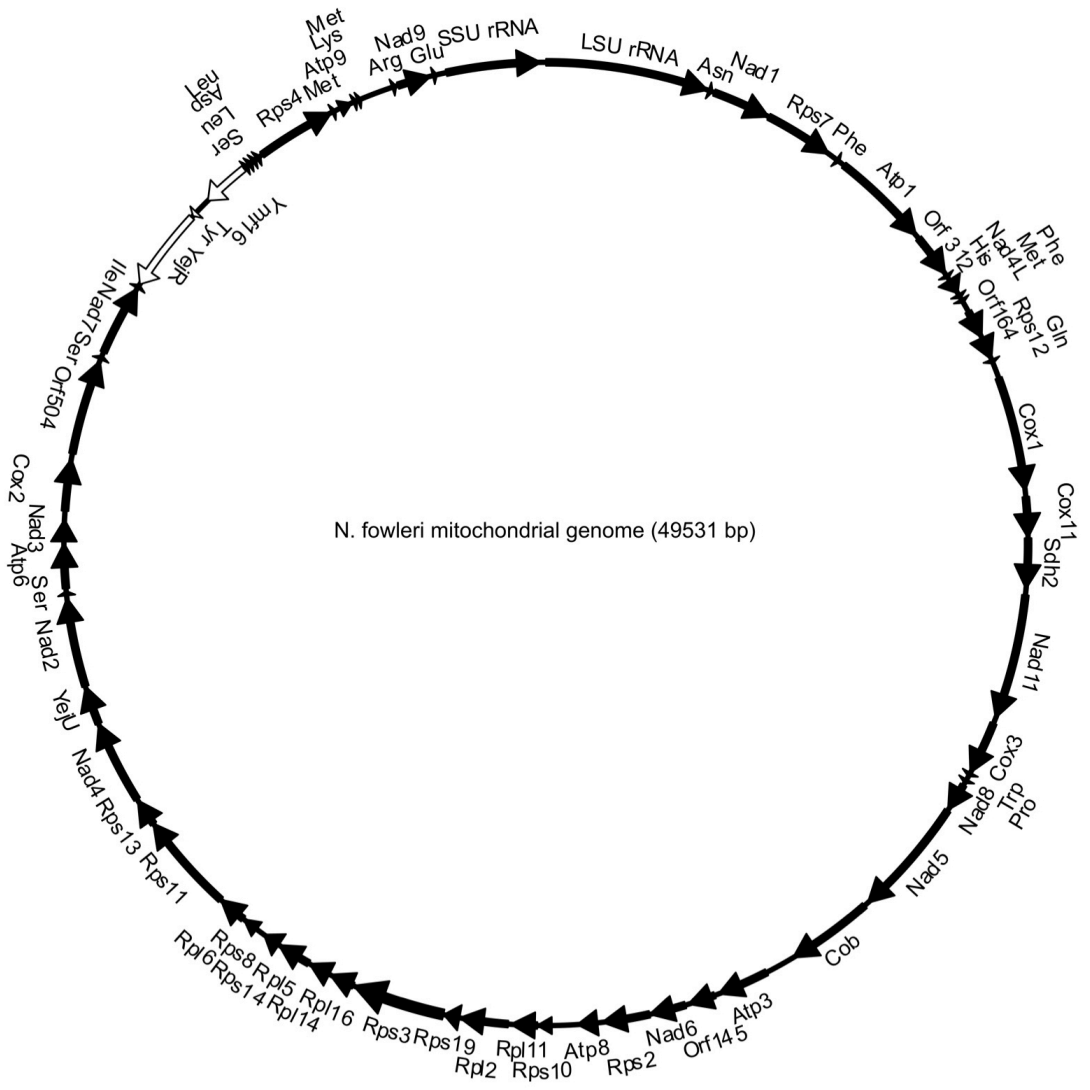

N. fowleri mitochondrial genome (49531 bp)
SSU rRNA
LSU rRNA
Nad9
Nad1
Rps7
Phe
Atp1
Orf 312
Nad4L
Met
Phe
Rps12
Gln
Orf164
Cox1
Cox11
Sdh2
Nad11
Cox3
Trp
Pro
Nad8
Nad5
Cob
Atp3
Orf145
Nad6
Atp8
Rps2
Rpl11
Rps10
Rpl2
Rps19
Rps3
Rpl16
Rpl5
Rps8
Rpl14
Rps14
Rpl6
Rps11
Rps13
Nad4
YejU
Nad2
Ser
Atp6
Nad3
Cox2
Orf504
IleNad7
Ser
Tyr
YmfR
YmfI6
Ser
Leu
Asp
Leu
Rps4
Met
Atp9
Lys
Met
Arg
Glu

Figure 6.2

Figure 6.3. Nested PCR showing contiguity of a 60kb genomic region.

(Top) Eleven ~1000 bp amplicons from across the 60-kb segment in the nuclear genome were successfully recovered by PCR amplification. **(**Bottom**)** Diagram of where the amplicons were located across the contig. M, ladder. PCR was performed by A. Greninger.

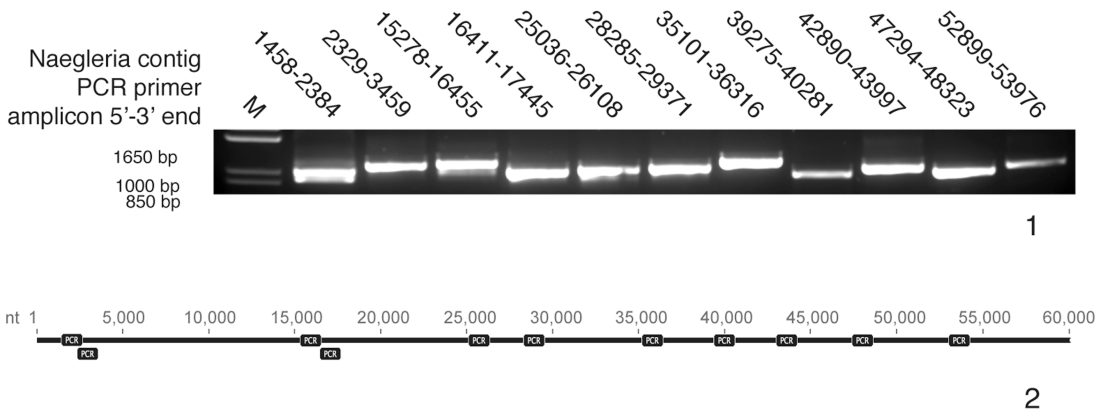

Naegleria contig
PCR primer
amplicon 5'-3' end
M
1458-2384
2329-3459
15278-16455
16411-17445
25036-26108
28285-29371
35101-36316
39275-40281
42890-43997
47294-48323
52899-53976
1650 bp
1000 bp
850 bp
1
nt 1
5,000
10,000
15,000
20,000
25,000
30,000
35,000
40,000
45,000
50,000
55,000
60,000
PCR
2

Figure 6.3

6.3.2 Gene prediction of nuclear genes

Using a combination of RNA-Seq evidence and *ab initio* methods, the gene prediction tool Augustus v.3.0.3 predicted 12,677 coding genes across 71.35% of the *N. fowleri* V212 genome (Table 6.1). 89% of the predicted V212 genes appear to be expressed (using pathogenicity transcriptomic data), based on at least five reads mapping to the gene. As a measure of genome completeness, a CEGMA analysis to identify eukaryotic core orthologous groups (KOGs) was performed. Out of 458 KOGs, 443 (97%) were identified in the *N. fowleri* V212 predicted proteome. This value is identical for the two other *N. fowleri* strains.

Using these same training data, gene prediction was also performed for strains 30863 and 986. The statistics of these gene predictions are compared with the corresponding publicly available data from the *N. gruberi* genome (Table 6.1). Protein predictions for *N. fowleri* V212, 30863, and 986 are found in Online Appendix Files 5, 6, and 7, respectively. Surprisingly, these values are highly variable within the *Naegleria* genus, and even within the *N. fowleri* species. *N. fowleri* V212 encodes the most genes at 12,677, with the 986 and 30863 strains encoding on average 11,600 genes. The *N. gruberi* genome, which is larger by roughly 13 Mbp, encodes 15,708 genes. In general, average genes lengths are greater in *N. fowleri* (1,785-1,984 bp) than *N. gruberi* (1,677 bp), and the *N. fowleri* genome has more exons per gene and a higher percent coding content (71.79-73.01% in *N. fowleri* versus 57.8% in *N. gruberi*). Despite these differences, the average length of exons is similar between *N. fowleri* and *N. gruberi*, with the *N. gruberi* genome encoding genes with longer exons, on average. Introns in *N. fowleri* genes are roughly 50-60% of the length of those in *N. gruberi* genes, although *N. fowleri* has more introns per gene.

6.3.3 Genome organization of three *N. fowleri* species and the related *Naegleria gruberi*

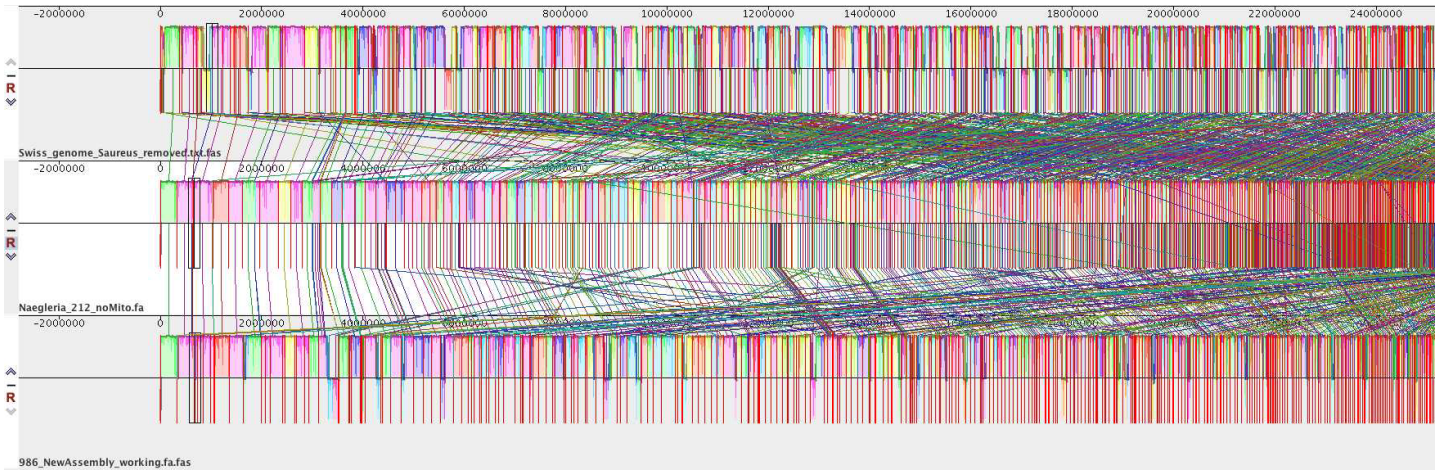
Further exploring large-scale genomic differences, the genome organization of the four *Naegleria* species were compared. Mauve was used to align the *Naegleria* genomes and identify regions of collinearity. Figure 6.4 shows the organizational differences between the three *N. fowleri* strains, and Figure 6.5 shows this including the *N. gruberi* genome. Even between strains there is an extremely high lack of synteny, and it is even more pronounced when the *N. gruberi* genome is included. This is likely not an artifact of short contigs or misassembly, since rearrangements and inversions can be seen between any two genomes within long stretches of contiguous sequence.

For comparison to other eukaryotes, similar analyses were performed with three *T. brucei* strains, and with two members of the *Saccharomyces* spp. species complex and one sister taxon, *Saccharomyces castelli*. In contrast to *Naegleria* spp., the genomes of the three *T. brucei* strains *T. brucei gambiense* (DAL972), *T. brucei brucei* (TREU927), and *T. brucei* Lister 427, are highly syntenic with minimal rearrangement (Supplementary Figure S6.1). Within the *Saccharomyces* spp. complex – between *S. cerevisiae* and *Saccharomyces uvarum* – there is more genomic rearrangement (Supplementary Figure S6.2). Including *Saccharomyces castellii*, which is a separate species outside of the *Saccharomyces* spp. complex, shows genome rearrangement to a similar extent as between the different *N. fowleri* strains and species. One explanation for the lack of synteny is sexual recombination events following the divergence of the *Naegleria* species from a common ancestor. The genes required for meiosis are present in *N. gruberi*, and population studies of *N. lovaniensis* show evidence of a sexual cycle.[361,617]

Figure 6.4. Genomic alignment of *N. fowleri* strains V212, 30863, and 986.

Mauve output showing synteny between strain 30863 (top), V212 (middle), and 986 (bottom); (A) full synteny comparison, (B) expanded region. Coloured blocks indicate regions of contiguous sequence on each strand (upper versus lower blocks). Lines connecting blocks between each genome show the extent of sequence rearrangement in the three strains. Red vertical lines delineate contigs or scaffolds.

A


Swiss_genome_Saureus_removed.txt.fas
Naegleria_212_noMito.fa
986_NewAssembly_working.fa.fas

B

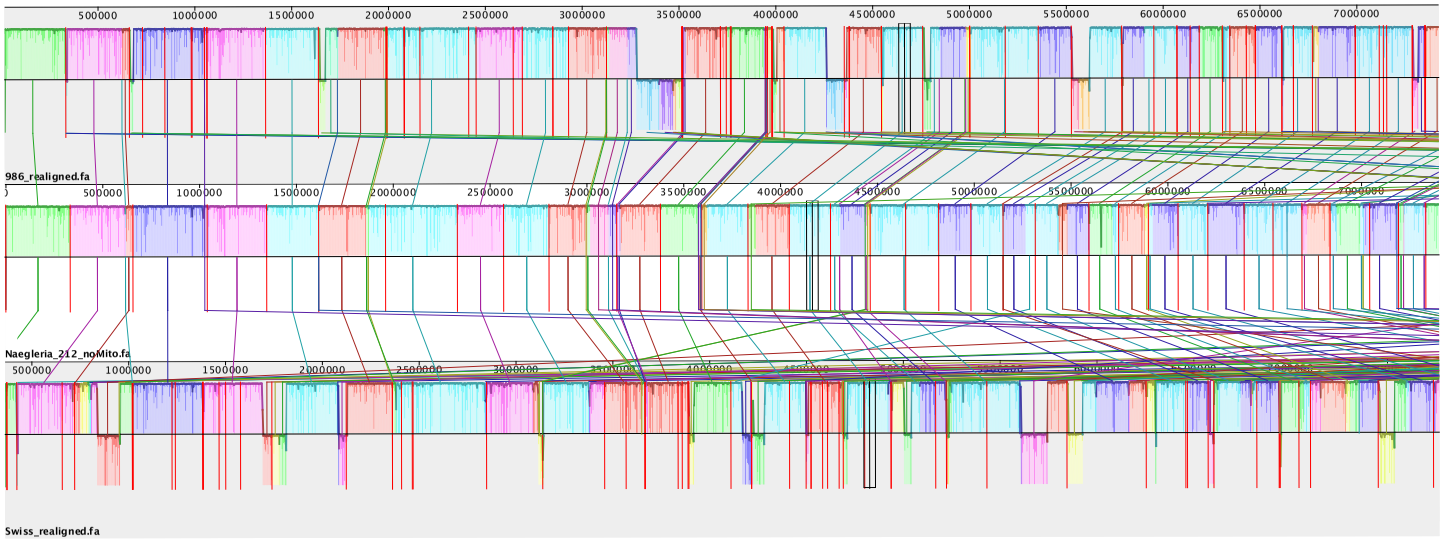

986_realigned.fa
Naegleria_212_noMito.fa
Swiss_realigned.fa

Figure 6.4

Figure 6.5 Genomic alignment of *N. gruberi* and the *N. fowleri* strains V212, 30863, and 986.

Mauve output showing synteny between *N. fowleri* V212 (top), *N. gruberi* (second from top), *N fowleri* 30853 (third from top), and *N. fowleri* 986 (bottom). Coloured blocks indicate regions of contiguous sequence on each strand (upper versus lower blocks). Lines connecting blocks between each genome show the extent of sequence rearrangement in the three strains. Red vertical lines delineate contigs or scaffolds.

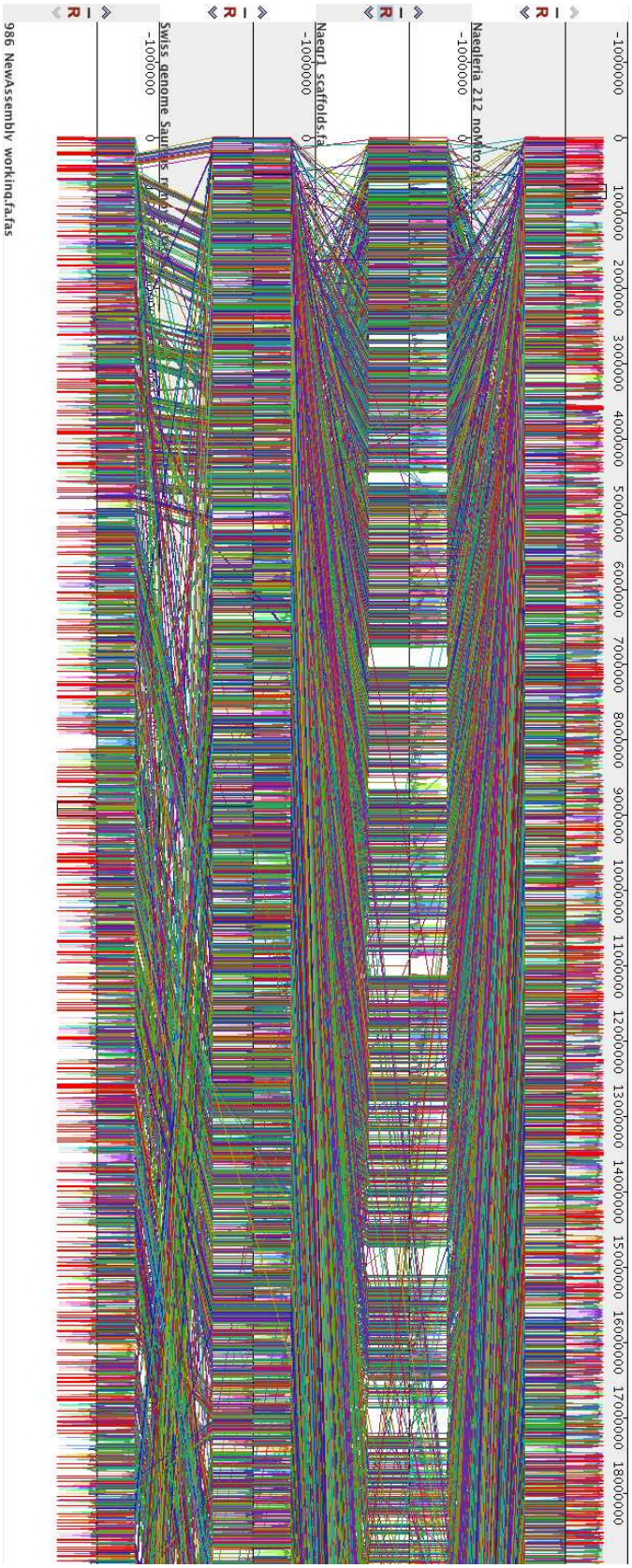

Naegleria_212_noMito.fa
Naegr1_scaffolds.fa
Swiss_genome_Saunas_raw
986_NewAssembly_workingfa.fas
-1000000
0
1000000
2000000
3000000
4000000
5000000
6000000
7000000
8000000
9000000
10000000
11000000
12000000
13000000
14000000
15000000
16000000
17000000
18000000

Figure 6.5

6.3.4 Identifying orthologous groups of genes

One of the questions central to understanding pathogenesis is how *N. fowleri* differs in terms of gene content from *N. gruberi*. Of 40 species, only *N. fowleri*, *N. australiensis,* and *N. italica*, which are not closely related,[606] are capable of infecting animals;[603] therefore, it is likely that pathogenesis is an independent gain-of-function in these taxa. To get a broad view of gene content, OrthoMCL was used to identify orthologous groups of genes shared between all four, or a subset of, the *Naegleria* genomes (Figure 6.6). Out of 11,399 groups, 7,656 (67%) are shared by all four *Naegleria* species, and 10,451 (92%) are shared by all three *N. fowleri* strains. There are 3,192 groups not in *N. gruberi* that are shared between two or more *N. fowleri* strains, and 2,795 groups not in *N. gruberi* that are shared by all three *N. fowleri* strains. BLAST searches into the *N. gruberi* scaffolds and predicted proteins, and manual verification of orthology allowed us to reduce the number of false positives in the latter set of orthologous groups. Using this approach, 936 proteins were identified with orthologues in all three *N. fowleri* strains, but not in *N. gruberi*. Their distribution and annotation based on NR BLAST search is found in Online Appendix Table 6.1. 625 (67%) of these are sequences unique to *N. fowleri*, i.e. with no detectable homologue in NCBI's non-redundant database. Of those with homologues in other organisms, the majority are proteins with unknown function. There were only 199 proteins (21%) identified that either could be functionally annotated based on NR BLAST results, or contained a characterized domain.

Genes that could be annotated based on the results of the NR BLAST search varied in function; some examples include GPI modification, purine metabolism, histone deacetylation, transcription, selenoproteins and selenium transport, and membrane trafficking proteins.

Figure 6.6. Orthogroup distribution in three strains of *N. fowleri* and in *N. gruberi*.

Venn diagram showing the shared orthogroups between different strains of *N. fowleri* and *N. gruberi*. Numbers in overlapping segments denote the shared orthogroups between various genomes, while numbers in non-overlapping segments denote in-paralogues for each genome.

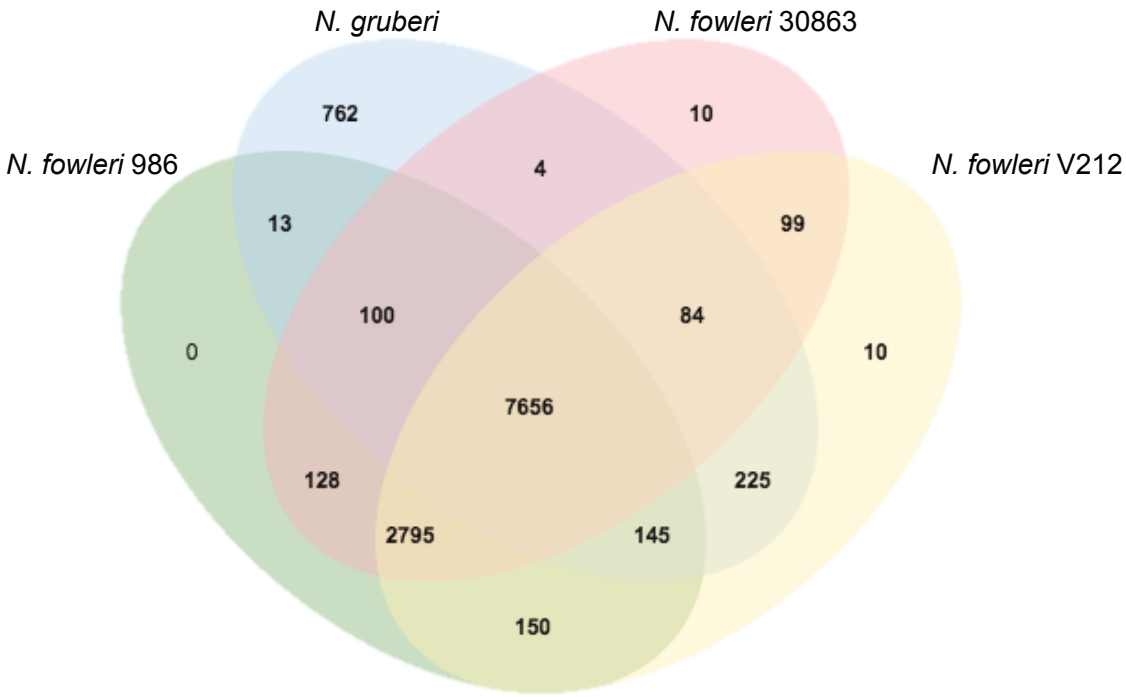

N. gruberi
N. fowleri 30863
N. fowleri 986
N. fowleri V212
762
10
4
13
99
100
84
0
10
7656
128
225
2795
145
150

Figure 6.6

However, there are no functional categories that appeared to be over-represented, and none of the genes are obviously related to pathogenesis.

Lateral gene transfer (LGT) events may be a source of *N. fowleri* pathogenicity factors, particularly from organisms that cause meningitis or other pathologies. LGT is the transfer of genetic material between unrelated organisms, i.e. not between mother and daughter cells. To determine the extent of LGT specifically associated with a gain of function, proteins unique to *N. fowleri* with top NR BLAST hits to bacteria or unrelated eukaryotes were analysed. The category of genes of bacterial origin without orthologues in any other eukaryotes is small; only 52 sequences were identified, and most are of unknown function. Annotation based on NR BLAST results did not identify any sequences that are likely pathogenicity factors in Bacteria. Furthermore, none of the bacterial lineages with orthologues most similar to the *N. fowleri* sequences (based on sequence similarity) are known to cause meningitis. Only two sequences had top BLAST hits to *Burkholderiales* spp. bacteria, which are opportunistic pathogens of humans that can live endosymbiotically within another free-living amoeba, *Acanthamoeba* spp. *Legionella pneumophila* is a known endosymbiont of *Naegleria*;[618] however, none of the potential bacterial LGTs had a top BLAST hit from this organism. Symbioses between amoebae and bacteria has been shown to benefit both organisms,[619] however, there is no evidence for the direct involvement of bacterial endosymbionts in human infection. There are also several genes that seem to be of viral or archaeal origin, or are similar to transposable elements, however, they cannot be annotated.

6.4 Comparative genomics in *Naegleria fowleri* and *Naegleria gruberi*

Several cellular systems were thought to be involved in pathogenesis *a priori* based on how infection occurs. To this end, comparative genomics was used to explore the differences between *N. fowleri* and *N. gruberi* related to the membrane trafficking system, cell stress response, autophagy machinery, and adhesion and cell-cell interacting factors.

6.4.1 Membrane trafficking

As the membrane trafficking system is responsible for movement of material into, out of, and around the cell, it is likely to be central to pathogenesis, which relies on secretion of pathogenicity factors such as cysteine proteases, and phagocytosis or trogocytosis (i.e. piece-meal eating) of host cells.

Both *N. fowleri* and *N. gruberi* have remarkably complete membrane trafficking machinery repertoires (Figures 6.7-6.9, Online Appendix Table 6.2). This strongly supports the notion that *N. fowleri* is not a parasite, nor is reliant on a human or animal host in any way. There also does not appear to be sculpting of the trafficking system, as there are no whole-complex losses. There are instances of single gene and whole (or nearly whole) complex duplication, but these are present in both *N. fowleri* and *N. gruberi*.

In addition to the vesicle formation machinery being nearly completely conserved in *Naegleria* spp., members of both the TSET coat and AP5 adaptor protein complexes were identified (Figure 6.7). TSET and AP5 have a patchy distribution in eukaryotes, and as such, were only recently discovered.[144,147] Both species have a complete TSET complex, which has been shown to function at the plasma membrane in *Dictyostelium*.[144] *N. fowleri* and *N. gruberi* also have a partial AP5 complex, including subunits AP5B and AP5M in both organisms, and

Figure 6.7 Comparative genomic survey of vesicle formation machinery in *N. fowleri* and *N. gruberi*.

Coulson plot showing the complement of ESCRT machinery, coat complexes, and adaptor proteins in *N. fowleri* V212, 986, and 30863, and *N. gruberi*. The vesicle formation machinery is highly complete between the two *Naegleria* species.

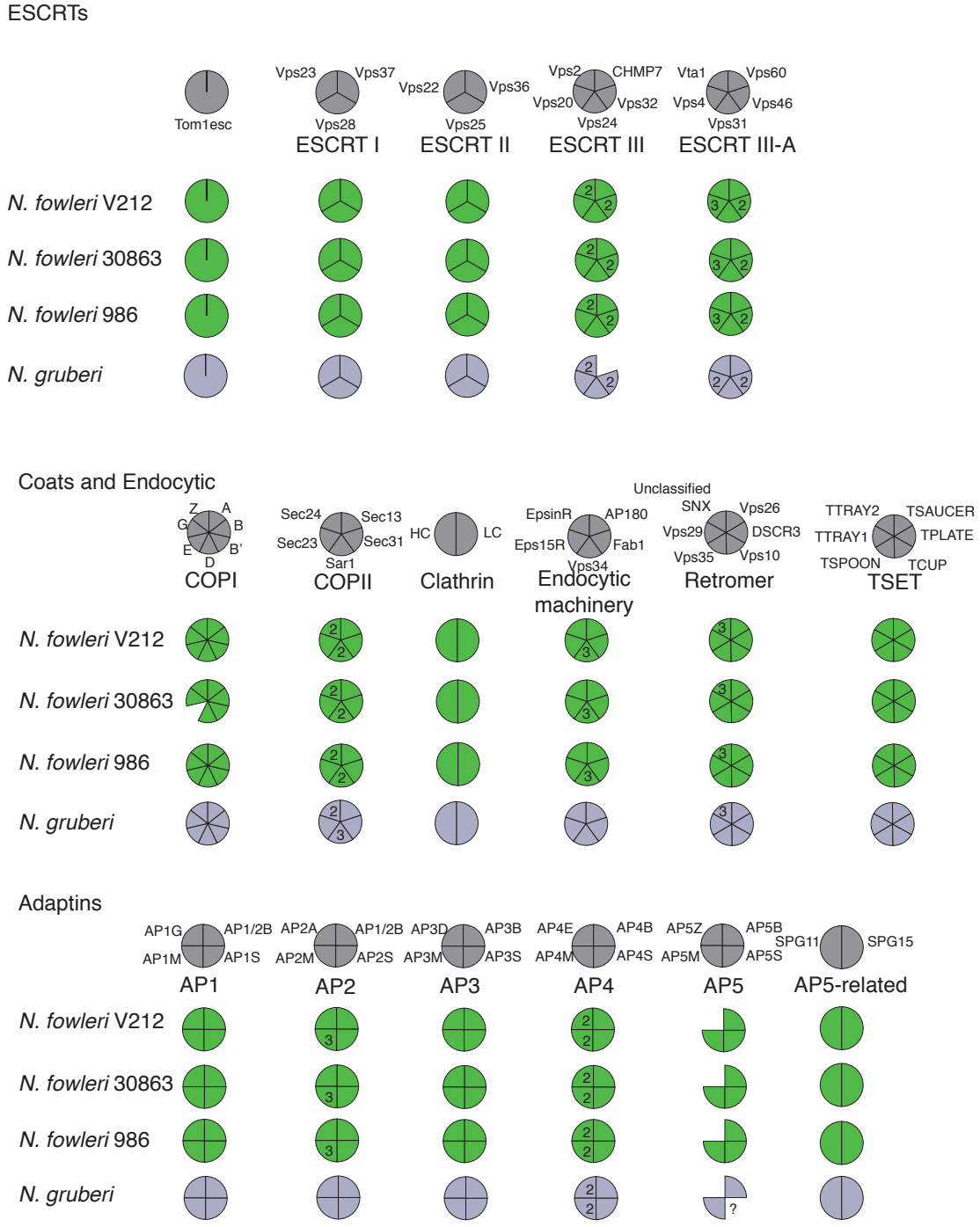

ESCRTs
Tom1esc
Vps23
Vps37
Vps28
ESCRT I
Vps22
Vps36
Vps25
ESCRT II
Vps2
CHMP7
Vps20
Vps32
Vps24
ESCRT III
Vta1
Vps60
Vps4
Vps46
Vps31
ESCRT III-A
N. fowleri V212
N. fowleri 30863
N. fowleri 986
N. gruberi
Coats and Endocytic
Z
A
G
B
E
B'
D
COPI
Sec24
Sec13
Sec23
Sec31
Sar1
COPII
HC
LC
Clathrin
EpsinR
AP180
Eps15R
Fab1
Vps34
Endocytic machinery
Unclassified
SNX
Vps26
Vps29
DSCR3
Vps35
Vps10
Retromer
TTRAY2
TSAUCER
TTRAY1
TPLATE
TSPOON
TCUP
TSET
N. fowleri V212
N. fowleri 30863
N. fowleri 986
N. gruberi
Adaptins
AP1G
AP1/2B
AP1M
AP1S
AP1
AP2A
AP1/2B
AP2M
AP2S
AP2
AP3D
AP3B
AP3M
AP3S
AP3
AP4E
AP4B
AP4M
AP4S
AP4
AP5Z
AP5B
AP5M
AP5S
AP5
SPG11
SPG15
AP5-related
N. fowleri V212
N. fowleri 30863
N. fowleri 986
N. gruberi

Figure 6.7

Figure 6.8 Comparative genomic survey of vesicle fusion machinery in *N. fowleri* and *N. gruberi*.

Coulson plot showing the complement of SNAREs, SM proteins, and multi-subunit tethering complexes in *N. fowleri* V212, 986, and 30863, and *N. gruberi*.

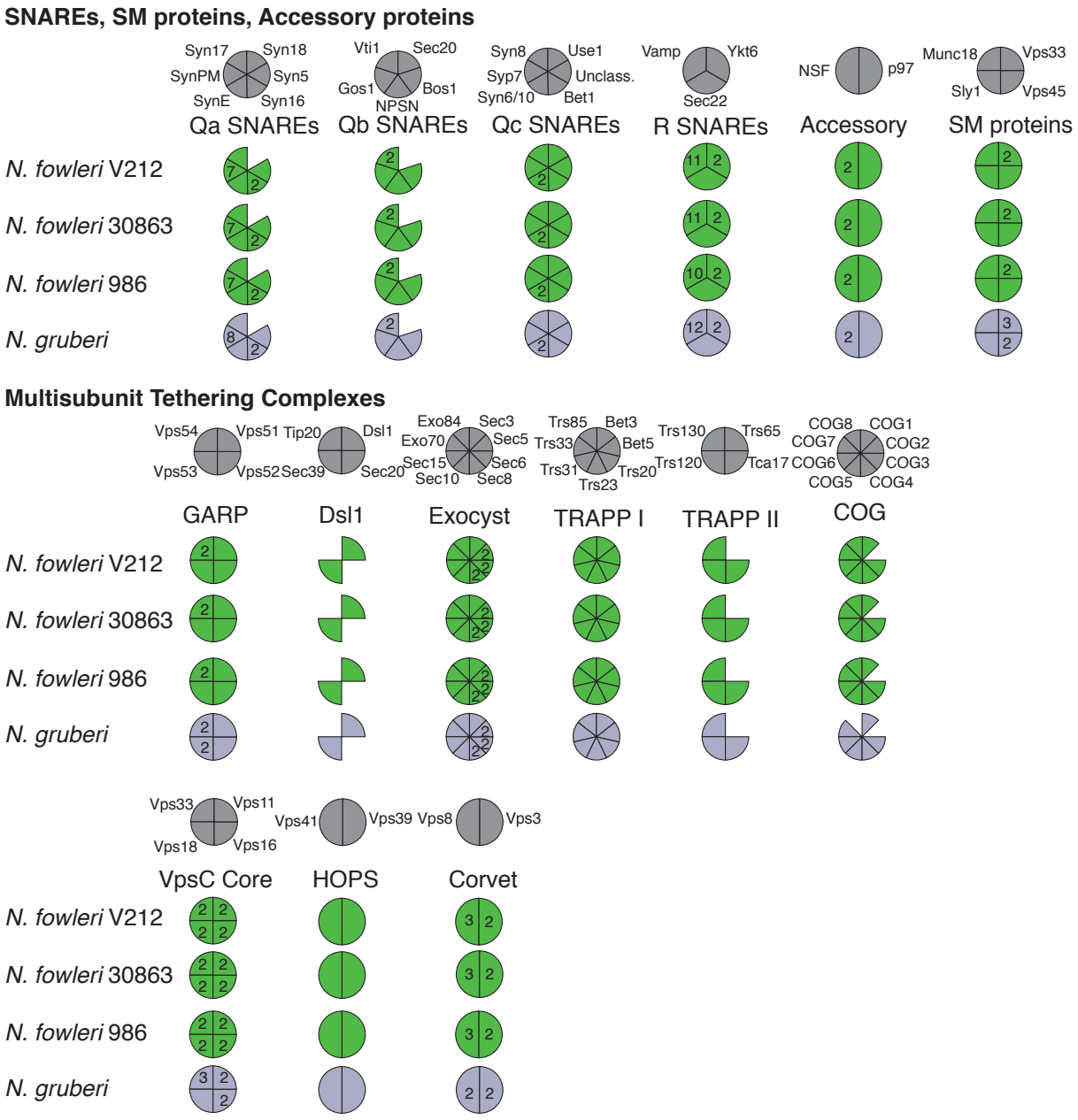

SNAREs, SM proteins, Accessory proteins
Syn17
Syn18
SynPM
Syn5
SynE
Syn16
Vti1
Sec20
Gos1
Bos1
NPSN
Syn8
Use1
Syp7
Unclass.
Syn6/10
Bet1
Vamp
Ykt6
Sec22
NSF
p97
Munc18
Vps33
Sly1
Vps45
Qa SNAREs
Qb SNAREs
Qc SNAREs
R SNAREs
Accessory
SM proteins
N. fowleri V212
N. fowleri 30863
N. fowleri 986
N. gruberi
Multisubunit Tethering Complexes
Vps54
Vps51
Vps53
Vps52
Tip20
Dsl1
Sec39
Sec20
Exo84
Sec3
Exo70
Sec5
Sec15
Sec6
Sec10
Sec8
Trs85
Bet3
Trs33
Bet5
Trs31
Trs20
Trs23
Trs130
Trs65
Trs120
Tca17
COG8
COG1
COG7
COG2
COG6
COG3
COG5
COG4
GARP
Dsl1
Exocyst
TRAPP I
TRAPP II
COG
N. fowleri V212
N. fowleri 30863
N. fowleri 986
N. gruberi
Vps33
Vps11
Vps18
Vps16
Vps41
Vps39
Vps8
Vps3
VpsC Core
HOPS
Corvet
N. fowleri V212
N. fowleri 30863
N. fowleri 986
N. gruberi

Figure 6.8

Figure 6.9 Comparative genomic survey of Arf and Rab GTPase regulators in *N. fowleri* and *N. gruberi*.

Coulson plot showing the complement of ArfGAPs, ArfGEFs, TBC RabGAPs, and DENN domain RabGEFs in *N. fowleri* V212, 986, and 30863, and *N. gruberi*.

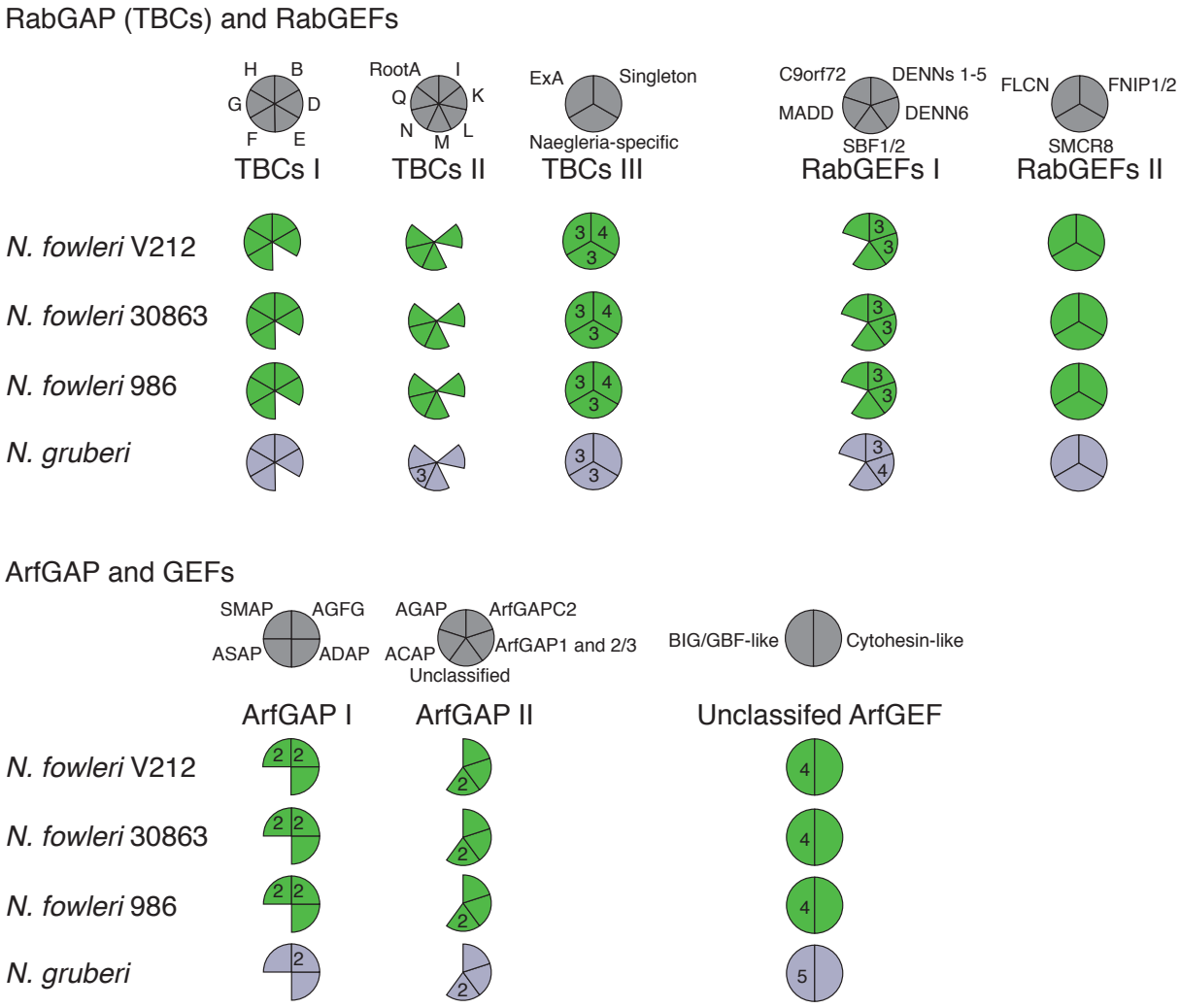

RabGAP (TBCs) and RabGEFs
H
B
G
D
F
E
TBCs I
RootA
I
Q
K
N
M
L
TBCs II
ExA
Singleton
Naegleria-specific
TBCs III
C9orf72
DENNs 1-5
MADD
DENN6
SBF1/2
RabGEFs I
FLCN
FNIP1/2
SMCR8
RabGEFs II
N. fowleri V212
N. fowleri 30863
N. fowleri 986
N. gruberi
ArfGAP and GEFs
SMAP
AGFG
ASAP
ADAP
AGAP
ArfGAPC2
ACAP
ArfGAP1 and 2/3
Unclassified
BIG/GBF-like
Cytohesin-like
ArfGAP I
ArfGAP II
Unclassifed ArfGEF
N. fowleri V212
N. fowleri 30863
N. fowleri 986
N. gruberi

Figure 6.9

AP5S in *N. fowleri*. AP5 has been localized to late endosomes in Hirst *et al.* (2013), and is only partially found in other excavate organisms.[169]

The ESCRTs are complete, with the absence of CHMP7 only in *N. gruberi*. Paralogue numbers are very similar between the two species. Similarly, the coats, adaptins, and endocytic machinery are complete (with the exception of AP5) in *Naegleria*. In general, there are few instances of gene duplication. In *N. fowleri*, three paralogues of AP2M were identified, while only one could be found in *N. gruberi*. AP2 is the adaptor complex involved in clathrin-mediated endocytosis, and the medium subunit functions in cargo binding.[620,621] It is possible that the multiple paralogues of AP2M allow for discrimination between different types of endocytic cargoes in *N. fowleri*. Furthermore, in both *N. fowleri* and *N. gruberi*, the medium (μ) and large (ε) subunits of AP4 are duplicated. AP4 functions at the TGN,[622] and again, these duplications may allow for functional divergence of the subunits and therefore increase the specificity/complexity of trafficking events at this organelle.

In terms of the vesicle fusion machinery, there are both patterns of loss and expansion (Figure 6.8). First, there is a general loss of trafficking factors associated with retrograde transport from the Golgi to the ER. The SNAREs Syntaxin 18 and Sec20 could not be identified in either *N. fowleri* or *N. gruberi*, both of which act in this step (Qa, Qb, and Qc SNAREs are classified in Supplementary Figures S3-S5, Online Appendix Table 3). Furthermore, only half of the Dsl1 MTC subunits are present, and therefore it is not clear whether this complex is functional. As mentioned in Chapter 3, the loss of Dsl1 is often concomitant with the loss or degradation of peroxisomes. Previous work on the peroxin genes of *N. fowleri* showed that although *Naegleria* lacks many of the typical peroxins, its genome does encode the metabolic

enzymes of peroxisomes.[610] Therefore, *Naegleria* may have modified peroxisomes, but experimental work is required to conclusively show this.

There is notable expansion of plasma membrane targeted trafficking machinery in both organisms. *N. fowleri* encodes 7 paralogues of Syntaxin PM-like proteins, and 11 VAMP-like SNAREs, while *N. gruberi* encodes 8 and 12, respectively. Furthermore, both have duplicated Sec5, Sec6, and Sec8 subunits of the Exocyst MTC. This expansion of exocytic machinery strongly suggests that there is increased complexity and specificity of trafficking in this pathway. There is also evidence of expansion of some endocytic and sorting machinery. Both species have duplicated the VpsC core and CORVET MTC subunits (endosome-lysosome fusion), as well as the SM protein Vps33 that also functions in early and late endosome trafficking. Some recycling machinery is also expanded. In addition to the duplication of two AP4 subunits mentioned above, the SNARE Syntaxin 16, and part of the GARP MTC are duplicated in *N. fowleri* and *N. gruberi*, and the corresponding SM protein Vps45 is also duplicated in *N. gruberi*. These expansions suggest that exocytosis, recycling, and endolysosomal degradation pathways have additional trafficking complexity.

Arf and Rab GTPase regulators are generally functionally uncharacterized outside of mammalian cells (Figure 6.9). ArfGAPs are classified in Supplementary Figure S6 (Online Appendix Table 3). Like other trafficking machinery, *N. fowleri* and *N. gruberi* have a similar repertoire. Both *N. fowleri* and *N. gruberi* encode ArfGAPC2, a newly discovered but ancient ArfGAP of unknown function.[199] This is the first example of an ArfGAPC2 protein present in the Excavata supergroup, further supporting its presence in the LECA. Two AGFG ArfGAPs, which function in endocytosis,[205] are also found in both organisms. Based on the complement of TBC Rab GAPs deduced to be present in the LECA,[330] both organisms encode a relatively complete

set. TBC proteins in *Naegleria* are classified in Supplementary Figure S7 (Online Appendix Table 3). *N. fowleri* also has four TBC proteins with no *N. gruberi* orthologue, one of the only large differences in trafficking component paralogue number. A large contingent of DENN domain-containing RabGEFs was found in both taxa, as well as several members of the BIG/GBF-like ArfGEFs, which work at the Golgi.[210]

6.4.2 Autophagy

It is possible that host infection is not an optimal life strategy for *N. fowleri*, and there may be differences in the cell stress responses of *N. fowleri* and *N. gruberi* that allow the former to cope with host conditions and survive. Autophagy, the unfolded protein response, and ER-associated degradation are all related to cell stress,[623–625] and were therefore the subject of comparative genomic analysis.

Autophagy is a process of degrading damaged organelles and other cytoplasmic material such as protein aggregates, and can be triggered by various cell stresses. It feeds into the lysosomal degradation pathway, as autophagosomes fuse with lysosomes to degrade their contents. The autophagy machinery was queried in *Naegleria* spp. for two reasons: to gain a deeper understanding of processes related to late endosome trafficking, as genes in this pathway have undergone duplication events, and to determine whether there are genomic differences between *N. gruberi* and *N. fowleri* that might be relevant to host infection, if host infection is a stressor for *N. fowleri*.

In macroautophagy, the first step is the formation of phagophore assembly sites (PAS) in the cytosol. The source of this membrane is unclear, but there is evidence to suggest that it is

derived from the ER.[626] Although autophagy induction machinery has been identified, only ATG1 is broadly conserved in eukaryotes.[473] Following induction, a macromolecular initiation complex containing Vps34, Atg6, Atg14, and Vps15 functions in the early stages of autophagy to generate phosphatidylinositol 3-phosphate and function in PAS formation.[627] As the autophagosome membrane grows, two ubiquitination reaction pathways promote the phosphatidylethanolamine (PE) lipidation of the protein LC3-I to form membrane-associated LC3-II (Atg3, Atg4, Atg7, Atg8; Atg12, Atg7, Atg10, Atg5).[627] The presence of LC3-II on autophagosome membranes may allow for autophagosome expansion, as it mediates membrane tethering and fusion. Mature autophagosomes fuse with lysosomes by HOPS-mediated tethering. Additional machinery is required to retrieve elongation factors from the outer membrane (Atg2, Atg9, Atg18).

Like other machinery in *Naegleria*, the autophagy system is highly complete, with very few differences between *N. fowleri* and *N. gruberi* (Figure 6.10, Online Appendix Table 6.4). With the exception of ATG17 and ATG20, the induction machinery is present. These are members of scaffold complexes that are involved in cytoplasm-to-vacuole (Cvt) protein targeting, or selective, non-starvation induced autophagy, which also include ATG13 and ATG1. ATG13 is conserved in *N. fowleri*, but could not be identified either in the predicted proteins or scaffolds of *N. gruberi*, suggesting that has been lost. The Cvt pathway genes ATG11 and ATG19 are basically restricted to *Saccharomyces*,[473] therefore their absence in *Naegleria* spp. is unsurprising. Together, these results suggest that some aspects of the Cvt pathway in yeast are lineage-specific, while others are more widely conserved, but have been lost in *Naegleria* spp.

Figure 6.10. Comparative genomic survey of autophagy machinery in *N. fowleri* and *N. gruberi*.

Autophagy machinery can be divided into machinery involved in autophagosome induction, cargo packaging, vesicle nucleation, vesicle expansion and completion, component retrieval, vesicle breakdown, and other related autophagy factors. Presence is indicated by a dot, and numbers indicate the number of paralogues identified. A grey dot indicates a putative homologue.

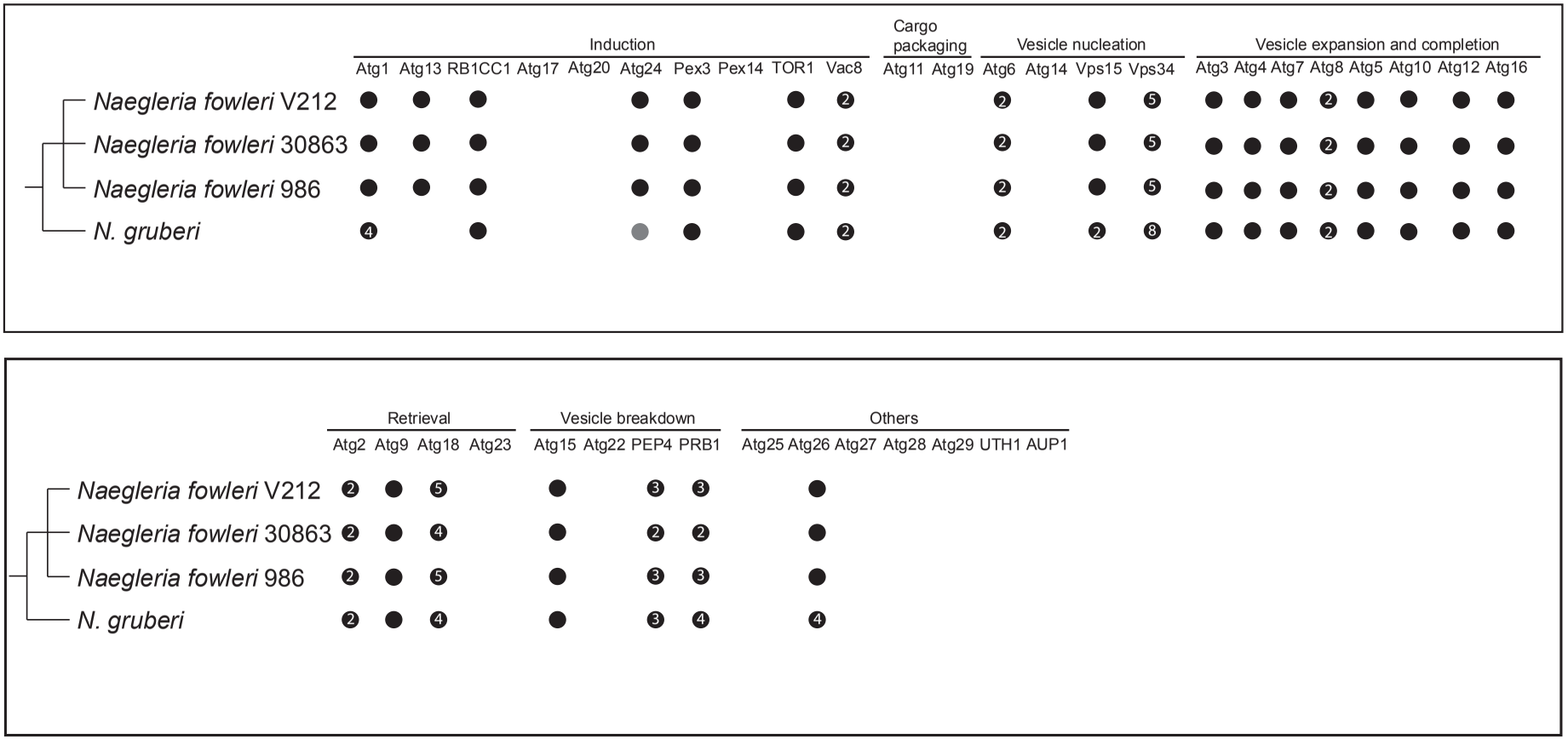

Induction
Cargo packaging
Vesicle nucleation
Vesicle expansion and completion
Atg1 Atg13 RB1CC1 Atg17 Atg20 Atg24 Pex3 Pex14 TOR1 Vac8
Atg11 Atg19
Atg6 Atg14 Vps15 Vps34
Atg3 Atg4 Atg7 Atg8 Atg5 Atg10 Atg12 Atg16
Naegleria fowleri V212
Naegleria fowleri 30863
Naegleria fowleri 986
N. gruberi
Retrieval
Vesicle breakdown
Others
Atg2 Atg9 Atg18 Atg23
Atg15 Atg22 PEP4 PRB1
Atg25 Atg26 Atg27 Atg28 Atg29 UTH1 AUP1
Naegleria fowleri V212
Naegleria fowleri 30863
Naegleria fowleri 986
N. gruberi

Figure 6.10

Pexophagy is a specific type of autophagy in which the peroxisome is selectively degraded. The existence of peroxisomes in *Naegleria* has been inferred from the presence of PEX genes, and genes that encode the metabolic enzymes of peroxisomes.[610] However, losses of certain machineries suggest that peroxisomes could be modified in *Naegleria*. This would be in line with the partial loss of the Dsl1 tethering complex, shown above. It is not surprising that some pexophagy machinery is present (ATG24 and PEX3). Both species are missing Pex14, which is also involved in matrix protein import;[628] however, this gene is also not identified in the peroxisome-containing Stramenopile *Phaeodactylum tricornutum*.[629] Identifying pexophagy machinery supports in *Naegleria* spp. supports the presence of peroxisomes in this genus.

The vesicle nucleation, expansion and completion machinery, protein retrieval, and vesicle breakdown machinery is largely present, and in some cases expanded in both *Naegleria* spp. Gene duplications have occurred in ATG6/Beclin, Vps34, ATG8, ATG2, ATG18, PEP4, and PRB1. The functional relevance of these duplications is unclear, but it suggests some additional complexity in the autophagy system in both *Naegleria* species, particularly in duplicated genes that encode interacting proteins (ATG6 and Vps34, ATG18 and ATG2). There are few examples of genes where *N. fowleri* and *N. gruberi* have large differences in paralogue number. These include ATG1 and ATG26, with four paralogues each in *N. gruberi*, compared to one sequence in the three *N. fowleri* strains. ATG1 is a scaffold protein that functions in the induction of autophagy;[630] this may allow for differences in complex assembly or induction efficiency in *N. gruberi*. ATG26 is involved in the synthesis of sterol glucoside membrane lipids,[631] and is involved in autophagy in the fungi *Pichia pastoris*, so the relevance of these duplications specifically to autophagy in the unrelated *N. gruberi* is unclear.

Based on these comparative genomic results, it appears that autophagy in *Naegleria* spp. occurs with much of the same machinery that is functionally characterized in human and yeast cells. Furthermore, there is no evidence of changes to the autophagy machinery complement in *N. fowleri* that seem to be related to pathogenesis; on the contrary, *N. fowleri* and *N. gruberi* encode a similar repertoire of autophagy genes. The only exceptions are a potential loss of ATG13 in *N. gruberi* (which may be a false negative), and the expansions of ATG1 and ATG26 that appear to be *N. gruberi*-specific.

6.4.3 ER-Associated Degradation machinery and Unfolded protein response machinery

One potential argument for *N. fowleri*'s ability to infect humans and animals is the ability to survive the stresses associated with infection, as the host environment of the nasal tissue and brain are vastly different from water and soil. Stresses could include changes in oxygen levels, temperature, and pH, as well as assault from immune cells. Potential differences in the UPR and ERAD machinery encoded in the genomes of *N. fowleri* and *N. gruberi* may help to explain pathogenicity in *N. fowleri*.

In the ER, N-glycosylation controls the nascent protein's interaction with chaperones protein disulphide isomerase (PDI), calnexin, and calreticulin. Refolding of an improperly folded protein is induced by the addition of a glucose moiety by UDP-glucose:glycoprotein glucosyltransferase (UGGT1).[632] If the protein cannot refold, it interacts with ER-associated Degradation (ERAD) machinery OS-9 and XTP3-B, as well as ER hsp70 (BiP) and hsp90 (Grp94).[633] This targets the misfolded protein for ubiquitylation by the Sel1L-Hrd1 complex, retro-translocation to the cytosol, and proteasome degradation. While no comprehensive comparative genomic analysis of ERAD machinery has been done in eukaryotes, there are at

least two studies of ERAD machinery in protists. *Plasmodium falciparum*, the apicomplexan parasite that causes malaria, has a duplicated ERAD system that is targeted to the apicoplast – a red algal secondary endosymbiotic plastid – to aid in protein transport.[634] This is also the case in diatoms, where ERAD machinery targets preproteins across the second outermost plastid membrane.[635]

Many components of the human and yeast ERAD machinery could be identified in *Naegleria* (Figure 6.11, Online Appendix Table 6.5). The only unidentifiable component is Usa1, a scaffold for the Hrd-ubiquitin ligase,[636] which does not appear to be present outside of the Opisthokonta. In both organisms, there have been gene duplication events, including the ubiquitination proteins Uba1, Ubc7, Doa10; Yos9, the lectin that detects misfolded proteins; Derlin-like proteins, which are involved in degrading soluble substrates; retrotranslocation protein Ufd1; and Png1, which functions in deglycosylation.[637–641] However, there are few paralogue differences between *N. gruberi* and the three *N. fowleri* strains, suggesting that any additional complexity within the ERAD system is common to both *Naegleria* species.

When misfolded proteins overwhelm the ERAD pathway, the UPR is activated to prevent cellular damage caused by protein aggregation. The UPR has been most extensively studied in human cells, in which three signaling pathways have been identified: IRE1, PERK, and ATF6. Spycher and colleagues (2013)[642] performed comparative genomics to identify homologues of the major proteins involved in each of the three known stress response pathways in a diversity of eukaryotes. They found that approximately one half of the factors involved in the mammalian stress responses are distributed in eukaryotes in a pattern that would suggest their presence in the LECA, while others appear to be taxon-specific.[642] For example, IRE1 is found across eukaryotic

Figure 6.11. Comparative genomic survey of ER-associated degradation machinery in *N. fowleri* and *N. gruberi*.

ERAD machinery can be classified into the following categories: machinery supporting normal ER function and misfolded protein detection, machinery involved in ubiquitination, and machinery involved in protein retrotranslocation out of the ER and protein degradation. Presence is indicated by a dot, and numbers indicate the number of paralogues identified. Plus signs in the Hsp70 and Hsp110 dots indicate that multiple sequences were identified that are orthologous to heat shock proteins, but could not be confidently annotated as Hsp70 or Hsp110.

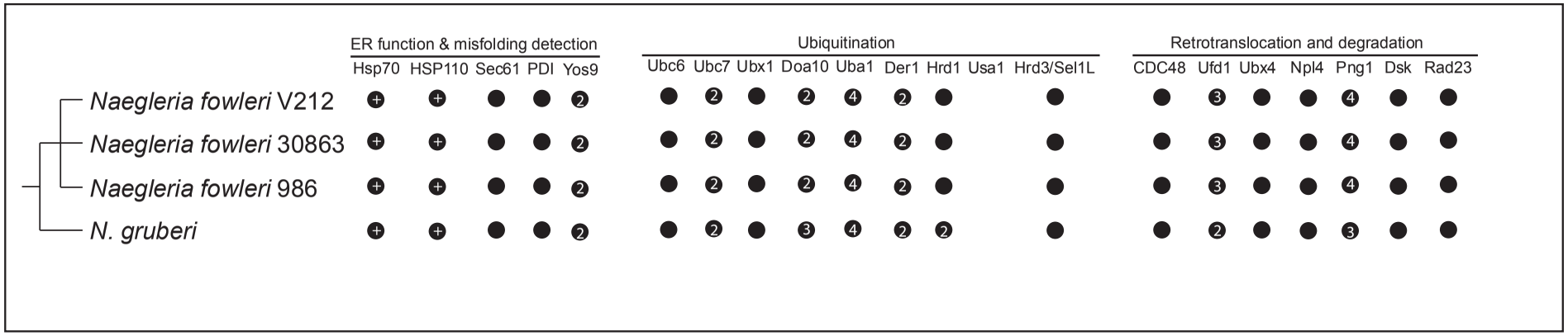

ER function & misfolding detection
Hsp70 HSP110 Sec61 PDI Yos9
Ubiquitination
Ubc6 Ubc7 Ubx1 Doa10 Uba1 Der1 Hrd1 Usa1 Hrd3/Sel1L
Retrotranslocation and degradation
CDC48 Ufd1 Ubx4 Npl4 Png1 Dsk Rad23
Naegleria fowleri V212
Naegleria fowleri 30863
Naegleria fowleri 986
N. gruberi

Figure 6.11

taxa, but its effector transcription factors are lineage-specific. PERK and PERK-like proteins are also widespread, and the latter appear to be functional in *Leishmania*[643] and *Toxoplasma gondii*.[644] The ATF6 transcription factor is again restricted to metazoa, while the S1P and S2P proteases are found across eukaryotes, although they are responsible for cleaving non-UPR related transcription factors.[645] Spycher *et al*. (2013) also showed that *N. gruberi* encodes IRE1, a potential Xbp1 transcription factor, several PERK pathway proteins including PERK-like, and site-1 protease.

Similar results were found for the UPR machinery in *N. fowleri* as compared with *N. gruberi* (Figure 6.12, Online Appendix Table 6.5). While IRE1 is present across eukaryotes, the transcription factors it activates appear to be lineage-specific. bZIP domain-containing transcription factors are known to be involved in the IRE1 pathway in human cells, some fungi, and *A. thaliana*,[646–649] therefore bZIP domain proteins identified in *Naegleria* are included in the Coulson plot (Figure 6.12). Parts of the PERK pathway are generally conserved, such as PERK-related proteins, EIF2A, and DNAJC3. These are upstream of the effectors ATF4 and GADD34, which are highly restricted, so it is possible that other lineage-specific effectors exist in non-metazoan eukaryotes. In the ATF6 pathway, S1P and S2P cleave ATF6 in the Golgi, from where it translocates to the nucleus as an active transcription factor.[650] Again, as S1P and S2P are conserved in *Naegleria* and other eukaryotes while ATF6 is animal-specific, it suggests that other organisms might employ one or more other transcription factors in this pathway. Of course, it is also possible that these transcription factor activators have functions outside of the stress response in non-metazoans. Because paralogue numbers were not shown in Spycher *et al*. (2013), only presence and absence of UPR factors is shown in Figure 6.12. However, only one homologue was identified in *N. fowleri* and *N. gruberi* for each UPR component, with the

Figure 6.12. Comparative genomic survey of machinery involved in the unfolded protein response in *N. fowleri*.

The IRE1 pathway, PERK pathway, and ATF6 pathways are three conserved arms of the UPR. Asterisks indicate putative subunits identified in *Aspergillus nidulans*. Figure modified from Spycher *et al.* 2013;[642] all organism data other that *N. fowleri* is published there.

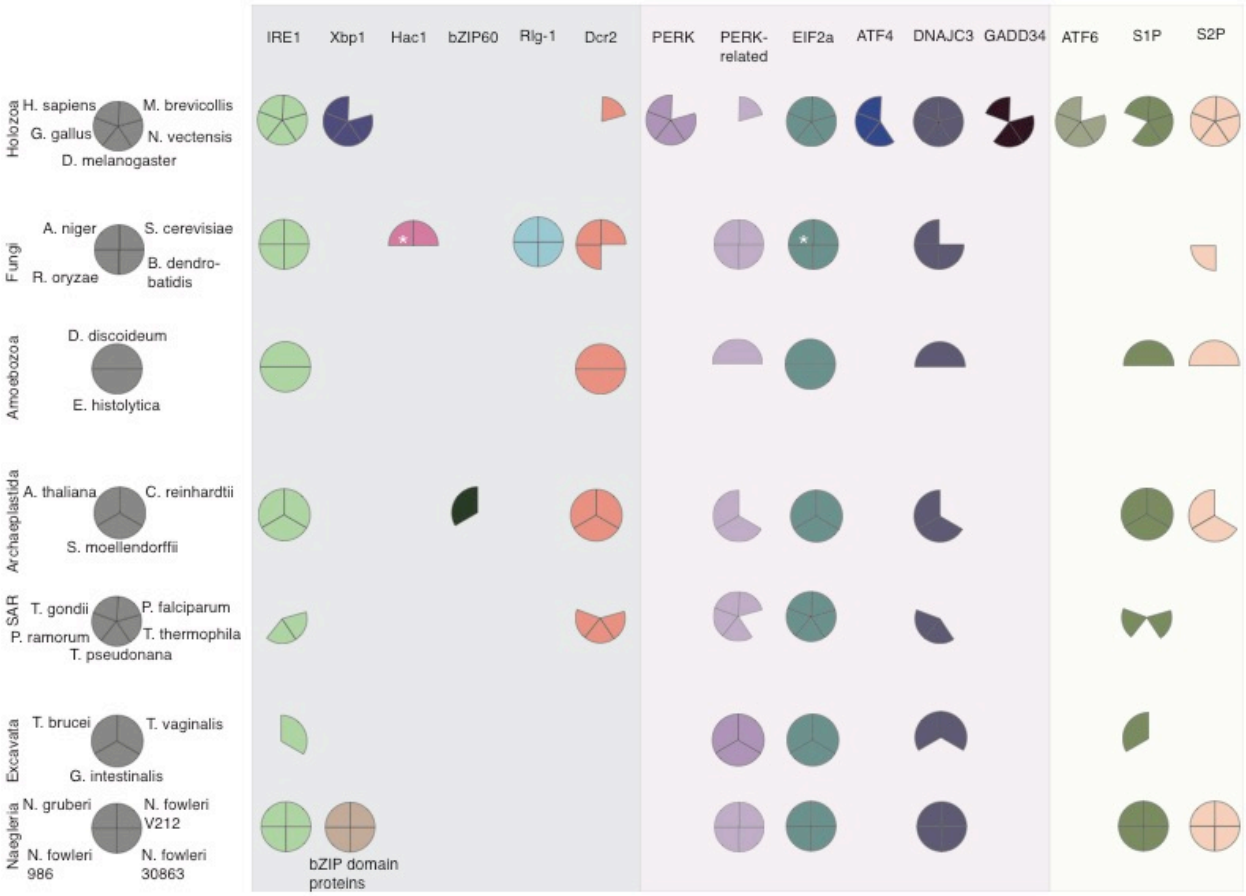

IRE1
Xbp1
Hac1
bZIP60
Rlg-1
Dcr2
PERK
PERK-related
EIF2a
ATF4
DNAJC3
GADD34
ATF6
S1P
S2P
Holozoa
H. sapiens
M. brevicollis
G. gallus
N. vectensis
D. melanogaster
Fungi
A. niger
S. cerevisiae
R. oryzae
B. dendro-batidis
Amoebozoa
D. discoideum
E. histolytica
Archaeplastida
A. thaliana
C. reinhardtii
S. moellendorffii
SAR
T. gondii
P. falciparum
P. ramorum
T. thermophila
T. pseudonana
Excavata
T. brucei
T. vaginalis
G. intestinalis
Naegleria
N. gruberi
N. fowleri V212
N. fowleri 986
N. fowleri 30863
bZIP domain proteins

Figure 6.12

exceptions of bZIP domain containing proteins and PERK-like proteins. Three bZIP domain-containing proteins were identified in all three *N. fowleri* strains, while only two were found in *N. gruberi*. Both *N. fowleri* and *N. gruberi* have three PERK-like proteins. As with the ERAD machinery, there are no obvious differences between the *N. fowleri* and *N. gruberi* UPR that would suggest a differential ability to cope with cell stress. However, our minimal of knowledge of the UPR outside of the opisthokonts and archaeplastids makes this assessment incomplete.

6.4.4 Adhesion and cell-cell interaction factors

Since infection requires the ability to attach to and crawl along cells of the nasal epithelium, differences between *N. fowleri* and *N. gruberi* in cell-cell adhesion factors may be relevant to pathogenesis. Jamerson and colleagues (2012)[651] have shown that an integrin-like adhesion protein localized to focal adhesions is expressed at higher levels in *N. fowleri* compared to the non-pathogenic *N. lovaniensis*. This protein was identified using a polyclonal antibody to a human β-1 integrin subunit. However, initial BLAST searches using human integrin queries did not retrieve any clear orthologues, and other work has shown that integrin-mediated adhesion likely evolved in the ancestor of the Opisthokonta (animals and fungi).[652] It is therefore likely that the anti-β-integrin antibody recognized another surface protein, although it may well be involved in adhesion.

Other adhesion and cell migration proteins in animals are the adhesion G protein-coupled receptors. They have an extremely large extracellular region that interacts with other cells and the extracellular matrix, which is attached to a seven pass transmembrane region. Binding of the extracellular region to a ligand induces a change in receptor conformation, causing activation of a G protein and downstream intracellular signaling. AGPCRs are known to be important in

tumor metastasis,[653] and importantly, are not restricted to multicellular animals. The sequences that are retrieved by using the human AGPCR protein sequences to search the *N. fowleri* and *N. gruberi* genomes are repeat containing proteins; these include tenascin-like proteins, leucine rich repeat proteins, NHL repeat proteins, epidermal growth factor domain proteins, von Willebrand domain proteins, and some G-protein coupled receptors (listed in Online Appendix Table 6.6). Proteins with these domains, particularly tenascins or EGF family proteins, are known virulence factors *Giardia intestinalis,*[654] and may be involved in mediating adhesion to the host gut. While their function is unclear in *N. fowleri,* this list is a starting point against which antibodies can be generated for further functional characterization. Interestingly, ~225 were identified in *N. gruberi*, while only ~50 were found in *N. fowleri*. As the function of these is unknown, it is not clear why *N. gruberi* has expanded gene families with repeat domains, but it does partly explain the difference in gene content between these two organisms.

Because of the large evolutionary distance between *Naegleria* and *H. sapiens*, this approach may ignore the potential diversity of AGPCRs in *Naegleria* that are unrelated to those in human cells. To identify more AGPCRs based on protein structure and domain organization, *Naegleria* sequences with the correct organization of transmembrane domains and large, repeat-containing extracellular regions were identified. The CBS TMHMM server was used to predict transmembrane regions in all four *Naegleria* datasets. Then, proteins with transmembrane regions were passed to the CDD domain prediction server. Table 6.3 shows all *Naegleria* sequences that have both an extracellular domain (assuming correctly predicted topology in the membrane) and at least one transmembrane region. Seven sequences were identified in *N. fowleri* that are predicted to be GPCRs, while ten were found in *N. gruberi*, based on the

Table 6.3. Putative AGPCRs: Sequences identified with ~7 transmembrane domains, with G protein-coupled receptor (GPCR) and receptor for G signaling (RGS) domains in *N. fowleri* V212 and *N. gruberi*

N. fowleri	Protein ID	Domains	Number of transmembrane domains
	g3853	lactonase, GPCR, NHL, RGS, ATPase	7TM
	g10229	GPCR, RGS	7TM
	g8432	EGF, GPCR, RGS	7TM
	g9651	KELCH, EGF, fibronectin, GPCR, RGC	8TM
	g8795	GPCR, EGF, ephrin-like receptor and protein kinase domain	3TM *
	g4693	GPCR, LRR, RGS	7TM
	g6748	EGF, GPCR, RGS	7TM
	g12446	GPCR, RGS, oleosin	8TM

* May not be adhesion GPCR, but ephrin-like receptor domain suggests signalling relevant to migration

N. gruberi	Protein ID	Domains	Number of transmembrane domains
	80427	GPCR, RGS, NHL	7TM
	63528	NHL, RGS, GPCR	7TM
	72726	NHL, GPCR, RGS, SMP30/Gluconolactonase	7TM
	49726	EGF, GPCR, RGS	8TM
	52385	EGF, GPCR, RGS, claudin domain	6TM
	75738	EGF, RGS, GPCR	7TM
	67572	NHL, RGS, SMP30/Gluconolactonase, GPCR, esterase-like activity of phytase	7TM
	64936	NHL, GPCR, RGS	7TM
	74387	GPCR, RGS, EGF	8TM
	65586	GPCR, RGS, alpha beta hydrolase	15TM**

** Potential modified adhesion GPCR

presence of ~7 transmembrane domains and, at the least, a Regulator of G protein Signalling (RGS) domain and a G Protein Coupled Receptor proteolysis site (GPS). The extracellular domains of these proteins include NHL (NCL-1, HT2A and Lin-41) repeats, the Epidermal Growth Factor domain, and leucine-rich repeats, all known to be involved in adhesion.[655] Interestingly, an additional sequence was identified with only three transmembrane regions in both *Naegleria* spp.; however it has a domain that resembles the ephrin receptor in human cells, and the *Giardia* variant-specific surface protein. Ephrin and ephrin receptors are expressed during neuronal development, but work in non-human primates suggests that these proteins are expressed in the adult brain as well.[656] While it is tempting to speculate that this protein is able to bind a protein found on neurons, it is more likely to be a general adhesion factor, given the large evolutionary distance between *Naegleria* and humans.

TM9/Phg1, SadA, SibA, and SibC proteins have been identified in the amoeba *Dictyostelium discoideum* as involved in adhesion.[657] TM9 and SadA are regulators of SibA, in that they control cell surface expression, abundance of SibA transcripts, intracellular transport, and protein stability. They may also be involved in adhesion in yeast and *Drosophila* phagocytic cells.[658] To determine if these proteins are also present in *Naegleria*, the *D. discoideum* sequences were used as BLAST queries. In both *N. fowleri* and *N. gruberi*, only orthologues of the TM9 protein could be reliably identified (Online Appendix Table 6.7). It is possible that the *Naegleria* version may be involved in cell-cell adhesion, but with other downstream effectors. Cell biological work in *Naegleria* is required to determine if this protein is indeed an adhesion molecule.

6.5 Differential expression of genes in high pathogenicity versus normal pathogenicity *N.*

fowleri

While comparative genomics is a useful way to gain a broad view of gene family or system-wide differences between different organisms, differential expression analyses allow us to understand cellular dynamics under a particular condition. In this case, we are interested in the *N. fowleri* genes that are involved in pathogenesis. To identify the genes whose expression changes as a consequence of host infection, a comparative transcriptomic analysis was performed on 'high pathogenicity' and 'regular pathogenicity' *N. fowleri* LEE strain. These strains were developed by Whiteman and Marciano-Cabral (1987), who showed that *N. fowleri* LEE passaged through 50 mice (high pathogenicity) has a lower LD_{50} in guinea pigs than *N. fowleri* LEE grown in axenic culture (normal pathogenicity).[607] High pathogenicity *N. fowleri* LEE has been continuously passaged through mice, and then grown in axenic culture to remove brain tissue, after which mRNA is isolated and sequenced. Normal pathogenicity *N. fowleri* LEE is only grown in axenic culture. Therefore, genes differentially expressed in highly pathogenic *N. fowleri* can be pathogenicity factors, but may also be genes that are more tenuously related, for example if the systems they work in are modulated by the environmental – and subsequent cell biological – changes associated with infection. Additionally, there is a chance that briefly growing the mouse-passaged samples in culture to reduce the amount of human cell DNA reduces the expression of pathogenicity factors, potentially giving false negatives. Further transcriptomic work on *N. fowleri* must be performed to determine the influence of this particular experimental condition on gene expression.

Initially, this experiment was attempted using the V212 strain rather than the LEE strain, and passaged through two, four, and six mice. However, no meaningful differential expression data was generated using this approach, even when comparing the axenically cultured sample

with one extracted after six mouse passages. There were likely two reasons for this; first, mouse LD_{50} data suggested that the V212 strain passaged through six mice was less virulent than the mouse-passaged LEE strain and very similar to the V212 sample grown axenically, making differences between mouse-passaged and axenic samples more difficult to detect, and second, biological replicates were not used, seriously limiting the statistical power of differential expression tests. Therefore, mRNA from three biological replicates each of mouse-passaged LEE (LEE-MP) and axenically cultured LEE (LEE-Ax) were sequenced. This is the first analysis of global transcriptomic expression associated with pathogenicity, and will give insight into both pathogenesis and the underlying cell biology of host invasion.

6.5.1 Transcriptomic analysis of *N. fowleri* LEE-Ax and *N. fowleri* LEE-MP

mRNA was extracted from three cultures of *N. fowleri* LEE-Ax (axenic, regular pathogenicity), grown axenically in culture. *N. fowleri* LEE-MP (mouse-passaged, high pathogenicity) was inoculated into three mice, and following sacrifice, the amoebae were extracted and grown in culture briefly to remove mouse brain tissue, after which mRNA was extracted from each of the three cultures. Illumina MiSeq sequencing produced ~2-3 million reads, which were mapped to the genome and predicted proteome of *N. fowleri* V212. Although genome organization is highly variable between strains, nucleotide identity is similar enough that reads from the LEE strain map to V212 sequence relatively well, with ~60% of reads mapping prior to filtering of mitochondrial and extrachromosomal plasmid reads (see Supplementary Table ST6.2 for transcriptomics statistics). Additionally, transcripts were generated from unmapped reads, and when present, were added to the dataset for differential expression (DE) analysis.

Inter-sample comparisons revealed that one of the LEE-MP replicates was highly dissimilar to both the other LEE-Ax replicates and the LEE-MP replicates, and was therefore discarded from DE analyses (Figure 6.13). DE analysis was performed with EdgeR through the Trinity software package, and the general results are visualized in MA plots and volcano plots Figure 6.14.

315 genes were identified as differentially expressed, using a false discovery rate maximum of 0.1. The false discovery rate is a metric to control for type I errors when making many comparisons (in this case, thousands; one comparison for every expressed gene). With a false discovery rate of 0.1, 10% of the DE genes are not actually differentially expressed. Additionally, for all DE genes, the P-value for each comparison is several orders of magnitude lower than the false discovery rate; none is on the verge of statistical significance (Online Appendix Table 6.7).

6.5.2 Up-regulated genes associated with pathogenesis

208 genes are up-regulated in high pathogenicity versus normal pathogenicity *N. fowleri*. Several large categories of DE genes are immediately obvious (Figure 6.15) and each is discussed below.

Nearly 15% of all up-regulated genes are involved in lysosomal processes. Twenty-two of these are lysosomal proteases, nine of which are members of the cathepsin B subfamily; these are known pathogenicity factors in both *N. fowleri* and other.[354,654,659–662] Chapter 6.5.6.1 contains an expanded analysis of the proteases in *Naegleria* spp. In addition to peptidases, a lysosomal rRNA degradation gene is up-regulated, as well as three vacuolar ATPase proton pump subunits (116 kDa, 21 kDa, and 16 kDa). The vacuolar ATPase is responsible for

Figure 6.13. Sample correlation matrix heatmap of LEE-Ax and LEE-MP samples.

Normalized differential expression data for each biological replicate are compared. Inset indicates Pearson correlation coefficient, visualized by colour. Red indicates more similar datasets, while green indicates less similar datasets. Cladograms represent sample relationships based on similarity. Yellow outlining boxes indicate the two sample groups to be compared, LEE-Ax and LEE-MP. The MP2 sample was discarded in further differential expression analyses.

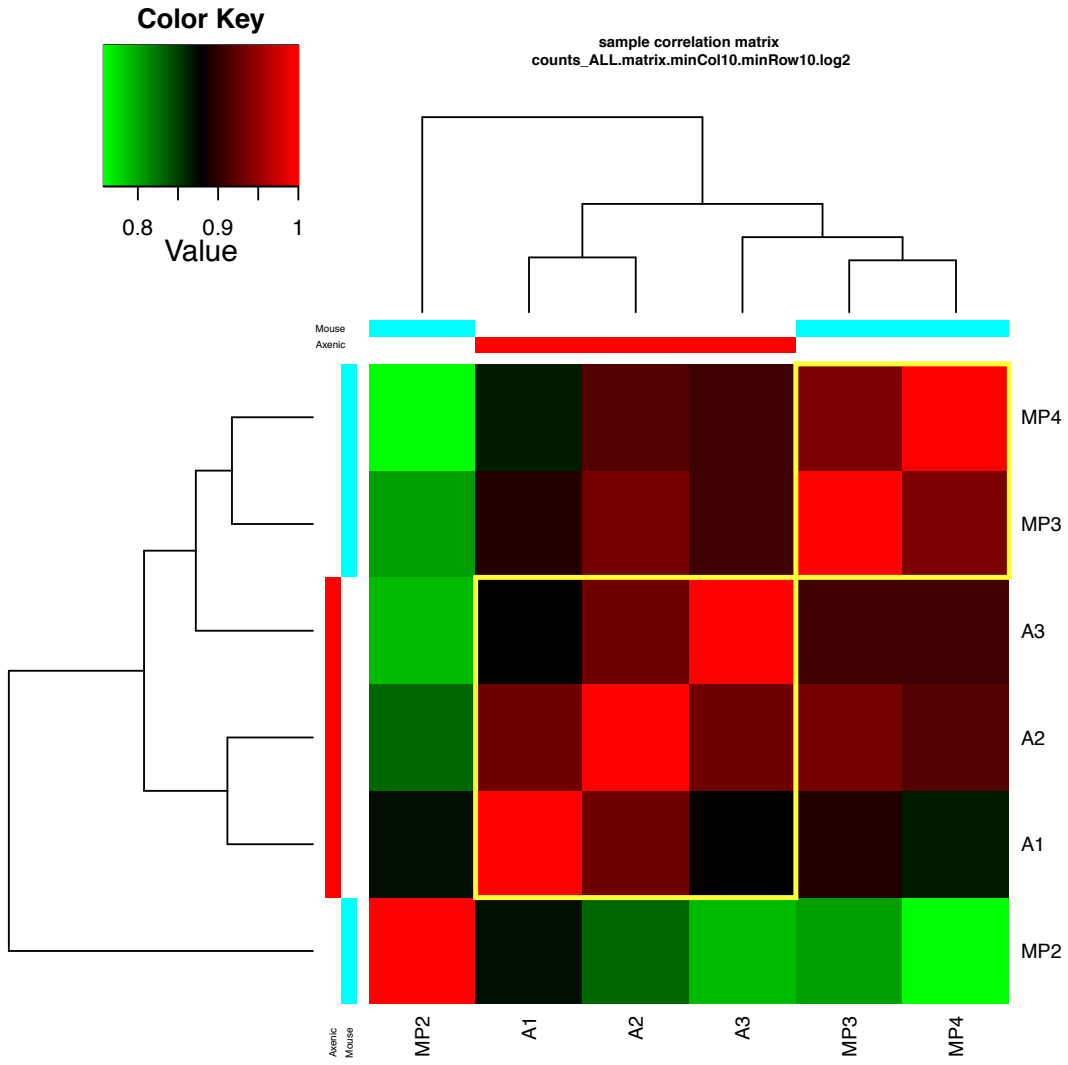

Color Key
0.8
0.9
1
Value
sample correlation matrix
counts_ALL.matrix.minCol10.minRow10.log2
Mouse
Axenic
MP4
MP3
A3
A2
A1
MP2
Axenic
Mouse
MP2
A1
A2
A3
MP3
MP4

Figure 6.13

Figure 6.14. MA and Volcano plots of differentially expressed genes in LEE-MP versus LEE-Ax samples.

Data points represent individual genes. (Left) Gene expression log fold change is plotted by read count in MA plots. (Right) The log of the false discovery rate for each differentially expressed gene is plotted by the log fold change in volcano plots. Red dots indicate genes with a false discovery rate < 0.05.

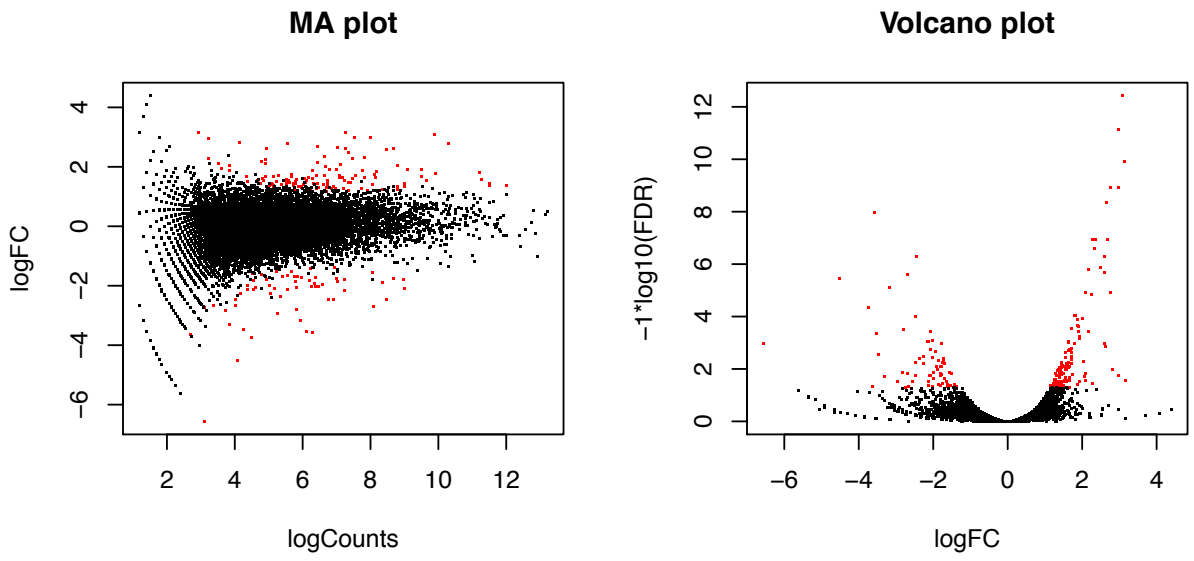
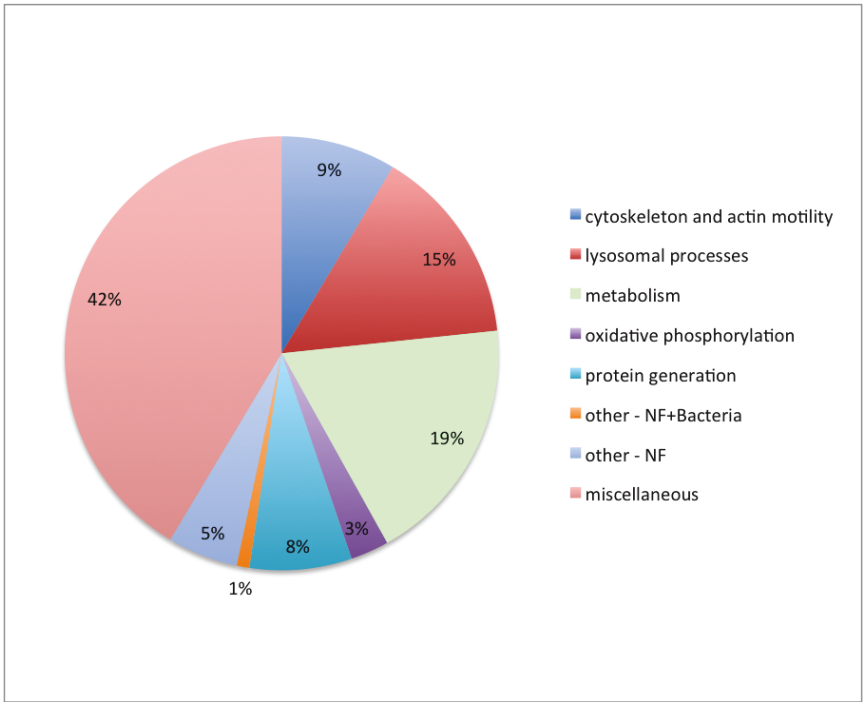


Figure 6.14

Figure 6.15. Categories of up-regulated and down-regulated genes in mouse-passaged *N. fowleri* LEE versus axenically grown *N. fowleri* LEE.

Functional categories of genes that are up-regulated in highly pathogenic *N. fowleri* (top) or down-regulated in highly pathogenic *N. fowleri* (bottom). The categories other – NF+Bacteria include sequences found in *N. fowleri* and bacteria, while other – NF includes *N. fowleri*-specific sequences.

Functional categories of up-regulated genes



Functional categories of down-regulated genes

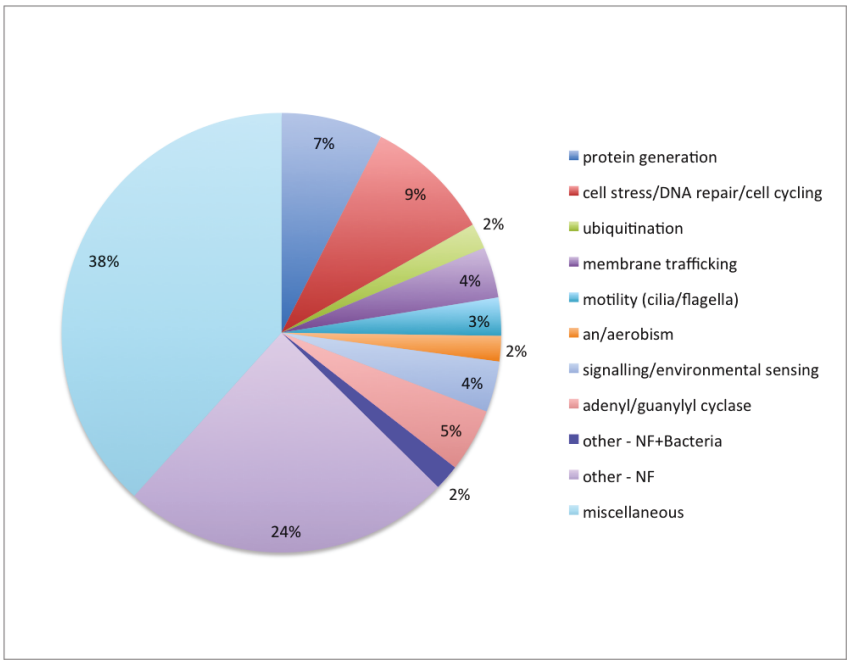


Figure 6.15

acidification of both lysosomes and secretory vesicles,[125] and is made up of the V1 subcomplex involved in ATP hydrolysis, and the membrane-spanning V0 subcomplex that functions in proton translocation (driven by V1 function). The up-regulated subunits are the a, c'', and c subunits, respectively. These are all members of the V0 complex; the membrane-spanning portion of the ATPase involved in proton translocation.[663] In opisthokonts, the V0 and V1 subcomplexes dissociate and re-associate based on nutrient availability.[664,665] Up-regulation of these V0 subunits strongly suggests that more proton pump subcomplexes are produced in response to mouse passage.

Endo-lysosomal trafficking genes are also up-regulated, including the Rab GTPase Rab32 and the retromer component Vps35. Although retromer is primarily involved in TGN-endosome recycling, work in *Drosophila* has shown that in Vps35 knockdown cells, Cathepsin L (a Cathepsin B family lysosomal enzyme) does not colocalize with the lysosomal marker Lamp1.[666] This raises the possibility that the retromer complex may play a role in lysosomal enzyme trafficking. Additionally, the strumpellin subunit of the WASH complex is up-regulated; WASH and retromer interact,[667] and the WASH complex is involved in lysosome maintenance and vacuolar ATPase recycling (a part of vacuolar maturation) in addition to its role in branched actin filament formation.[668]

19% of up-regulated genes are involved in metabolism. Both catabolic and anabolic processes of lipid metabolism are represented: phospholipase B-like genes, genes involved in beta-oxidation, phosphatidate/phosphatidylethanolamine synthesis, fatty acid synthesis, long chain fatty acid elongation, and sterol biosynthesis. Interestingly, one up-regulated gene is squalene synthase, which catalyzes one of the first steps in sterol biosynthesis. In trypanosomatid parasites, the sterol ergosterol is an essential membrane component, and as such, squalene

synthase has been explored as a drug target in treatment of Chagas disease.[669] Ergosterol or related sterols may be membrane components in *Naegleria,* and potentially involved in stabilizing cell membranes in *N. fowleri* as a result of the shift to a relatively higher temperature of the infected host. Sphingosine, ceramide, and inositol metabolism genes are also up-regulated, as are eight genes involved in amino acid metabolism. Mitochondrial and energy production genes are up-regulated, such as ubiquinone biosynthesis genes, cytochrome p450, isocitrate dehydrogenase (tricarboxylic acid cycle), Complex I and Complex III genes (oxidative phosphorylation), and a mitochondrial ADP/ATP translocase. The pattern of up-regulated metabolism and energy production genes suggests that *N. fowleri* is thriving during infection.

Eleven actin motility and adhesion genes are up-regulated in high pathogenicity *N. fowleri*. In addition to the WASH complex subunit strumpellin mentioned above, other actin-related motility genes include Arp3, a flotillin domain-containing protein, twinfillin, and a gelsolin-like protein. There is also an up-regulated mucin 4-like gene that may be involved in cellular protection or cell-cell interaction,[670] and RhoGAP22 and the serine/threonine protein kinase PAK3 that are involved in RAC1-induced cell migration.[671,672] Additionally, two genes with RGS (regulator of G protein signaling) domains are up-regulated. Although they do not appear to be adhesion GPCRs, they may be involved in environment sensing and cell signaling.

Sixteen transcription, translation, and protein modification genes are up-regulated in high pathogenicity *N. fowleri* (~8% of up-regulated genes). These include three transcription factors, three proteins that are either part of the ribosome or are involved in ribosome biogenesis, and two elongation factor 1-alpha genes. Four genes are involved in proper protein folding, including the chaperone DNAK and several glycosylation genes. One curiously up-regulated gene is the protein-lysine methyltransferase METTL21D, which methylates the protein p97/CDC48, an

ATPase with numerous cellular functions.[316] Methylation of p97 by METTL21D has been shown to negatively impact its ATPase activity,[673] suggesting a decrease in p97 function in highly pathogenic *N. fowleri*.

There are many other genes that are up-regulated, but do not fall into one of the categories outlined above. These include four genes that function in the mitochondria, specifically as peptidases and in respiratory chain assembly. Three other genes are involved in mitosis and cell cycle regulation (MOB1, cyclin dependent kinase inhibitor 3, Nek2). There are also several genes involved in cell signaling, in intracellular trafficking (dynein light chain), and the autophagy gene ATG18. As discussed in previous sections on autophagy evolution, ATG18 plays a role in both pre-autophagosomal structure (PAS) formation and vacuolar maintenance in yeast.[674] Its increased expression in highly pathogenic *N. fowleri* therefore either provides further evidence of heightened lysosomal/autophagosomal function.

Finally, there are 62 genes that could not be reliably annotated, or are specific to either *N. fowleri* or both *Naegleria* spp. Although their functions are unknown, they may be relevant to pathogenicity, especially those that are specific to *N. fowleri*. However, only 7 genes appear to be *N. fowleri*-specific, with no clear homologues in any other organism. These represent unique potential targets against which anti-*Naegleria* therapeutics may be developed.

6.5.2.1 Expression of prosaposins and their evolutionary history

Saposins are degradative pore-forming proteins found in many parasites. Two prosaposin genes were identified in the up-regulated gene set. Previous work by Herbst et al. (2004)[358] identified one of these genes as Naegleriapore A, a heavily glycosylated and protease-resistant

pore-forming protein. Saposins, or sphingolipid activator proteins, are known to function both in lysosomes and on the cell surface in a degradative manner towards eukaryotic and bacterial cells.[675,676] Saposins of liver flukes (*Fasciola hepatica* and *Clonorchis sinensis*), hookworms (*Ancylostoma duodenale* and *Necator americanus*), and the gut pathogen *Entamoeba histolytica* are extremely hemolytic,[677–679] and play a key role in host cell digestion.[680] In order to better understand their evolutionary history, saposins were searched for in a diversity of eukaryotes using the human sequence as a BLAST query. Saposin homologues were identified in Animals (but lost in Fungi), in the Amoebozoa, in the Archaeplastida, and in the Excavata, also suggesting that they were lost at the base of the SAR clade (Online Appendix Table 6.8).

Although saposins appear to be ancient and present in the LECA, it is possible that one or more saposin homologues in *N. fowleri* are the result of lateral gene transfer with a parasite such as *Entamoeba*. To test this, phylogenetics was used to determine the evolutionary relationships between the identified saposin sequences. The two up-regulated prosaposin proteins in *N. fowleri* have clear orthologues in *N. gruberi* (NfSap1 and NfSap2), and one of these is related to a prosaposin orthologue in the amoebozoan *Filamoeba nolandi* (Figure 6.16). *F. nolandi* is a non-pathogenic freshwater and soil amoeba.[681] The second prosaposin that is up-regulated is not closely related to any sequence other than an orthologue in *N. gruberi*. Together, these results suggest that the ability of saposin proteins to be involved in pathogenicity in *N. fowleri* is not specifically related to virulence in any other eukaryote.

Homology searching and phylogenetic analysis of prosaposins also identified a well-supported clade of saposin-like sequences, with representatives in the opisthokonts, Amoebozoa, and Excavata (including *Naegleria* spp.). This clade is likely ancient and potentially present in the LECA, but lost from either all archaeplastids, or at least the archaeplastids sampled here.

Figure 6.16. Phylogeny of eukaryotic saposin domain-containing proteins.

Includes pro-saposin sequences, saposin-like sequences, and the saposin domains of plant phytepsins. Node values are listed as MrBAYES/RAxML (posterior probability/bootstrap) and as symbols indicating a minimum level of support as shown in the inset. Node values are shown on the best Bayesian topology. Phytepsin sequences appear to branch within a clade of amoebozoan saposin sequences; this larger clade is boxed. Significantly up-regulated *N. fowleri* saposins are marked with an asterisk. Tree is rooted arbitrarily on a clade of saposin-like proteins.

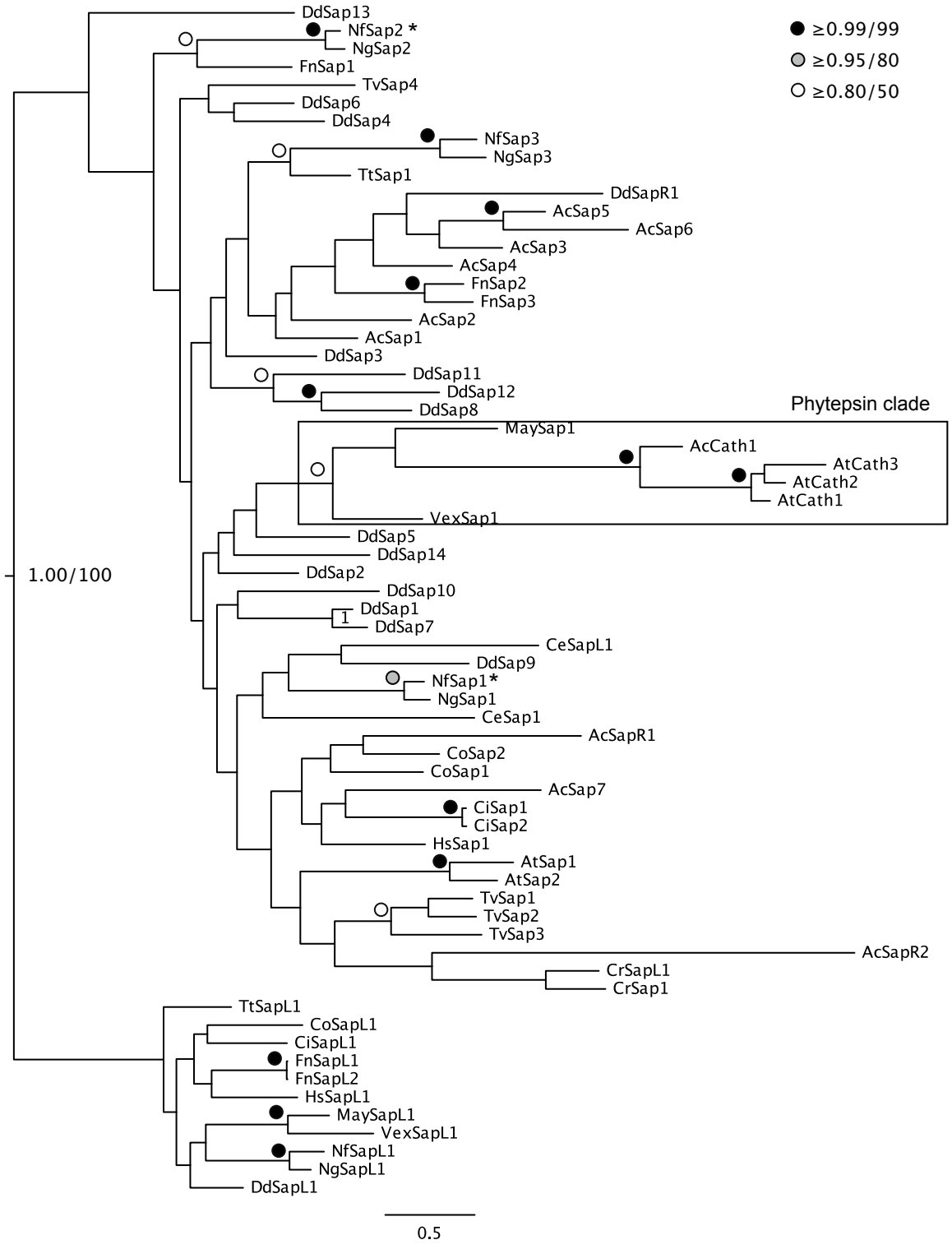

DdSap13
NfSap2 *
NgSap2
FnSap1
TvSap4
DdSap6
DdSap4
NfSap3
NgSap3
TtSap1
DdSapR1
AcSap5
AcSap6
AcSap3
AcSap4
FnSap2
FnSap3
AcSap2
AcSap1
DdSap3
DdSap11
DdSap12
DdSap8
MaySap1
AcCath1
AtCath3
AtCath2
AtCath1
VexSap1
DdSap5
DdSap14
DdSap2
DdSap10
DdSap1
DdSap7
CeSapL1
DdSap9
NfSap1*
NgSap1
CeSap1
AcSapR1
CoSap2
CoSap1
AcSap7
CiSap1
CiSap2
HsSap1
AtSap1
AtSap2
TvSap1
TvSap2
TvSap3
AcSapR2
CrSapL1
CrSap1
TtSapL1
CoSapL1
CiSapL1
FnSapL1
FnSapL2
HsSapL1
MaySapL1
VexSapL1
NfSapL1
NgSapL1
DdSapL1
1.00/100
≥0.99/99
≥0.95/80
≥0.80/50
Phytepsin clade
0.5

Figure 6.16

Secondly, homology searching retrieved *A. thaliana* saposin domain-containing sequences (known as "phytepsins"). In plants, phytepsins are involved in cell autolysis during developmentally regulated programmed cell death, as a part of organismal development.[682] Phytepsins are Cathepsin D-like aspartyl proteases, however, they contain a saposin domain insertion in the middle of the sequence.[683] The saposin domain of the phytepsins was included in the phylogeny, and surprisingly, grouped not with other plant saposins but within a larger amoebozoan clade containing sequences from *Acanthamoeba castellanii*, *Vexillifera* sp. and *Mayorella* sp. Domain analysis of the amoebozoan sequences showed that the *A. castellanii* sequence (ACA1_173660) has identical domain organization to plant phytepsins, and BLAST searching shows that it is highly similar (but not identical to) the phytepsin in the moss *Physcomitrella patens*. While the E-value is low (6E-158), these sequences only share 49% sequence identity, suggesting that this is not the result of contamination. This result raises the possibility that phytepsin was laterally transferred between amoebae and plants, although the directionality of the transfer cannot be inferred from this phylogeny.

6.5.3 Down-regulated genes associated with pathogenesis

Only 107 genes are significantly down-regulated in high pathogenicity versus normal pathogenicity *N. fowleri* (Online Appendix Table 6.8). Additionally, there are fewer cellular systems with multiple down-regulated genes.

Three genes are down-regulated that are involved in transcriptional and translational inhibition, further supporting the idea that more protein synthesis is occurring in highly pathogenic *N. fowleri*. However, somewhat counter to this, is the down-regulation of multiple translational genes, such as EIF4E and EIF5A, and three spliceosomal proteins. Homology

searches show that in all cases, multiple paralogues with unchanged expression levels were identified for each of these down-regulated genes involved in translation. This suggests that instead of simple up- or down-regulation, there is complexity in the protein synthesis machinery that may be relevant to pathogenesis.

Eight genes involved in cell stress and DNA damage repair/cell cycling are down-regulated in highly pathogenic *N. fowleri*, including 26S proteasome non-ATPase regulatory subunit 10, a DNAJ homologue, and the cell checkpoint protein CHK1,[684,685] which perhaps indicates that these processes are less relevant during infection. Several membrane trafficking genes are down-regulated: a VAMP7 homologue, an Arf GTPase protein, and an unconventional myosin heavy chain. As with the protein synthesis genes that are down-regulated, each of these trafficking proteins has multiple paralogues with different expression patterns, again suggesting complex regulation of these systems.

Machinery involved in cilia and flagella-based cell motility is down-regulated, including cilia and flagella associated protein 100, δ-tubulin, and a sperm associated antigen 16-like protein.[686,687] Because *N. fowleri* infects as an amoeba, it is congruent that flagellar machinery is down-regulated in relation to host infection.

Two genes involved in anaerobism that are likely of bacterial origin (but also present in *N. gruberi*) are down-regulated. These are a hemerythrin-like gene and a dyp-type peroxidase YfeX-like protein. The former contains a heme domain that may be involved in binding oxygen in low-O2 conditions, and YfeX is up-regulated in anaerobiasis in bacteria.[688,689] The down-regulation of these proteins when *N. fowleri* is in the high-oxygen environment of the host is consistent with a potential role in anaerobiasis.

At least three signal transduction genes are down-regulated, although several other genes of this category are also up-regulated; this is another example of the complexity of gene regulation. At least five adenylyl/guanylyl cyclases are downregulated. These proteins are generally involved in cAMP/cGMP production, which play many roles in the cell.

Approximately 70% of the down-regulated genes not in these categories are genes of unknown function, and many are specific to *N. fowleri* or *Naegleria* spp.

6.5.4 Identification of a regulatory motif upstream of up-regulated genes associated with pathogenesis

Differential gene expression as part of infection might be controlled by regulatory elements upstream of DE genes. Of course, not all DE genes are necessarily regulated by the same mechanism; there may be several transcriptional programs regulating different systems resulting in gene co-regulation during infection. Nonetheless, it does not preclude one or more regulatory elements controlling the expression of DE genes. To identify such elements, a region 200 bp upstream of the predicted translation start site of each gene was extracted, and potential transcriptional regulatory motifs were identified using RegRNA2.0. Because regulatory elements can be fairly short, in order to reduce the amount of noise in the predictions, only the upstream regions of up-regulated genes involved in lysosomal function were used to identify potential regulatory elements.

Two potential elements were identified in 11 and 9 of the 30 lysosomal upstream regions, which were ZNF333 and KID3, respectively. Other elements were identified, but with much lower frequency (<6/30, Online Appendix Table 6.10). These elements, ATAAT (ZNF333) and

CCACC (KID3), were searched for in the upstream regions of all up-regulated genes, and were found with a frequency of 34/201 and 17/201, respectively. However, ZNF333 is found even more frequently among the entire cohort of *N. fowleri* genes. It has a general frequency of 0.34, compared to 0.17 in up-regulated genes and 0.36 in lysosomal up-regulated genes. KID3 is found upstream of all *N. fowleri* genes with a frequency of 0.14, compared to 0.08 for up-regulated genes; however, its frequency in lysosomal up-regulated genes is still relatively high (0.3). It is possible that the CCACC motif is a transcriptional regulatory element relevant to lysosomal proteases and general lysosomal function; however, it is found upstream of nearly 1800 genes that are not DE, rendering these results inconclusive at best.

RegRNA requires the user to select an organism database for transcriptional regulatory element prediction. After several iterations with databases of *D. discoideum*, several plant taxa, *P. falciparum*, and *Candida glabrata*, the database that gave the most promising results was for human. The evolutionary distance between *N. fowleri* and *H. sapiens* is large, and this approach may give false positive results, while ignoring other *N. fowleri*-specific motifs. Therefore, the program MEME was used to identify motifs *de novo* in the lysosomal dataset, and a 100-gene subset of all *N. fowleri* genes. Two motifs were identified within the lysosomal dataset (Figure 6.17). The C(A)7 motif is found in the majority of both lysosomal genes and in the random subset (83% and 86% respectively), and in most cases within 15 bp of the predicted gene start site. The other motif, RDAKTTTYHGKWGTT (see Supplementary Table ST6.3 for base frequencies), is found in 20 of the 30 lysosomal sequences, but is not found in the random gene subset. In searching for the motif in all the up-regulated genes, no addition upstream sequences with this motif were identified. Searching the upstream region of all genes predicted in *N. fowleri*, 1206 regions with this motif were identified with a p-value less than 1E-4, which is less

Figure 6.17. Motif logos found upstream of lysosomal genes up-regulated in highly pathogenic *N. fowleri.*

Letter height shows frequency at that position in lysosomal gene dataset. (A) The C(A)7 motif is found up-stream of most *N. fowleri* genes. (B) The RDAKTTTYHGKWGTT motif is found up-stream of two-thirds of genes in the lysosomal dataset, and <10% of all *N. fowleri* genes. For base frequencies in (B), see Supplementary Table ST6.3.

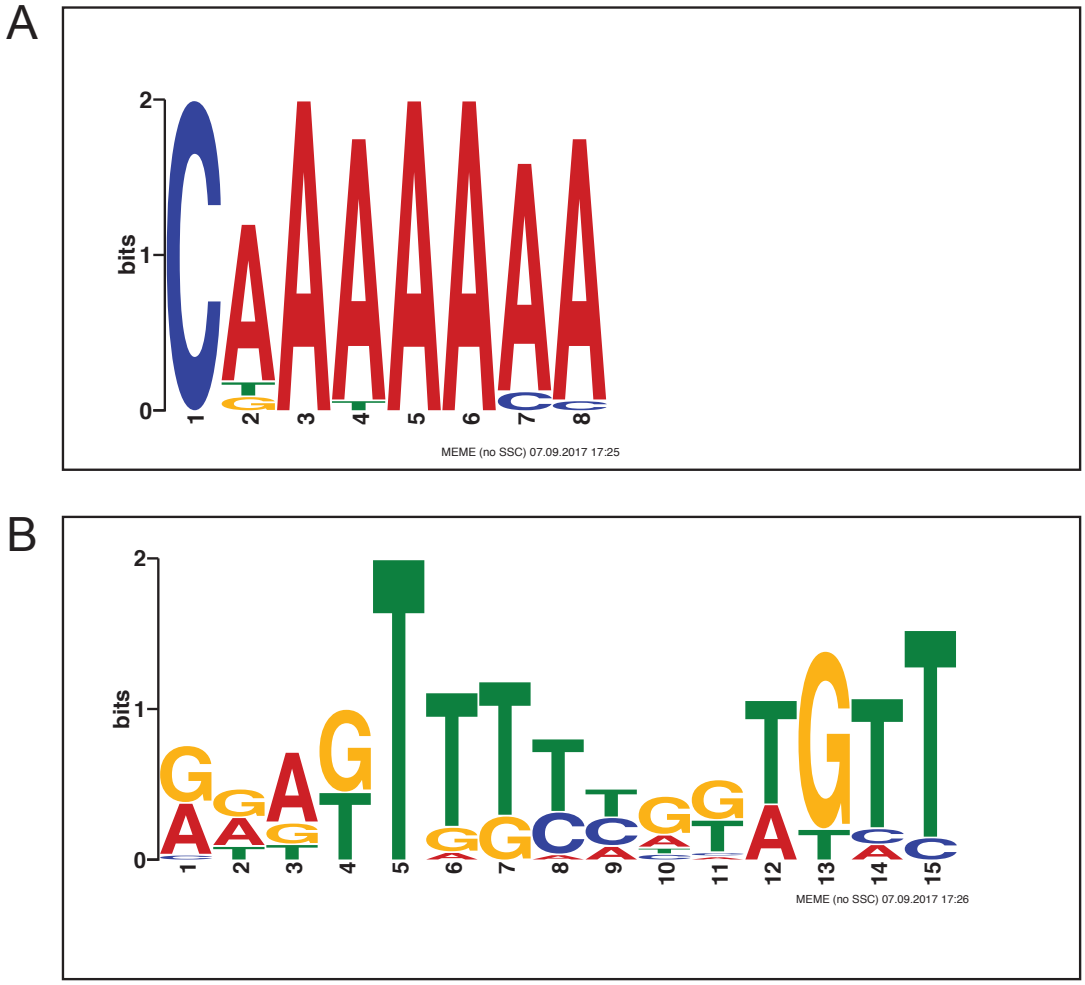

A
bits
2
1
0
1
2
3
4
5
6
7
8
MEME (no SSC) 07.09.2017 17:25
B
bits
2
1
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
MEME (no SSC) 07.09.2017 17:26

Figure 6.17

than 10% of all genes. It is possible that this motif is a *cis*-regulatory transcriptional element upstream of some lysosomal genes, however, there are many other genes with this motif that are not DE. If it is a transcriptional regulatory element, it may be part of a larger system of regulatory features that control the expression of lysosomal genes under different conditions.

6.5.5 The influence of bacterial lateral gene transfer events on pathogenesis

Lateral gene transfer (LGT) in *N. fowleri* related to gain-of-function pathogenicity factors was explored in Chapter 6.3. To focus further on potential bacteria-derived pathogenicity factors, all genes with no *N. gruberi* orthologue and with a bacterial top hit when searching the NR database were examined more closely (Online Appendix Table 6.11). Most genes do not have homologues in pathogenic bacteria, and for those that do, their predicted gene functions are not obviously relevant to pathogenesis. Only two genes on this list are up-regulated in highly pathogenic *N. fowleri*. One is found in members of the Burkholderiales bacteria and in *Dictyostelium*, and the other is a Membrane Associated Proteins in Eicosanoid and Glutathione metabolism domain (MAPEG)-containing protein found in cyanobacteria. Together, these results suggest that LGT is not a source of pathogenicity factors in *N. fowleri*.

6.5.6 Comparative genomics and transcriptomics of proteases

Proteases are a well-known category of pathogenicity factors in many parasitic and pathogenic eukaryotes, including *N. fowleri*.[354,690] Twenty-eight proteases are up-regulated in highly pathogenic *N. fowleri*, making up more than 10% of all up-regulated genes. Differences in the number of proteases between *N. fowleri* and *N. gruberi* may be relevant to pathogenesis. To

better understand the evolution of proteases in the *Naegleria* spp., a comparative genomic analysis of all proteases in the three strains of *N. fowleri* and in *N. gruberi* was undertaken.

Proteases are classified by the residue that acts as the nucleophile in the catalytic reaction, giving rise to the following superfamilies: serine proteases, cysteine proteases, threonine proteases, aspartic proteases, glutamic proteases, metalloproteases, and asparagine peptide lyases. These are defined by the MEROPS classification scheme based on structure, mechanism, and catalytic residue.[691] There is also a 'mixed' superfamily, which contains proteases with similar protein folds, or similarly arranged catalytic residues, but the members have different nucleophilic residues. Within superfamilies are collections of families, and the proteases within each family are evolutionarily related.

The *N. gruberi* proteases are already classified in the MEROPS database; paralogue numbers were compared to those for the three *N. fowleri* strains that were identified here. For some families, the number of paralogues in *N. gruberi* was much larger or smaller than in *N. fowleri*, and therefore the family was re-annotated manually in *N. gruberi*. Table 6.4 (Online Appendix Table 6.12) shows the number of paralogues for each superfamily and family for the four taxa.

Glutamic proteases and asparagine peptide lyases were not identified in either *Naegleria* spp. In general, *N. fowleri* and *N. gruberi* have similar numbers of proteases in each family, and there was only one case where a protease family has a representative in *N. fowleri* but not *N. gruberi*, the serine protease family S81. The sole representative of the S81 family in the MEROPS database is a destabilase protein in *Hirudo medicinalis*, the European medicinal leech. This protein has a destabilase and peptidoglycan-binding domain; it is found in the saliva of the leech, and although its function is unclear, it may have lysozyme activity and be capable of

Table 6.4. Overview of proteases in *N. fowleri* and *N. gruberi*

Boxes indicate subfamilies with genes that are up-regulated in highly pathogenic *N. fowleri*, and red text indicates subfamilies with more members in *N. fowleri* than *N. gruberi*

Type	Family	N. fowleri V212	N. fowleri 986	N. fowleri 30863	N. gruberi	Notes	Contains up-regulated genes?
Aspartyl	A1	3	3	2	3		YES
	A22	1	1	1	1		
	A28	1	1	1	1	DNA damage inducible protein	
Cysteine	C01	21	21	20	36		YES
	C02	12	12	12	19		
	C12	2	2	2	2		
	C13	2	2	2	2		
	C15	1	1	1	1		
	C19	23	23	23	18		
	C26	6	6	7	6		YES
	C39	3	2	3	2	bacterial peptidase	YES
	C40	3	2	3	2		
	C44	4	5	6	2	asparagine synthase	
	C45	2	2	2	2		
	C48	3	3	3	2		
	C50	1	1	1	1		
	C54	1	1	1	1		
	C56	5	5	5	4		
	C65	2	2	2	1	otubain	
	C78	1	1	1	1		
	C83	2	2	2	2		
	C85	7	7	7	4		
	C86	2	2	2	2		
	C89	3	3	3	2		
	C95	4	4	4	6	phospholipase B-like	YES
	C97	3	3	3	3		
	C110	0	0	0	3		
	C115	0	0	0	1		
Metallo							
	M01	5	5	5	6		Down-regulated
	M03	2	2	2	3		
	M08	3	3	3	5		
	M14	6	6	6	6		
	M16	9	9	9	12		YES
	M17	2	2	2	2		
	M18	1	1	1	1		
	M19	2	2	2	2		
	M20	6	6	6	7		
	M24	6	6	6	9		
	M28	6	6	6	6		
	M32	1	1	1	1		
	M38	4	4	4	2		
	M41	3	3	3	3		
	M42	0	0	0	1		
	M48	2	2	2	2		
	M49	1	1	1	2		
	M50	1	1	1	1		
	M54	2	2	2	4		
	M60	1	1	1	1		
	M67	9	9	9	8		
	M76	5	5	5	4		
	M79	1	1	1	2		
	M98	1	1	1	2		
Mixed	P01	1	1	1	1		

Serine	S01	11	10	11	16		
	S08	25	24	20	18		
	S09	33	29	31	32		YES
	S10	9	9	9	12		YES
	S12	4	3	3	5		YES
	S13	1	1	1	1		
	S15	0	0	0	5	includes non-peptidase homologues	
	S16	1	1	1	1		
	S26	2	2	2	3		
	S28	4	4	4	4		YES
	S33	22	21	20	20		
	S45	7	7	7	6		YES
	S49	1	1	1	1		
	S51	0	0	0	1		
	S53	5	5	5	4		YES
	S54	9	8	8	8		
	S59	1	1	1	2		
	S63	0	0	0	0		
	S81	1	1	1	0	destabilase	
Threonine							
	T01	15	15	15	14		
	T02	2	1	1	1		
	T03	2	2	2	3		

dissolving fibrin.[692] A single orthologue is found in all three *N. fowleri* strains, but not in the predicted proteins or genome of *N. gruberi*. The *N. fowleri* V212 sequence was used as a query to search the NCBI non-redundant database, and retrieved no hits, however, a BLAST search into the *H. medicinalis* predicted proteins did retrieve the homologue present in MEROPS (E-value of 8E-12). Zavalova and colleagues (2000) originally identified the protein in *H. medicinalis*, and found homologues in *Caenorhabditis elegans* and bivalve molluscs.[693]

Although it is not differentially expressed in high versus normal pathogenicity *N. fowleri*, it is relatively highly expressed under both conditions, with FPKM values between 500-800. The *N. fowleri* V212 sequence has both destabilase and peptidoglycan-binding domains, suggesting that it may also have dual functions in bacterial cell wall and fibrin breakdown. This makes the *N. fowleri* S81 family protease a prime candidate for future cell biological work to investigate its function and potential role in pathogenesis.

The proteases that are up-regulated in highly pathogenic *N. fowleri* are distributed in 12 families within the aspartyl, cysteine, metallo, and serine superfamilies. Many of these are cathepsin proteases, as well as phospholipase B-like proteins, tripeptidyl peptidases and serine carboxypeptidases. Half of the protease families that have up-regulated members are either localized to lysosomes or are secreted, whereas the others have proteolytic activities in other organelles or in the cytoplasm (for example, a mitochondrial processing peptidase is up-regulated). The most substantially represented types of lysosomal/secreted protease in the up-regulated genes are the cathepsin proteases. Cathepsin proteases represent several unrelated families of proteases that generally function in the lysosome, but can also be secreted. Two cathepsin D aspartyl proteases are up-regulated in highly pathogenic *N. fowleri*. Cathepsins are found in endolysosomes and phagosomes, and plays numerous roles in the cell, including protein

turnover in the lysosome, processing enzymes, prohormones, and growth factors, and functioning in cell survival/apoptosis.[694] It is likely that the role of cathepsin D in *N. fowleri* is restricted to lysosomal degradation, as its additional functions in human are part of a multicellular lifestyle. However, functional work is necessary to corroborate this.

The other major category of cathepsin proteases with several up-regulated members is the C01 family of cathepsin cysteine proteases, with 10 out of 21 up-regulated genes in highly pathogenic *N. fowleri*. The C01 family includes cathepsins B, C, L, Z, and F; cathepsin B in particular is a known pathogenicity factor in *N. fowleri*, as well as *Leishmania*, *Giardia*, *Trichomonas*, *Entamoeba*, and schistosome worms.[354,654,659,695–698] Human cathepsin B has been shown to degrade the extracellular matrix, and is important in cancer metastasis.[699] Each of these families has multiple members, up to 10 in the case of cathepsin B. Only members of cathepsin B, Z, and F are up-regulated; none of the five cathepsin C orthologues are differentially expressed. Despite the large number of C01 family cathepsin proteases in *N. fowleri*, *N. gruberi* encodes even more members. While the three *N. fowleri* strains have 20-21 C01 family members, *N. gruberi* has 35. In order to get a better understanding of the expansions and/or losses in this family within *Naegleria* spp., phylogenetic analysis of the C01 cysteine protease family was performed.

6.5.6.1 Evolution of cathepsin cysteine proteases in Naegleria

To determine the lineage-specific evolution of the C01 cysteine proteases in *N. fowleri* V212 and *N. gruberi*, a phylogenetic analysis of these proteins was performed. Figure 6.18 shows the clades of cathepsin B, C, Z, and L-like (F) proteins in *Naegleria* spp. Many of the *N. fowleri* and *N. gruberi* cathepsins have 1:1 orthology, with at least three expansions that have

Figure 6.18. Phylogenetic analysis of the C01 cysteine protease subfamily in *N. fowleri* and *N. gruberi*.

Node values are listed as Phylobayes/RAxML (posterior probability/bootstrap), and as symbols indicating a minimum level of support as shown in the inset. Node values are shown on the best Bayesian topology. Sequences with signal peptides have red text, those with potential signal peptides (score near cutoff) have orange text, and those without identifiable signal peptides have blue text. Asterisks (*) indicate genes that are up-regulated in highly pathogenic *N. fowleri*.

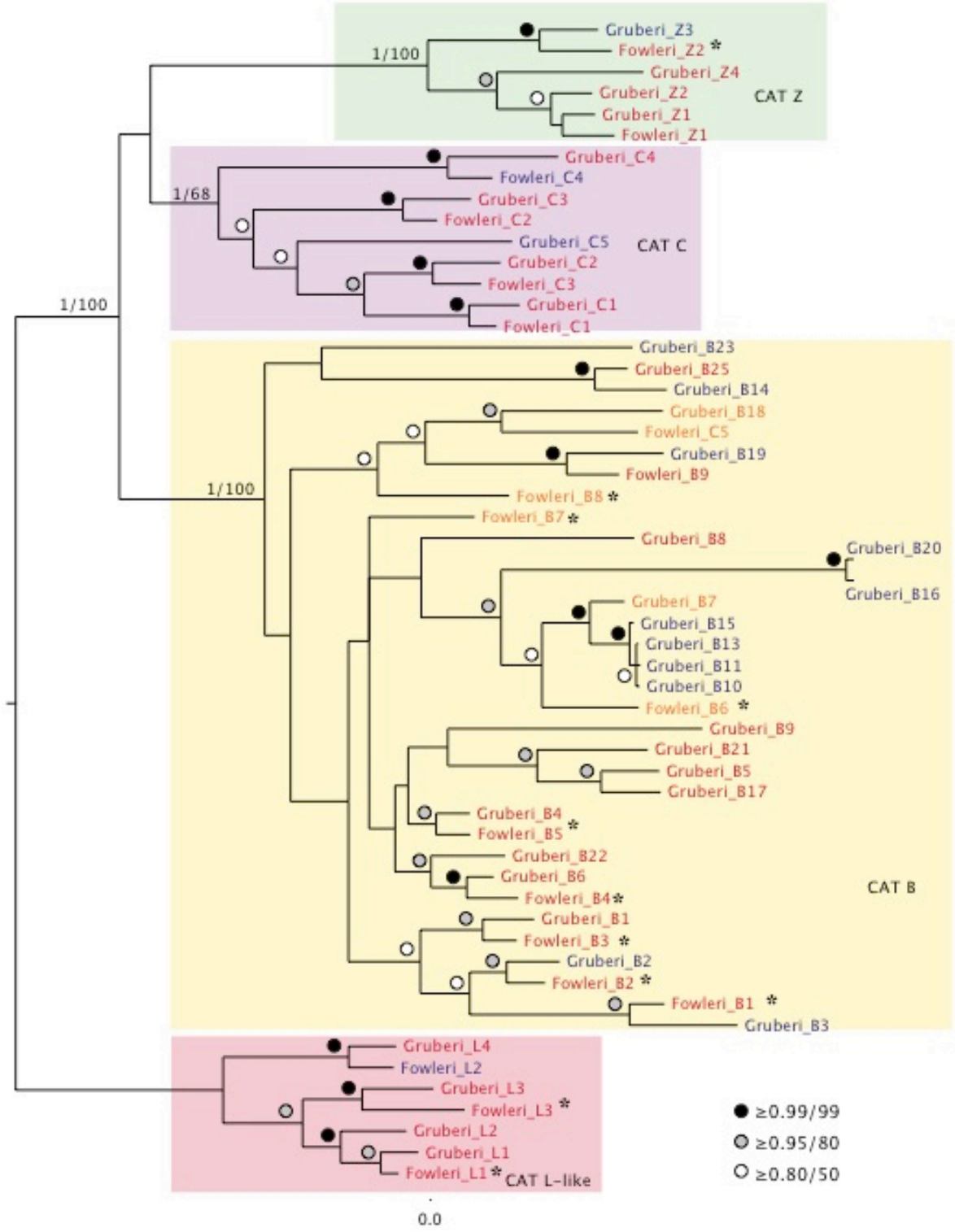

Gruberi_Z3
Fowleri_Z2 *
Gruberi_Z4
Gruberi_Z2
Gruberi_Z1
Fowleri_Z1
CAT Z
1/100
Gruberi_C4
Fowleri_C4
Gruberi_C3
Fowleri_C2
Gruberi_C5
Gruberi_C2
Fowleri_C3
Gruberi_C1
Fowleri_C1
CAT C
1/68
1/100
Gruberi_B23
Gruberi_B25
Gruberi_B14
Gruberi_B18
Fowleri_C5
Gruberi_B19
Fowleri_B9
Fowleri_B8 *
Fowleri_B7 *
Gruberi_B8
Gruberi_B20
Gruberi_B16
Gruberi_B7
Gruberi_B15
Gruberi_B13
Gruberi_B11
Gruberi_B10
Fowleri_B6 *
Gruberi_B9
Gruberi_B21
Gruberi_B5
Gruberi_B17
Gruberi_B4
Fowleri_B5 *
Gruberi_B22
Gruberi_B6
Fowleri_B4 *
Gruberi_B1
Fowleri_B3 *
Gruberi_B2
Fowleri_B2 *
Fowleri_B1 *
Gruberi_B3
CAT B
1/100
Gruberi_L4
Fowleri_L2
Gruberi_L3
Fowleri_L3 *
Gruberi_L2
Gruberi_L1
Fowleri_L1 *
CAT L-like
0.0
≥0.99/99
≥0.95/80
≥0.80/50

Figure 6.18

occurred in the cathepsin B clade in *N. gruberi*. These expansions account for most of the difference in paralogue number between the two species, making up 12 of ~16 *N. gruberi*-specific C01 homologues. Interestingly, *N. fowleri* Cathepsin B7 and B8 are up-regulated and do not have orthologues in *N. gruberi*, and could help to explain *N. fowleri*'s pathogenicity. Functional work is therefore necessary to understand their role during infection.

When aligning these sequences in preparation for phylogenetic analysis, it was observed that the *N. gruberi* sequences had a shorter N-terminus than the *N. fowleri* sequences. As the cysteine protease cathepsins are known to be either lysosomal or secreted, the N-terminus of the protein must contain a signal peptide that targets it to be translated into the ER, destining it for the secretory pathway. The shortened N-termini of the *N. gruberi* sequences raised the possibility that the signal peptide was potentially missing. To assess this, signal peptides in the *N. gruberi* and *N. fowleri* C01 cathepsins were predicted using the SignalP 4.0 prediction server (Online Appendix Table 6.13). In the majority of cases, signal peptides were identified, but when they were not, the genes were manually assessed to ensure that no region upstream of the gene was mis-predicted and excluded from the gene model. In Figure 6.18, the cathepsin tree is shown with the nodes colour-coded based on the presence of a predicted signal peptide. Most cathepsin sequences did contain signal peptides, however, there was a notable absence of signal peptides from sequences within one of the large expanded clades of *N. gruberi* cathepsin B sequences. Therefore, it is unclear whether these proteins are able to enter the secretory system to be trafficked to the lysosome, or to be secreted from the cell. It is possible that they may have a novel function, as a human cathepsin L lacking a signal peptide has been shown to localize to the nucleus and function as a transcription factor.[700] In contrast to this, only two *N. fowleri* sequences appear to be missing a signal peptide (cathepsin Z and cathepsin L-like homologues),

while four other *N. fowleri* cathepsin B sequences had a signal peptide prediction score near the cutoff value. One of the latter sequences is in the same clade as multiple *N. gruberi* sequences that lack signal peptides; however, this *N. fowleri* sequence is up-regulated in highly pathogenic *N. fowleri*. This raises questions about the localization and function of these proteins in *Naegleria* spp.

6.6 Confirming the presence of a Golgi body in *Naegleria gruberi*

One distinctive cellular feature of *N. gruberi* is that it lacks a visible Golgi organelle. In mammalian cells and many diverse eukaryotes, the Golgi appears as a stack of flattened membranes, or cisternae, inside which proteins are modified via glycosylation and transported to the plasma membrane or to endocytic organelles. In mammalian cells, disrupting the stacked structure of the Golgi leads to numerous defects in these processes, and Golgi fragmentation is observed in autoimmune diseases, cancer, Huntington's, Parkinson's, and Alzheimer's diseases.[701] However, examples of eukaryotes with unstacked Golgi are found across the evolutionary tree, and include the budding yeast *Saccharomyces cerevisiae*,[95] as well as parasitic taxa such as *Plasmodium falciparum, Entamoeba histolytica, and G. intestinalis*.[480] There is an array of unstacked Golgi morphologies. In *S. cerevisiae*, the Golgi compartments are dispersed in the cytoplasm, appearing as punctae in immunofluorescent staining of Golgi markers.[96] *E. histolytica* was originally thought to not have a Golgi, but an ultrastructural study showed that this was an artefact of the fixation process for transmission electron microscopy, and presented micrographic evidence of dispersed cisternae in the cell.[97] Finally, Golgi functions in *G. intestinalis* are stage specific, and are carried out in encystation-specific vesicles that form

dynamic, tubular structures.[98] Unlike typical Golgi, they are not steady state organelles, but arise in response to cyst wall material formation in the ER.[702,703]

The Golgi of *N. gruberi* has never been visualized, but it is predicted to exist based on the presence of Golgi-related membrane trafficking proteins encoded in its genome (Figures 6.7-6.9, Online Appendix Table 6.2). These include the coats COPI and retromer; adaptor complexes 1, 3, and 4; MTCs TRAPPI, GARP, and COG; SNAREs Ykt6, Syntaxin 5, GS28, Bos1, GS15, Syntaxin 6, Syntaxin 16, and Vti1; and the SM proteins Sly1 and Vps45. Using these predicted protein sequences, antibodies were generated to markers of the Golgi (COPB), ER (Sec31), and plasma membrane (Syntaxin PM). After Western blotting subcellular fractions of *N. gruberi* with these antibodies to show sensitivity and specificity, they were used in immunofluorescence microscopy (IFM) and immune-electron microscopy (IEM) to survey the organellar landscape of *N. gruberi*.

6.6.1 Detecting membrane trafficking markers of endomembrane organelles by Western blotting

Subcellular fractionation of *N. gruberi* was performed to separate whole cells, membrane fragments and nuclei (whole cell lysate), mitochondria, membrane-bound organelles, and cytosol. These fractions were analysed by SDS-PAGE followed by a Western blot to validate the sensitivity and specificity of the antibodies (Figure 6.19). Blots probed with the polyclonal chicken α-NgCOPB antisera shows a band near the predicted molecular weight of 114.5 kDa, largely in the whole cell and cytoplasmic fractions. Blots probed with the polyclonal rabbit α-NgSynPM antisera and polyclonal rat α-NgSec31 antisera show bands near the predicted molecular weights of 39kDa and 145.7 kDa, respectively. These Western blots were preliminary work, and have since been further optimized by L. Yiangou. Based on these data, COPB, Sec31,

Figure 6.19. Western blots of NgCOPB, NgSec31, and NgSynPM from subcellular fractions of *N. gruberi* cells.

Whole cell lysates (WC), membrane fractions (Me), mitochondrial fractions (Mt), and cytoplasmic fractions (Cy) are run on gels and transferred to membranes. Membranes are blotted with anti-NgCOPB antibodies at a dilution of 1:200 (A), anti-NgSec31 antibodies at a dilution of 1:400 (B), and anti-NgSynPM antibodies at a dilution of 1:2500 (C). *L*, ladder.

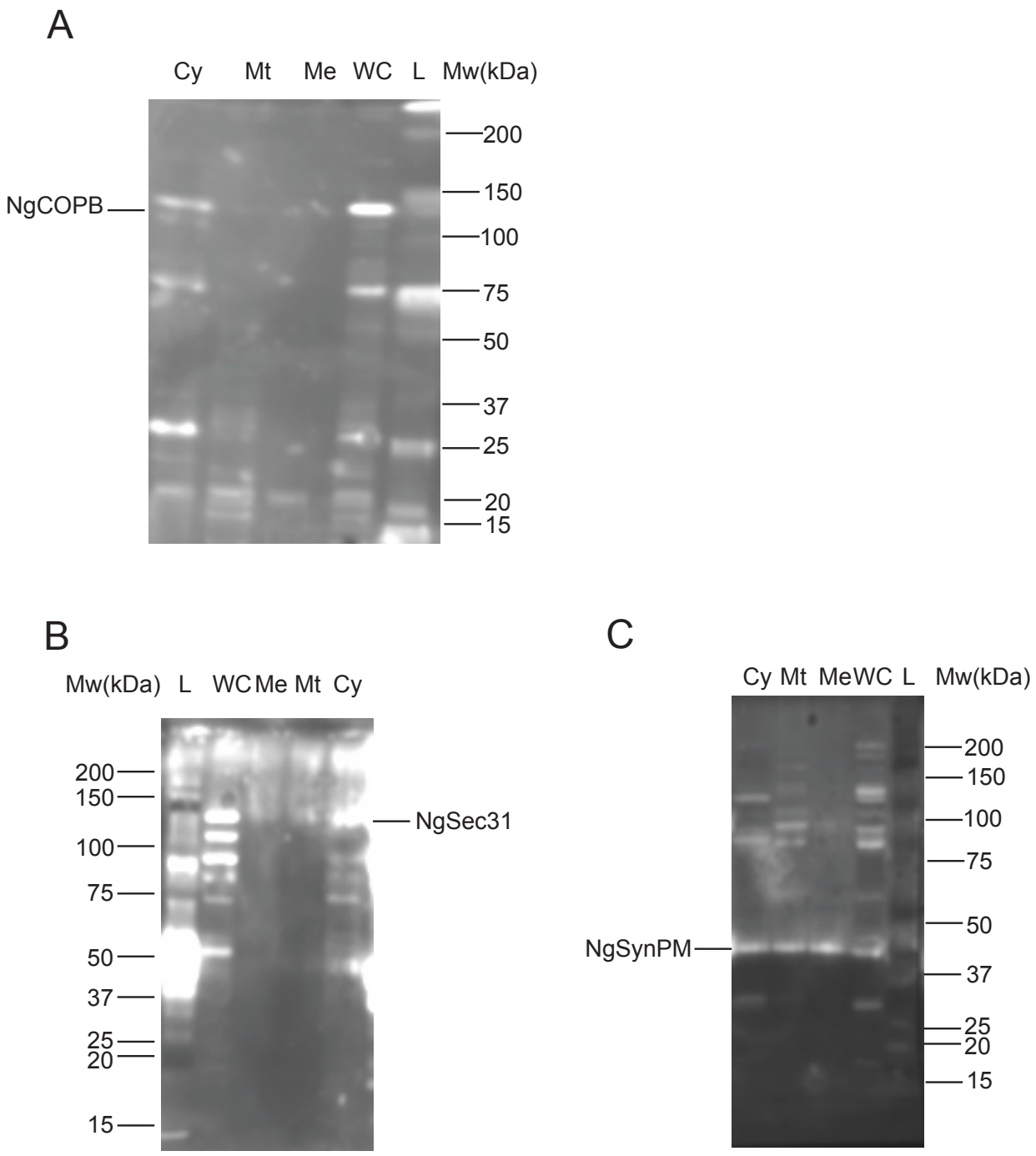

A
Cy
Mt
Me
WC
L
Mw(kDa)
NgCOPB
200
150
100
75
50
37
25
20
15
B
Mw(kDa)
L
WC
Me
Mt
Cy
NgSec31
200
150
100
75
50
37
25
20
15
C
Cy
Mt
Me
WC
L
Mw(kDa)
NgSynPM
200
150
100
75
50
37
25
20
15

Figure 6.19

and Syntaxin PM proteins are expressed in *N. gruberi*, and in the appropriate subcellular location.

6.6.2 Immunofluorescence microscopy of organellar markers

To understand the structure of the *Naegleria* Golgi, as well as visualize other parts of the endomembrane system, these antibodies were used in fluorescence microscopy of *Naegleria* cells. After incubating fixed, permeabilized *N. gruberi* cells with the primary antibodies, secondary fluorescent anti-IgG antibodies were added to the corresponding primary antibodies. DNA was stained with DAPI. Figure 6.20A shows representative fluorescence micrograph images showing NgCOPB, NgSec31, and NgSynPM localization.

NgCOPB appears as small puncta, although there is a high level of background staining. Again, these experiments were repeated and optimized, and the final figure showing COPB localization are shown alongside the initial image in Figures 6.20B and 6.20C. In Figure 6.20B, the COPB antibody appears to localize to tubular structures in the cytoplasm of the cell. In Figure 6.20C, NgCOPB and NgSec31 appear to have partially overlapping but distinct staining. Finally, NgSynPM is diffuse in the cytoplasm, and does not overlap with NgSec31 signal. These images show that NgCOPB labels an organelle that appears to be distinct from the ER (NgSec31), with some clearly overlapping regions.

6.6.3 Immuno-electron microscopy of organellar markers

To better investigate the localization of these endomembrane proteins, immuno-gold electron microscopy was performed. *N. gruberi* cells were suspended in resin, sectioned, and

Figure 6.20. Immunofluorescence and confocal microscopy of NgCOPB, NgSynPM, and NgSec31.

(A) Original immunofluorescence images of NgCOPB, NgSynPM, and NgSec31. Antibodies to these proteins are shown in green, nuclei are stained blue (DAPI), and mitochondria (mitotracker) are shown in red. (B) Confocal micrograph of anti-NgCOPB (green), with DAPI in blue, optimized by L. Yiangou and D. Cantoni. (C) Confocal micrographs showing costaining of NgCOPB with NgSec31 and NgSynPM, as well as ER-Tracker (staining the ER) and DAPI (staining DNA). These images were produced by L. Yiangou and D. Cantoni. *N. gruberi* cell sizes range from 10-30 μm.

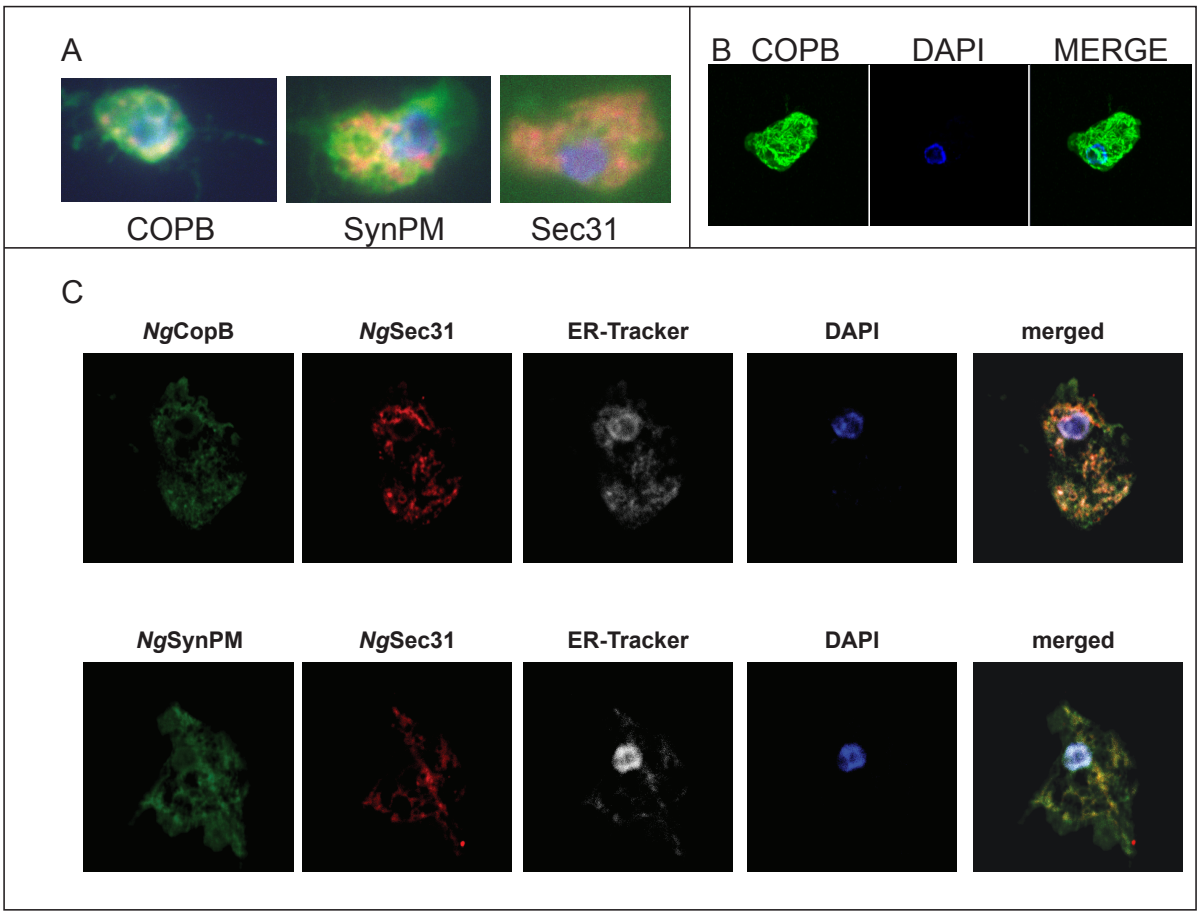

A
COPB
SynPM
Sec31
B COPB
DAPI
MERGE
C
NgCopB
NgSec31
ER-Tracker
DAPI
merged
NgSynPM
NgSec31
ER-Tracker
DAPI
merged

Figure 6.20

Placed on gold EM grids labeled using primary antibodies, and then incubated with the corresponding gold-labeled secondary antibodies. Figure 6.21 shows NgCOPB labeling membrane-bound organelles in the cell. Images of NgSec31 and NgSynPM show non-specific labeling throughout the cell, and are therefore not included here.

6.7 Discussion

In this Chapter, comparative genomics, transcriptomics, and molecular biological techniques to understand the biology of the neuropathogenic amoeba *Naegleria fowleri*, in comparison with its harmless relative *Naegleria gruberi*. Many facets of the *N. fowleri* and *N. gruberi* genomes were explored using comparative genomics, in order to identify any differences between the two organisms that may explain why *N. fowleri* alone is able to infect humans. In the Introduction, this type of analysis was suggested to be predictive of cell biological function, and furthermore has been successfully used to understand pathogenesis in the *Candida* genus.[704] Building on comparative genomics work, these predictions can be confirmed or expanded on by gene expression analyses, which pinpoint the genes that are differentially expressed under a particular condition. To this end, comparative transcriptomics was performed to home in on the cell biology of pathogenesis, with normal, pathogenic *N. fowleri,* and *N. fowleri* grown to be even more virulent. Finally, the results of comparative genomics and transcriptomics in *Naegleria* spp. were then used as a basis for cell biological work to elucidate the endomembrane landscape in this organism. The wealth of data produced by the analyses in this chapter is a testament to the power and utility of large-scale '-omics', and will be the source of further downstream cell biological investigations.

Figure 6.21. Immuno-electron microscopy showing NgCOPB in *N. gruberi.*

Immuno-gold labeled NgCOPB is shown in vesicular-tubular structures in *N. gruberi*. Inset image shows clear labeling of these organelles at a higher magnification. The lower right graph shows the average amount of gold labeling of different *N. gruberi* compartments, suggesting that it is mainly localized to membrane-bound organelles.

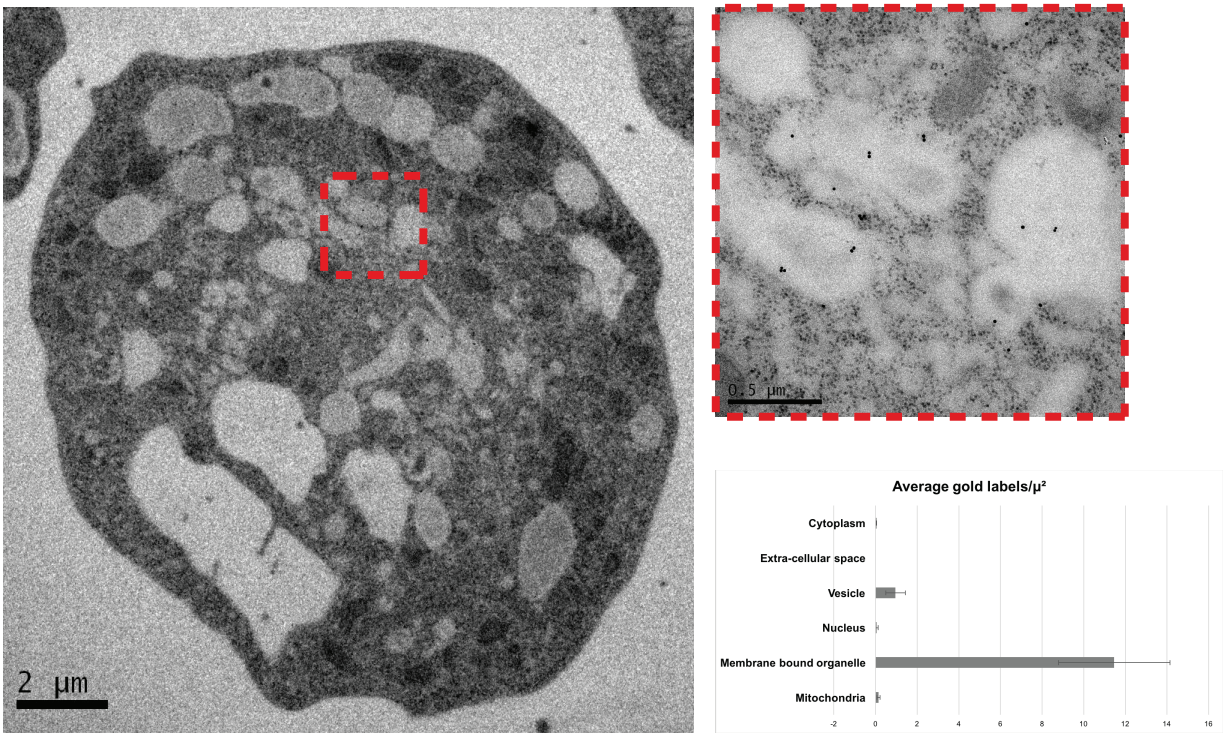

2 µm
0.5 µm
Average gold labels/µ²
Cytoplasm
Extra-cellular space
Vesicle
Nucleus
Membrane bound organelle
Mitochondria
-2
0
2
4
6
8
10
12
14
16

Figure 6.21

The mitochondrial genomes of *N. fowleri* and *N. gruberi* are highly similar in both gene content and organization, the latter being in stark contrast to the nuclear genomes. The nuclear genome of *N. gruberi* is approximately 13 Mb larger than any of the three *N. fowleri* strains. From the orthologue clustering analysis, only 762 orthologue groups appear to be specific to *N. gruberi*, despite it encoding ~3000 more protein-coding genes than *N. fowleri*. Manual inspection of these groups suggest that the majority are not highly paralogous, and therefore the 'extra' *N. gruberi* proteins are likely to be paralogous duplications of genes shared between the two *Naegleria* species. This appears to be the case when looking at particular cellular systems; one clear example is the cathepsin B family of cysteine proteases, in which *N. gruberi* encodes nearly twice as many paralogues as *N. fowleri*.

The genes in the *N. fowleri* genome are also more tightly packed, generally with very little intergenic space, and shorter intron lengths compared to *N. gruberi.* Often genome compaction is seen in parasites, but this pattern is also observed in the vertebrate *Fugu rubripes* (Japanese pufferfish) and other teleost fish.[705] However, with no other sampling points within the *Naegleria* genus, it is just as likely that *N. gruberi* has increased its non-coding DNA, in parallel with gene duplication events throughout the genome. One surprising finding was the almost complete lack of shared genome organization between the four *Naegleria* genomes. Despite high sequence similarity and similar paralogue number between the *N. fowleri* strains, the level of genomic rearrangement observed is more typical of an inter-species comparison, as exemplified by the analysis of three *Saccharomyces* species. It is likely that this is due to recombination during meiosis. *N. gruberi* is known to be tetraploid, and although it is primarily asexual, it has maintained apparently functional copies of genes required for meiosis.[361] Additionally, there is strong evidence for a sexual cycle in *N. lovaniensis*.[617] As more genomes of closely related

eukaryotes are sequenced, it will be interesting to see how typical this extensive rearrangement is in other lineages.

Analysis of orthologous groups showed that there are approximately 3,000 groups of genes that are found in *N. fowleri*, for which there is no clear orthologue in *N. gruberi*. While this type of automated analysis is a good first pass, additional BLAST searches dropped this number to fewer than 1,000. Because of the difficulty in determining orthology versus paralogy on the scale of thousands of genes, this type of two-step filtering is necessary to remove false positives. The majority of the sequences that make up the *N. fowleri*-specific orthogroups have no BLAST hit in the non-redundant database. This leads to the question of whether these are mis-predicted genes. Only eight of these genes have no transcriptomic data associated with them from *N. fowleri* LEE-Ax or LEE-MP samples, suggesting that the vast majority is expressed at some point in *N. fowleri*. Many of them have no hits in any other organism, and make up a not insignificant fraction of differentially expressed genes in highly pathogenic *N. fowleri*. Although their functions are unknown at this point, they represent an attractive category of genes for further functional work, as they could very well be unknown pathogenesis factors.

Of the genes that are found in *N. fowleri* and other organisms, but not *N. gruberi*, these fall into two main categories: genes that were present in the common ancestor of *N. fowleri* and *N. gruberi* but have been lost from *N. gruberi*, and genes in *N. fowleri* that are the result of lateral gene transfer. (Of course, this ignores the case of an LGT in a *Naegleria* ancestor and subsequent loss from *N. gruberi*, although this is indistinguishable from straight LGT into *N. fowleri*). Overall, it appears that there are few *N. fowleri*-specific LGT events that are shared by all three strains, and therefore LGT is likely not a major factor in pathogenesis. This is in contrast to what has been observed in other organisms such as *Blastocystis* sp., in which bacterial

LGT has been critical for adapting to the anaerobic gut environment.[706] It again points to the fact that *N. fowleri* does not require a host, and there is no obvious evolutionary advantage associated with infection. This is not meant to discount any potential involvement of individual laterally transferred gene in host infection, however, as there are a number of putative LGTs identified in the up-regulated genes associated with pathogenesis. One major concern is the lack of sampling points in this lineage; more *Naegleria* genomes, particularly those more closely related to *N. fowleri*, will give a better idea of the timing of any putative LGT events. Additionally, phylogenetics is also needed to confirm these putative findings.

When starting this project, systems suspected to be involved in pathogenesis were selected for comparative genomic analysis, including the membrane trafficking system, autophagy, stress responses, and adhesion factors. In general, there are no major differences between *N. fowleri* and *N. gruberi* in these systems that are likely to play a role in pathogenesis. Particularly in the cases of membrane trafficking and autophagy, these systems are largely complete, and there are only a few loss events in both taxa. This suggests that there is no evolutionary pressure to 'streamline' the genome, and that these processes are, in general, required for normal cellular function in this lineage. This is not surprising, as neither *N. fowleri* nor *N. gruberi* are parasitic or permanently host-associated; instead, they are normally soil-associated heterotrophs that do not inhabit a niche environment like the gut, or support a unique cell biological process like scale formation. This conservation of proteins shown to be present in the LECA makes *N. gruberi* an excellent potential model system, as there are currently no free-living model systems within the supergroup Excavata. However, despite no large differences in membrane trafficking gene complement between the two organisms, it is clear from the

differential expression analysis that the trafficking system is critical to pathogenesis, particularly the lysosomal trafficking pathway.

Both *N. fowleri* and *N. gruberi* have duplicated some exocytic machinery, SynPM and Vamp7 SNAREs, and subunits of the Exocyst MTC, as well as recycling machinery, such as CORVET and AP4. While expansion of the recycling machinery is more common, and suggests complexity and subfunctionalization in this pathway, it is more rare to see late exocytic components that have been duplicated. Extensive duplications of exocytic SNAREs and the Exocyst tethering complex are characteristic of multicellular land plants. While most land plants encode only 1-2 copies of most exocyst components, they can have ~3-7 Exo84 subunits, and shockingly, copies of Exo70 numbering in the teens, twenties, thirties and forties.[297] Furthermore, the SynPM-related Qa SNAREs and VAMP7-like SNAREs, which can function in both plasma membrane and lysosomal transport are highly expanded in *Arabidopsis thaliana*.[707] Experimental work in *A. thaliana* shows that these SNAREs do not share a redundant function, and are speculated to traffic to different domains of the plasma membrane. It is possible that this is also the case in *Naegleria* spp., especially as this lineage has retained other plasma membrane trafficking factors such as TSET and the SNAREs NPSN and Syp7. However, it is difficult to imagine how spatial control of secretion in an amoeboflagellate is as beneficial as it is to a multicellular plant. Perhaps the diversity in this system is more relevant to temporal control of vesicle fusion due to different reaction kinetics, or perhaps regulated secretion of specific vesicles. Further experimental work will be required to determine the localization and function of these proteins.

The ERAD machinery has been studied less extensively in eukaryotes; however, it appears to be well-conserved in both *Naegleria* species. And in the UPR, although the

transcription factors in eukaryotes (other than IRE1) are generally unknown, *N. fowleri* and *N. gruberi* have identical gene complements that contain the pan-eukaryotic members of this pathway. Again, in both cases, there are no clear differences that might make *N. fowleri* better able to cope with the stress of the host environment. However, it may be that *Naegleria* spp. encodes part of a – or a completely distinct – lineage-specific UPR, as is the case in *Giardia*. This seems unlikely, given how typical the *Naegleria* species are in comparison with the highly derived *Giardia*. It has been speculated that a stress response is the key to *N. fowleri*'s ability to infect the host. Contrary to this, several stress response genes are down-regulated in highly pathogenic *N. fowleri*, which suggests that infection does not induce a stress response. These include a proteasome chaperone protein, a DNAJ/HSP40 protein, and an HSP20 (bacteria-like) protein.

Much like the expanded plasma membrane-associated SNARE complement and potential LGTs, work to identify adhesion proteins resulted in a list of factors to explore experimentally. Many of the adhesion GPCR proteins in both organisms have extracellular domains typical in adhesion proteins in human cells. For example, EGF and NHL repeats are found in the extracellular portion of the teneurin family of proteins functions, which function as cell-cell adhesion receptors in neuronal networks in *Drosophila*, mouse, and human cells.[655] While these domains are pan-eukaryotic, it appears that specific adhesion proteins are lineage-specific, and identifying homologous sequence beyond these domains between organisms of different lineages is not possible.

The differential expression analysis of genes and systems involved in pathogenesis brought forth a surfeit of data to try to piece together into coherent pathways. The three major themes of up-regulated genes are lysosomal degradation, cell housekeeping (metabolism,

transcription, translation), and amoeboid movement. A large proportion of the up-regulated genes encode products that function in lysosomal proteolysis, are involved in lysosome acidification, traffic cargo to the lysosome, etc. In terms of cellular function, this suggests that a key part of *N. fowleri* infection is the breakdown of internalized material. However, it is also possible that a population of these lysosomes is secreted as exosomes to damage nearby host cells, particularly when moving through nasal tissue and extracellular matrix. Release of cytolytic molecules has previously been shown to be a way in which *N. fowleri* destroys target cells. Additionally, proteins such as cathepsin B can be trafficked directly to the cell surface. Both scenarios may explain the exocytic trafficking machinery.

Genes involved in metabolism and protein production are another major category of up-regulated genes associated with pathogenesis. These are likely to be downstream consequences of *N. fowleri* consuming host tissue, and further support the idea that *N. fowleri* is not in a stressful environment. Actin-based motility genes are up-regulated in highly pathogenic *N. fowleri*, while several flagellar genes are down-regulated. Nf-Actin has specifically been classified as a pathogenicity factor due to its role in phagocytosis and trogocytosis via food cup formation.[355] Previous work has shown that only the amoeboid form of *N. fowleri* is capable of infecting a host, therefore seeing this pattern is further evidence that the differential expression analysis is biologically meaningful.

Another way to test the quality of the data is to look for pathogenicity factors described by others. Naegleriapore A and B are pore-forming proteins that have antibacterial activity against both Gram positive and Gram negative bacteria, but are also capable of lysing a variety of target cells.[358] Not only did the differential expression analysis identify the saposin precursor that generates Naegleriapore A and B as up-regulated in association with pathogenesis, but

another homologous sequence was identified in the up-regulated dataset. Other previously identified pathogenicity factors found to be up-regulated in relation to pathogenesis include phospholipases[359] and Nf314 (Cathepsin A).[356] Another protein found at food cups, Nfa1,[708,709] was not up-regulated, but it is highly expressed in both mouse-passaged and axenically cultured *N. fowleri* LEE (>1000 FPKM), providing further evidence for its role as a pathogenicity factor. In 2014, Zysset-Burri and colleagues published a proteomic screen of highly virulent versus weakly virulent *N. fowleri*, as a function of culturing cells with different types of media.[613] While the factors that they found had higher protein levels in more virulent *N. fowleri* were generally non-overlapping with the sequence set generated by the DE analysis here, there were some shared pathways. For example, they identified villin and severin as being more abundant in high virulence *N. fowleri*, which are involved in actin dynamics; a process represented in the up-regulated transcriptomic dataset, albeit by different genes. Other pathogenicity factors have been identified in *N. fowleri* whose gene sequence is unknown, such as glycoproteins involved in complement-mediated lysis,[350] and a putative CD59-like protein.[352] Using the genomic and transcriptomic data, the identity and function of these proteins can start to be pieced together.

A major pathogenicity factor in *N. fowleri* as well as other pathogens is the secretion of the C01 family of cathepsin cysteine proteases. Nearly all of the cathepsins in this family are up-regulated in association with pathogenesis. However, while most other protease families are approximately equivalent in number between *N. fowleri* and *N. gruberi*, *N. gruberi* encodes more members of these cysteine proteases than *N. fowleri*. *In silico* predition of signal peptides showed that many of these 'extra' *N. gruberi* paralogues do not appear to have signal peptides, which direct nascent polypeptides to the ER for translation in order to enter the secretory pathway (although some *N. fowleri* sequences also appear to be missing them). Work by others has shown

that signal peptide-less cathepsins can function as transcription factors,[700] which may also be the case for these paralogues.

In looking at the down-regulated gene dataset, it appears at first glance that there are conflicting patterns. For example, several eukaryotic translation initiation factors are down-regulated, despite an overall pattern of up-regulated of mRNA translation. Closer inspection of these results showed that these genes have paralogues whose expression does not significantly change in relation to pathogenesis. Another example is that of a VAMP7-like protein, which can be involved in both endo-lysosomal and plasma membrane transport. However, it is also one of many paralogues in *N. fowleri*. This echoes the findings of Chapter 4, in which many of the membrane trafficking genes in *Entamoeba invadens* are paralogous, and have different and even sometimes opposing expression patterns during encystation. Again, it appears that cellular system modulation is a key factor in pathogenesis, which can be controlled by having genes with multiple, functionally distinct paralogues, and varying expression in response to the environment.

This leads to the elephant-in-the-room question of what, exactly, are the necessary and sufficient pathogenicity factors in *N. fowleri*? One of the often-made assumptions is that these genes must be present only in *N. fowleri,* and not in *N. gruberi*. But this does not have to be the case. It has been previously observed that differences in virulence of different *N. fowleri* strains can come down to slightly higher or lower levels of protein expression, rather than all-or-nothing gene expression.[359 *inter alia*] In this analysis, several known pathogenicity factors, e.g. the saposins and cathepsins, have clear orthologues in *N. gruberi*. Together, these results suggest that a major aspect of pathogenicity in *N. fowleri* is about "working with what you've got;" in that gene expression and protein levels are key determinants of infection ability. An attempt to identify a

regulatory sequence upstream of the up-regulated lysosomal genes was largely unsuccessful, however, determining the mechanism of gene expression regulation in terms of pathogenesis will be key to understanding this process.

However, there are still some factors that are likely to be specific to *N. fowleri*, and important for virulence. One is thermotolerance. A single potential LGT was proposed as potentially aiding thermotolerance in *N. fowleri*. The question of thermotolerance should be addressed in future analyses as a separate, but related aspect of pathogenesis. This may involve the organism *Naegleria lovaniensis*, which is the outgroup taxon to *N. fowleri*, and is thermotolerant, but not pathogenic. Comparative genomic analyses of these *Naegleria* species could shed light on the relationship between thermotolerance and pathogenesis. Secondly, this analysis largely ignores the fact that dozens of genes are differentially expressed, but could not be meaningfully annotated. It is possible that relevant pathogenicity factors are found in these "unknown" proteins, and without functional characterization in *Naegleria* spp. or other eukaryotes, they represent still missing pieces of the pathogenicity puzzle. An example of what might be found is a protein responsible for trogocytosis, which describes *N. fowleri* pulling membrane off of host cells in a piece-meal fashion using its 'food cup' membrane extensions. This way of cellular eating is also performed by *Entamoeba,* and recently, the protein specifically responsible for trogocytosis was shown to be AGC family kinase 1.[710] However, this protein is not conserved in *Naegleria* spp. It likely uses a different lineage-specific mechanism, which may involve proteins that are – at this point – broadly annotated or un-annotated. Functional characterization of these proteins is therefore critical to getting the full picture of pathogenesis in *N. fowleri*.

This type of molecular biological analysis is the next logical step from data produced by comparative genomics and transcriptomics. Both comparative genomic and transcriptomic data suggest that *N. gruberi* cells have Golgi bodies. However, because they do not have the typical 'stacked' morphology, they had never previously been observed. In order to visualize the Golgi in *N. gruberi* – as well as the endomembrane organellar landscape – antisera to NgCOPB (Golgi), NgSec31 (ER), and NgSynPM (plasma membrane) were generated, and used in immunofluorescence microscopy and immunoelectron microscopy. Fluorescence microscopy images show NgCOPB-labeled structures as punctate organelles, and have a different staining pattern than that of NgSec31 or NgSynPM. EM images further show that NgCOPB gold-labeled beads are associated with distinct organelles, rather than the cytoplasm. In EM images published by Stevens *et al.* (1980), a tubular-vesicular structure in *Naegleria* is labeled as a 'primitive Golgi complex'.[608] Without antibody staining, it is not possible to know whether this is truly a Golgi body, but if it were, its structure is not inconsistent with the immuno-EM data presented here. The *N. gruberi* Golgi likely does exist, and appears as discrete, membranous compartments, not unlike other organisms lacking a stacked Golgi, such as *Giardia* and *Entamoeba.* Tubular Golgi structures have also been described in *Encephalitozoon*,[711] suggesting that there exists a range of unstacked Golgi morphologies. Further analysis is required to determine the precise morphology of the *N. gruberi* Golgi. Localization of these three marker proteins opens the door to colocalizaton experiments and other molecular work, and it represents the first step on the path to building a model system in *N. gruberi*.

7. Significance

This thesis has explored the evolution of the membrane trafficking system in four eukaryotic lineages of cell biological, medical, evolutionary and ecological importance. It has also illustrated the types of questions that can be addressed using genomics and transcriptomics, and how these approaches work synergistically to give a comprehensive understanding of the biology of these organisms and membrane trafficking diversity. This final section will discuss in a broad evolutionary context the findings of this thesis; conservation and novelty in the membrane trafficking machinery as detected by comparative genomics, how the trafficking systems of unrelated organisms have evolved to support similar lifestyles (i.e. parasitism/endobiosis, gut habitation, specialized secretion), the complexity of gene expression regulation with respect to specialized trafficking processes, and the utility of genomics and transcriptomics data in making and testing hypotheses about the cell biology of non-model eukaryotes.

7.1. Conservation and evolvability of membrane trafficking machinery in eukaryotes

In general, much of the membrane trafficking machinery is highly conserved in eukaryotes, and it is accepted that the LECA had a complex trafficking system. As more comparative genomic analyses are published, three patterns of conservation have been revealed: ubiquitous machinery, machinery with a limited, lineage-specific distribution, and machinery that is likely ancient but has been independently lost in multiple taxa (so-called "patchy proteins").[326] These latter proteins are most interesting, as this pattern implies that they are more 'evolvable' than highly conserved or lineage-specific machinery. The TSET complex and AP5 complex are two such examples of machinery with a patchy distribution; in fact, this led to such difficulty in identifying their components that they were only recently discovered. Both

complexes are part of the vesicle formation machinery, which, in general, is more conserved than vesicle fusion machinery, based on the findings of this thesis as well as other work. For example, COPI, COPII, clathrin, and retromer coat complexes are almost never lost,[196,337,468] with the exception of some reduced parasitic lineages. Adaptor protein complexes AP1 and AP2 are similarly retained (and are also the most recently evolved complexes),[144] while the more ancient AP4 is lost both in organisms at the base of Fungi and multiple times throughout this lineage, as well as in invertebrates and *Leishmania*,[175,712] and AP3 is lost in some members of the Apicomplexa, the green alga *Chlamydomonas reinhardtii*,[50] and in the haptophytes.[175,183] The ability to lose these complexes suggests that their function is compensated for by another trafficking factor, or that the endocytic/recycling pathway of these organisms is modified.

The ESCRTs are another set of machinery that could be defined as 'patchy', since although all ESCRT machinery is never lost, certain subunits or complexes are. In this thesis, ESCRTs I and II were shown to be missing from the haptophytes, similar to the losses observed in the Apicomplexa.[50] *Blastocystis* sp. and the gut parasite *Giardia intestinalis* have also lost ESCRT I.[174] Even in these taxa, certain subunits of ESCRTs III and IIIA are retained, suggesting that their function in cytokinesis and/or other cellular pathways is still required. The most well-conserved ESCRT III and IIIA subunits are Vps4 and Vps60, an AAA-type ATPase and protein that stimulates ATPase function through interaction with Vta1, respectively.[237,557] The ATPase function of Vps4 certainly raises the possibility of these proteins moonlighting in another cellular pathway. Meanwhile, at least two ESCRT subunits are frequently lost, even in organisms with relatively complete ESCRT complements. These are Vps37 and CHMP7. CHMP7 functions with ESCRT components in nuclear breakdown during cytokinesis, but its role in MVB biogenesis in unclear.[713] Vps37 aids in Vps23 binding to the membrane,[231] however, it is not clear what other

role(s) this protein plays in the ESCRT I complex. Because of the multi-functionality of this system, it is difficult to identify the minimal ESCRT complement required for MVB biogenesis.

Loss of various SNARE and multisubunit tethering complex components is generally more common than losses of vesicle formation machinery (with the exception of the ESCRTs). This is likely for two reasons. First, because the members of the Qa, b, c, and R SNARE subfamilies are closely related paralogous genes, the composition of SNARE complexes is not as prescribed as that of the COPI complex, for example. In yeast, there is evidence of SNARE promiscuity, where SNAREs can participate in multiple trafficking steps by incorporation into different complexes.[714] Secondly, while MTCs aid in vesicle capture prior to SNARE engagement, and some may be effector proteins of Rabs on incoming vesicles, they are not all required for vesicle fusion. There is at least one described example of an MTC complex with a patchy distribution in eukaryotes. As noted by Klinger *et al*. (2013), the Dsl1 complex is often missing or degraded in organisms that have unconventional peroxisomes.[480] This was observed in *Blastocystis* sp. in this thesis, and potentially also in *E. huxleyi*. Therefore Dsl1 loss may serve as a barometer of peroxisomal function; indeed, it was shown in the *Blastocystis* sp. genome paper that nearly all of the peroxisome biogenesis machinery is absent. Preliminary work has shown that *P. lacertae* encodes several additional peroxisome biogenesis genes, and *C. roenbergensis* even more. This same pattern is seen in the parasitic Apicomplexa and their free-living sister taxa, the Chromerids. The Apicomplexa have lost peroxisomes and Dsl1 function,[50,504] while the chromerids have retained Dsl1 and the peroxisome biogenesis machinery (F. Mast, unpublished). Loss of peroxisomes and Dsl1 are clearly correlated; however, the relationship between this pattern and organismal biology, such as parasitism or low-oxygen

environment is less clear. Understanding this will require more genomic sampling points, and an updated analysis of peroxisome biogenesis machinery in eukaryotes.

Several other patterns of MTC evolution were observed, however, these lack a functional association. First, the exocyst complex has been lost in both the haptophytes and Apicomplexa.[50] Exocyst is the plasma membrane-associated MTC that tethers incoming vesicles as part of exocytosis.[295] In mammalian cells, it is specifically involved in spatially regulating secretion.[715] It was therefore surprising to see a loss of exocyst in the haptophytes, which regularly secrete large scales. Apicomplexa, on the other hand, are intracellular parasites that depend on secretion for survival in the host cell, although they secrete the contents of lysosome-related organelles.[716] Despite the vastly different lifestyles of these two organisms, they are united by the fact that they have unique secretory requirements. In the case of Apicomplexa, exocyst loss may be possible because the secretion of LROs does not involve the TGN-plasma membrane pathway; rather, endosome/lysosome machinery is used for biogenesis and unknown machinery for interaction with the plasma membrane in "kiss-and-run" fusion.[489,717] On the other hand, in the haptophytes, secretion of the scale via the Golgi-derived coccolith vesicle may represent the 'default' secretion pathway, and may not require the specificity imparted by tethering machinery or may use haptophyte-specific machinery. Regardless of the details of secretion in the haptophytes and Apicomplexa, their shared whole-complex losses are intriguing. These lineages have independently lost some of the same complexes that are normally retained in eukaryotes (AP3, ESCRTs I and II, exocyst), and yet are not affected by the same lifestyle constraint (i.e. parasitism). This suggests that these membrane trafficking components are predisposed to loss, and their losses can co-occur (in the case of the apicomplexan *Cryptosporidium*), or be independent (e.g. *Toxoplasma*). In human cells, AP3 is associated with trafficking to lysosome-

related organelles,[488,718,719] although in its absence in *Toxoplasma*, trafficking to the rhoptries and micronemes occurs through AP1.[720] Overlapping function of related trafficking components may therefore allow loss. However, there is no functionally similar machinery that one could imagine taking the role of the ESCRTs or exocyst tether, raising the question of whether there is lineage-specific machinery that acts instead, and if not, how functional these pathways are.

The second example of a patchy MTC complex is TRAPPII (Figure 7.1). Similar to HOPS/CORVET, TRAPPII is made up of the TRAPPI complex, plus three additional subunits. In mammalian cells, TRAPPI and TRAPPII both function at the *cis*-Golgi, while in yeast and plants TRAPPII has a role in trafficking at the TGN and recycling endosome,[286,305,523] suggesting that this is its function in other eukaryotes. However, the Coulson plot in Figure 7.1 shows that TRAPPII-specific subunits are often missing in eukaryotes. The MTC GARP has a similar function at TGN-endosomes,[302] and yet is almost never lost. This suggests that GARP can fulfill the role of TRAPPII in many eukaryotes, or that – if TRAPPII and GARP have specific non-overlapping functions at the same location – GARP performs some typically essential function while TRAPPII does not. This raises the question of the specific functions of these MTCs in eukaryotes, despite their functioning at the same cellular locations.

7.2 Gene family expansion as a means of innovation; or, multicellularity is not the only driver of complexity

Rather than mere presence and absence, another avenue of membrane trafficking diversification is the expansion of various gene families. Complexes with multiple lineage-specific paralogues can generate novelty in that pathway, and can be thought of as an extension of the Organelle Paralogy Hypothesis. This hypothesis, put forward by Dacks and Field (2007),[85]

Figure 7.1 A compilation of comparative genomic analyses of multi-subunit tethering complexes across select eukaryotes.

Comparative genomic data taken from the chapters in this thesis, as well as Koumandou *et al.* (2007) and Woo *et al.* (2015).[50,254] This Coulson plot is not meant to demonstrate tethering evolution across eukaryotes, but rather to focus on the systems explored here and their close relatives. Boxes indicate the relative conservation of EARP/GARP, contrasted with the relative loss of TRAPPII.

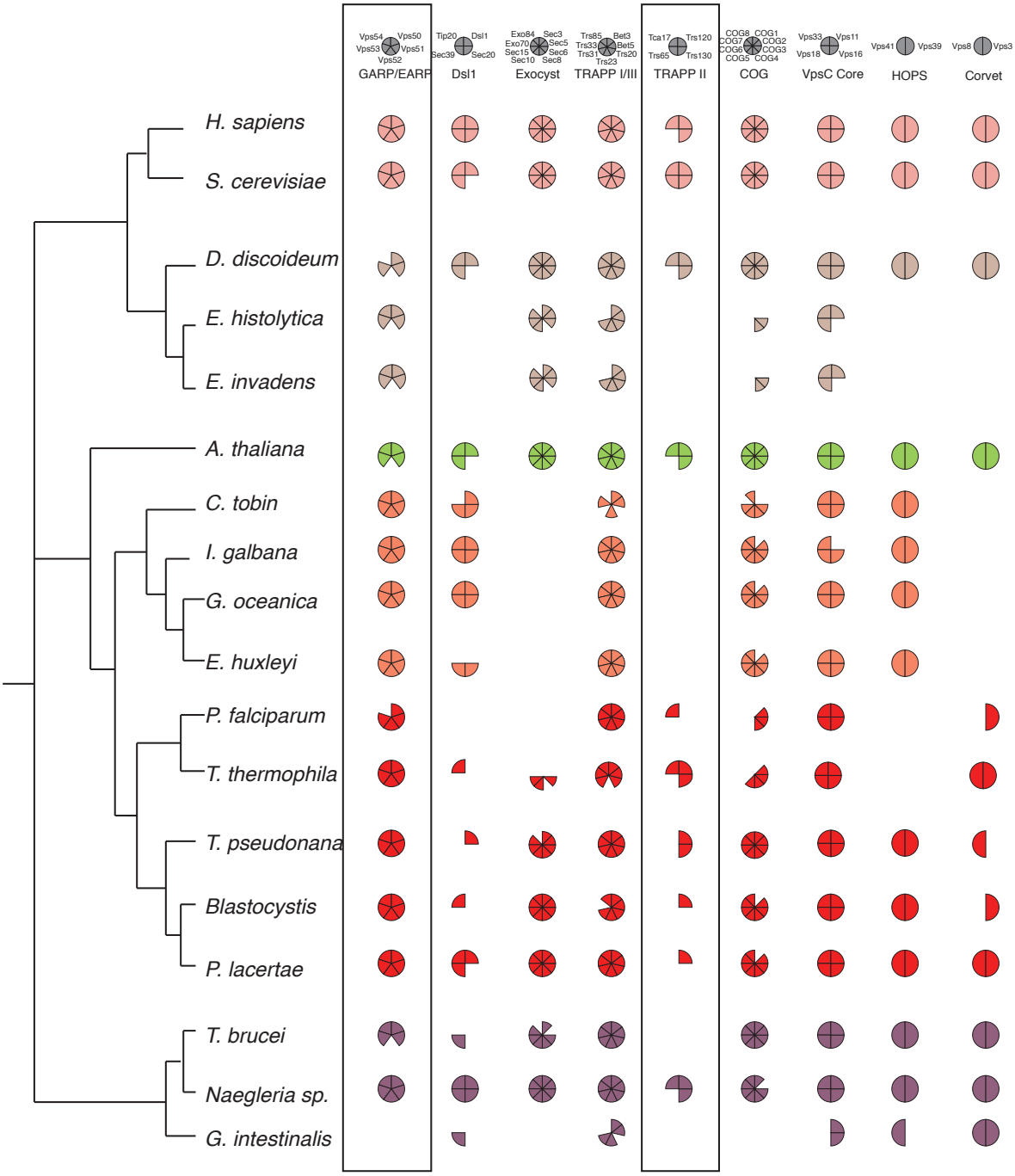

Vps54
Vps50
Vps53
Vps51
Vps52
GARP/EARP
Tip20
Dsl1
Sec39
Sec20
Dsl1
Exo84
Sec3
Exo70
Sec5
Sec15
Sec6
Sec10
Sec8
Exocyst
Trs85
Bet3
Trs33
Bet5
Trs31
Trs20
Trs23
TRAPP I/III
Tca17
Trs120
Trs65
Trs130
TRAPP II
COG8
COG1
COG7
COG2
COG6
COG3
COG5
COG4
COG
Vps33
Vps11
Vps18
Vps16
VpsC Core
Vps41
Vps39
HOPS
Vps8
Vps3
Corvet
H. sapiens
S. cerevisiae
D. discoideum
E. histolytica
E. invadens
A. thaliana
C. tobin
I. galbana
G. oceanica
E. huxleyi
P. falciparum
T. thermophila
T. pseudonana
Blastocystis
P. lacertae
T. brucei
Naegleria sp.
G. intestinalis

Figure 7.1

states that a core set of membrane trafficking proteins underwent duplication and co-evolution prior to the LECA, and are responsible for the biogenesis of different endomembrane organelles and trafficking pathways. This process can obviously continue in extant eukaryotic lineages, where evidence for membrane trafficking system novelty can be gleaned from comparative genomics. The majority of gene family expansions in the genomes studied in this thesis are in the endocytic system, notably AP4, the tethers GARP, HOPS, and CORVET, the SNAREs Syntaxin 6/10 and Vti1, and the ArfGAP AGFG. These proteins function in endocytosis, recycling, and endolysosomal degradation, and expansions in these factors may be relevant to novel trafficking pathways. The HOPS complex has been implicated in secretory lysosome formation in mammalian cells, and shown to specifically mediate trafficking to LROs in *C. elegans*.[487] Furthermore, it was recently shown that CORVET participates in a specialized secretory process in the ciliate *Tetrahymena thermophila*, and these subunits have been extensively duplicated.[721] This is one example of lineage-specific membrane trafficking evolution with clear functional relevance, and it raises the possibility that this phenomenon is wide-spread in eukaryotes (not just in relation to HOPS and CORVET, but potentially other trafficking factors). Interestingly, the vesicle fusion machinery appears to be more amenable to loss and duplication events, suggesting that there is more flexibility and therefore evolvability at this step of membrane trafficking.

There are also lineage-specific expansions that are likely to be functionally relevant, such as COPII in the haptophytes, which make large scales in Golgi-derived vesicles. An expansion of HOPS and CORVET machinery in *Blastocystis* may be related to the large central vacuole that occupies the majority of the cell's volume; these factors are not duplicated in the related *P. lacertae* that does not have a central vacuole. Massive expansion of the Rab family of GTPasees

in *Entamoeba* sp. has previously been published, and yet little is known of their cellular functions.[460] However, this finding mirrors the large Rab expansion in the human parasite *Trichomonas vaginalis*, which has a complement exceeding multicellular animals and plants.[722] So what can be made of the highly expanded secretory machinery in both *N. fowleri* and *N. gruberi*, which includes duplications of SyntaxinPM, VAMP-like proteins, and Exocyst? Perhaps they are related to novel function in the TGN-plasma membrane secretory pathway. These duplicates are found in both *N. fowleri* and *N. gruberi*, while they are not significantly differentially expressed in highly pathogenic *N. fowleri*, this does not preclude them from being involved in pathogenesis. SyntaxinPM duplications are typically lineage-specific, and have occurred independently in humans and plants.[246,723,724] It has been suggested that for both lineages, SNARE duplications play a role in the development of multicellularity. Multiple SyntaxinPM and VAMP-like proteins are found in all the organisms studied in this thesis, implying that SNARE subfamily expansions are not solely associated with multicellularity. It further raises the question of the complexity of the late secretory landscape in microbial eukaryotes, which has yet to be explored.

In general, expansions in these pathways strongly suggest that multicellularity is not the only evolutionary driver of complexity, and functional work in microbial eukaryotes is necessary to explore the functions of these paralogues.

7.3 Similar cellular behaviours with dissimilar underlying biology

Work presented in this thesis gives insight into how similar environmental pressures or lifestyles can have varying effects on the membrane trafficking systems of different eukaryotes. In terms of secretion, both *Entamoeba* sp. and the haptophytes have lifestyles that involve major

secretory events, specifically, encystation and scale secretion. Although cyst formation is a common eukaryotic response to adverse environmental conditions, transcriptomics of the encystation process pinpoints how the trafficking system supports this process. Haptophytes are not the only scale-secreting eukaryotes; members of both the Amoebozoa (arcellinids) and Rhizaria (euglyphids) secrete scales to form an outer shell. However, these are unrelated to the haptophytes, and therefore their mechanisms of scale secretion are not likely to be homologous. In comparing the transcriptomic findings of encystation and scale formation, there are clear similarities. Genes involved in early secretory processes are up-regulated in both taxa, as are endocytic recycling genes. At least in *E. huxleyi* and *G. oceanica*, this is concomitant with a down-regulation of endolysosomal degradation genes. One possible explanation for this is that receptor recycling is a downstream consequence of plasma membrane uptake as a homeostatic mechanism, rather than endocytosis as a feeding/environmental sampling mechanism. This is further supported by the up-regulation of endocytic genes in the haptophytes.

One interesting observation regarding the haptophyte genomic and transcriptomic data is the relationship between *E. huxleyi* and *G. oceanica*. The concatenated gene phylogeny shown here as well as work by others suggest that *G. oceanica* and *E. huxleyi* are not separate genera, but are actually more closely related.[545] An analysis of ultrastructure and life cycle show that these organisms are remarkably similar. However, their membrane trafficking system complements show key differences. While they have retained much of the same machinery, they often differ in paralogue number. The most obvious example is a duplication of all subunits of the retromer complex in *G. oceanica* that is not observed in *E. huxleyi*, although differences in paralogue number are found across the trafficking system in these organisms. Furthermore, most of the genes that are differentially expressed during biomineralizing conditions are different

between the two organisms. Despite this, the end result when focusing on pathways with up-regulated genes is similar. This raises questions about the closeness of *E. huxleyi* and *G. oceanica*, the relationship between genotype and phenotype, and the heritability of gene expression regulation, among other things. *E. huxleyi* has been shown to have a pan genome, which is thought to be responsible for its success in a wide variety of ocean environments.[62] Assuming a close relationship with *G. oceanica*, it may be the case that variability is also a feature of gene regulation, so long as trafficking pathway dynamics are modulated in a way that supports scale formation.

In both haptophytes and *E. invadens*, there are genes that are differentially expressed during the secretory events that encode multiple paralogues. This is most obvious in *E. invadens*, where ¼ of the ~400 membrane trafficking genes identified encode multiple paralogues where two or more were found in different subclusters. This suggests that both lineages employ surprisingly complex regulation of the membrane trafficking system.

7.4 Membrane trafficking system reduction is not a requirement of parasitism, nor is its conservation required in free-living taxa

One of the main questions addressed in this thesis is how the membrane trafficking system has been independently sculpted due to the pressures of parasitism and/or host gut association. Generally, parasite genomes are associated with streamlining because of the advantage it affords in terms of a lower energy requirement for replication. Two examples of this are *Giardia intestinalis* and the microsporidia,[68,175,725,726] although as more parasite genomes are sequenced, it appears that genome reduction is only one outcome of parasitism. *Entamoeba* sp.

are obligate parasites of animals, and have a relatively complete set of vesicle formation machinery, but have lost much of the vesicle fusion machinery that functions at the ER, Golgi, and involved in endo-lysosomal degradation. The latter finding was surprising, particularly because phagocytosis is a key mechanism of pathogenesis for *Entamoeba*.[727] This raises the question of how much novelty exists in the membrane trafficking systems of parasites; either in the sense of a homologous membrane trafficking protein that has diverged in sequence and function to the extent that it is not retrievable using even sensitive homology searching methods, or in the sense of truly novel factors that work in place of or interact with canonical trafficking proteins. In the case of the former, there is little that can be done to identify highly divergent sequences using informatics methods. However, both issues can be addressed by functional work in microbial eukaryotes, which again reinforces the importance of using both cell biological and *in silico* techniques together to inform our understanding of membrane trafficking function.

Novelty has been previously shown to be a feature of the membrane trafficking system of trypanosomes. In *T. brucei*, clathrin-mediated endocytosis occurs in the absence of the AP2 complex, due to the presence of novel, lineage-specific clathrin-interacting proteins called TbCAPs.[183] This is likely to also be the case in *Blastocystis* sp., which lacks almost all of the canonical autophagy machinery likely to be necessary for this process. Despite this, evidence of autophagosome biogenesis (albeit in the central vacuole)[486] suggests that this process occurs, and possibly with the help of lineage-specific factors. Lineage-specific adaptation may not only be a feature of parasites; any organism with membrane trafficking system losses could compensate for them in this way. The haptophytes are a likely candidate for this, as they are free-living, but have membrane trafficking system complement losses on par with apicomplexan intracellular parasites. They also have novel alpha and gamma adaptin ear proteins, as shown in this thesis.

While their function is unknown, they are likely haptophyte-specific. Even now, ancient membrane trafficking factors are still being discovered,[144,147,199] suggesting that there is still much to be learnt about general trafficking biology, let alone novel factors. This necessitates cell biological work in diverse model systems to identify both pan-eukaryotic and lineage-specific trafficking factors that are currently unknown.

Entamoeba, *Blastocystis* sp., and *P. lacertae* are all obligate gut endobionts, yet *Blastocystis* sp. and *P. lacertae* have remarkably complete membrane trafficking systems. Unlike in *Entamoeba* spp., SNAREs and MTCs are generally present in *Blastocystis* sp. and *P. lacertae*, and have undergone gene duplication events. This again supports the idea that reduction and simplification are not *de facto* qualities of parasite genomes. It also implies that similar environmental pressures (low oxygen, obligatory gut association) will not necessarily have the same selective effect on the evolution of two different organisms. One reason for this may be pre-adaptation; changes in trafficking machinery that pre-date the transition to parasitism. Unfortunately, there is no published genome of a close free-living sister taxon to *Entamoeba*, nor is there a genome available for *C. roenbergensis*; both of which would be necessary to determine if this is the case. Pre-adaptation has been put forward to explain the evolution of some trafficking machinery in the Apicomplexa (e.g. the loss of the exocyst complex).[480] Rather than being a result of parasitism, it is partially lost in the free-living *Perkinsus marinus*, and then fully lost in the Apicomplexa.[480] It is possible that pre-adaptive states (gene losses or gains) are common in parasitic genomes, and that reductive pressures are acting on an already-partial system, where loss events are then more likely. Again, sequencing more genomes of free-living relatives to parasites or other symbiotic eukaryotes may offer more insight into this scenario.

7.5 Understanding the biology of *N. fowleri*: from comparative genomics to pathogenicity-associated gene expression

It has always been clear that *N. fowleri* is a free-living amoeba that does not require a human host,[598] making it all the more curious how it is able to infect humans. *N. fowleri* and *N. gruberi* are highly similar in gene content, especially in particular systems that were thought to be involved in pathogenesis *a priori*, such as the membrane trafficking system. Because of this, comparative genomics alone would likely be insufficient to investigate pathogenesis, and it highlights the value of interpreting genomic data together with pathogenesis-associated gene expression data. Lysosomal function appears to be a key feature of pathogenesis, specifically cysteine protease production. At least one cysteine protease has been shown to be a secreted, tissue-degrading pathogenicity factor,[354,609] raising the possibility that more of the up-regulated cysteine proteases identified in this thesis are also secreted.

Other groups have identified factors such as the saposins as pathogenicity factors,[675,676,679,680] which were also retrieved in the differential expression analysis of this thesis. However, these proteins are also found in *N. gruberi*, and both harmless and parasitic eukaryotes across the tree. This is a common theme of potential pathogenicity factors in *N. fowleri*; none appear to be specific to the pathogen. This suggests several possibilities for what drives pathogenicity by this organism, which are not mutually exclusive. First, one or more pathogenicity factors may be uncharacterized proteins that are up-regulated during pathogenesis, with no conserved domains or other identifying information. Second, gene regulation differences between *N. fowleri* and *N. gruberi* may be responsible for permitting *N. fowleri* infection, for example, *N. fowleri* may have a higher base level of expression of a cell surface glycoprotein that improves adhesion when water is taken up into the nose, thus allowing it to infect. Third,

thermotolerance in *N. fowleri* may be key to infection, as *N. gruberi* is not able to survive at a host's body temperature. However, *in vitro* pathogenicity experiments show that this is not the only barrier to infection, as *N. gruberi* cannot kill tissue culture cells. These are only a few potential explanations for differences in pathogenesis, and in reality, it is likely a combination of these and other factors that allow *N. fowleri* to infect a host, and even become better at infecting hosts through multiple passages.

Although no "one true" pathogenicity factor could be identified, studying the transcriptomic response associated with pathogenesis has afforded insight into this process from a cell biological perspective. It allowed for hypotheses to be tested (e.g. is pathogenicity related to a stress response?), and generated a list of gene targets that can now be explored using cell biological techniques. The genomes of *N. fowleri* and *N. gruberi* were also invaluable, as they helped narrow the focus of pathogenicity question by ruling out those that are not likely to be explanations for pathogenicity. These datasets have generated a great deal of material to be explored, and they now serve to inform further cell biological work to understand how *N. fowleri* is so deadly.

7.6 Using comparative genomics and transcriptomics as a foundation for cell biological work in *Naegleria*

The final experiments to visualize the Golgi body in *N. gruberi* are the first functional studies of *Naegleria* membrane trafficking biology. From the analysis of the membrane trafficking system, it is clear that *Naegleria* encodes a highly conserved, LECA-like complement of trafficking machinery. The next step in understanding membrane trafficking in eukaryotes is

to address the problem of asymmetry; that much of our understanding comes from three model systems, two of which are within the same supergroup. While membrane trafficking has been studied in the related trypanosomatids, their parasitic lifestyle has left them with some membrane trafficking system adaptations. *N. gruberi* is easy to culture and manipulate, and provides another sampling point in this supergroup with a more conventional membrane trafficking complement. The work in this thesis lays the foundation for future cell biological work in the previously unexplored *Naegleria*, and its development as a new model system.

7.7 Final Thoughts

This thesis dealt primarily with the evolution of the membrane trafficking system in diverse eukaryotes. From both the analyses presented here and those by others, it seems very likely that there are aspects of membrane trafficking biology that are still undiscovered. While the basics of membrane trafficking function have been mapped out, the discoveries from our lab alone of TSET, AP5, and ArfGAPC2 proteins in only the last decade are reminders that we probably still do not have the full membrane trafficking story. TSET and ArfGAPC2, particularly, are reduced in or lost from opisthokonts, and are an example of a consequence of the 'asymmetry problem' caused by our focus on animal and yeast cell biology. Another recent example is that of the tepsins, which are accessory proteins of AP4.[728] Tepsins were at first thought to be animal-specific, but have recently been shown to be present across eukaryotes.[729] Further development of non-opisthokont model systems will be key to addressing the relative disparity in data contributing to our understanding of general eukaryotic cell biology. Another aspect of membrane trafficking biology that is worthy of future study is the evolution of lineage-specific losses, duplications or novelty, particularly when these features have arisen

independently in different taxa (e.g. adaptin ear domain proteins in haptophytes and GGA proteins in opisthokonts).

Traditionally, comparative genomics has relied on the use of complete genomes to make confident assessments of protein presence and absence. However, there is a wealth of transcriptomic data available through projects such as MMETSP (Marine Microbial Eukaryote Transcriptome Sequencing Project), which is largely unexplored with regard to comparative genomics of membrane trafficking. While transcriptomic datasets are likely not complete representations of the coding content of a genome, they offer a distinct advantage to genomes. Often, they are the first source of sequence data published for many organisms, as transcriptome assembly is less resource-consuming than genome assembly. This is particularly true for the Amoebozoa, in which nine species have associated transcriptomic data through MMETSP, but only a handful of fully sequenced genomes. These additional sequence data can help improve taxonomic sampling in key areas with few representative genomes, much as expressed sequence tag (EST) datasets did when sequencing a single genome was a major technical effort. While full genomes should remain the gold standard for making statements about gene presence and absence, the potential of transcriptomic data also should not be overlooked.

This raises the question of the future of microbial –omics in an era of inexpensive sequencing and incredible computational power. In addition to large-scale transcriptome sequencing projects, single cell genomic and transcriptomic sequencing is poised to become a powerful sequencing tool. Rather than averaging population data across supposedly homogenous cultures, single cell sequencing has the potential to give insight into diversity at the cellular level and a high-resolution look at population dynamics. However, –omics data are only one part of the story. No matter the volume or sophistication of *in silico* data generated, integration with

functional work is – and will continue to be – critical for understanding the biology of microbial eukaryotes.

When I began contributing to this field in 2009, there were still major gaps in sequenced genome representatives across the tree of eukaryotes. I have helped to fill some of those gaps, in analyzing the membrane trafficking systems of the first sequenced rhizarian organism, *Bigelowiella natans*, the first sequenced cryptophyte, *Guillardia theta*, and the first sequenced haptophyte, *Emiliania huxleyi*.[46,62] At this point, the focus of genome sequencing was shifting towards asking questions about eukaryotes with unique biology, which often included the synthesis of multiple types of –omics data to further inform our understanding of specialized cellular processes. As described in this thesis, my work has contributed to our understanding of the enigmatic parasite *Blastocystis* sp.,[466] the process of encystation in *Entamoeba* spp.,[501] and coccolith secretion in the haptophytes. Although it was not presented here, I analysed part of the membrane trafficking system of *Monocercomonoides* sp., the first eukaryote shown to completely lack any remnant of a mitochondrion.[66] Outside of genome projects, I have also focused my research on specific aspects of the membrane trafficking system, and investigated their evolution across eukaryotes. This has included studying the evolution of the various retromer cargo proteins across eukaryotes,[196] and a comprehensive analysis of the evolution of the Vps9 family of Rab GEF proteins (Herman *et al.* in preparation). My work has also included an analysis of the unfolded protein response in eukaryotes.

The work compiled herein is a major contribution to our understanding of the neuropathogenic amoeba *Naegleria fowleri*. In the decades since *N. fowleri*'s discovery, efforts have been made by a handful of labs to identify pathogenicity factors. Almost all of those found to date have been secreted or surface molecules, and despite these advances, the biology of *N.*

fowleri pathogenesis remained a ‘black box’. The work presented in this thesis is the first look at the overall gene expression program associated with selection for enhanced pathogenesis, revealing how many cellular processes are affected by or involved in pathogenesis. Importantly, this work has generated a long list of targets that can be functionally studied and potentially exploited therapeutically. Studying these factors further requires a closely related model organism. To this end, I aided in the molecular biological work to assess the cellular landscape of the non-pathogenic *Naegleria gruberi*, starting with localization of three markers of membrane trafficking system components. Although this work is foundational, it is a first step in establishing *N. gruberi* as a model system.

Overall, my work represents a major contribution to understanding the evolution of the membrane trafficking system across eukaryotes, and understanding the biology of pathogenesis in a deadly amoeba.

Bibliography

1. Embley, T. M. & Martin, W. Eukaryotic evolution, changes and challenges. *Nature* **440,** 623–630 (2006).

2. O'Malley, M. A. The first eukaryote cell: An unfinished history of contestation. *Stud. Hist. Philos. Sci. Part C Stud. Hist. Philos. Biol. Biomed. Sci.* **41,** 212–224 (2010).

3. Lane, N. & Martin, W. The energetics of genome complexity. *Nature* **467,** 929–934 (2010).

4. Forterre, P. A new fusion hypothesis for the origin of Eukarya: Better than previous ones, but probably also wrong. *Res. Microbiol.* **162,** 77–91 (2011).

5. Dacks, J. B. & Doolittle, W. F. Reconstructing/deconstructing the earliest eukaryotes: how comparative genomics can help. *Cell* **107,** 419–425 (2001).

6. Roger, A. J. Reconstructing Early Events in Eukaryotic Evolution. *Am. Nat.* **154,** S146–S163 (1999).

7. Chernikova, D., Motamedi, S., Csürös, M., Koonin, E. V & Rogozin, I. B. A late origin of the extant eukaryotic diversity: divergence time estimates using rare genomic changes. *Biol. Direct* **6,** 26 (2011).

8. Koumandou, V. L. *et al.* Molecular paleontology and complexity in the last eukaryotic common ancestor. *Crit Rev Biochem Mol Biol* **48,** 373–396 (2013).

9. Eme, L., Sharpe, S. C., Brown, M. W. & Roger, A. J. On the Age of Eukaryotes: Evaluating Evidence from Fossils and Molecular Clocks. *Cold Spring Harb. Perspect. Biol.* **6,** (2014).

10. Philippe, H., Germot, A. & Moreira, D. The new phylogeny of eukaryotes. *Curr. Opin. Genet. Dev.* **10,** 596–601 (2000).

11. Adl, S. M. *et al.* The revised classification of eukaryotes. *J. Eukaryot. Microbiol.* **59,** 429–93 (2012).

12. Radek, R. *et al.* Morphologic and molecular data help adopting the insect-pathogenic nephridiophagids (Nephridiophagidae) among the early diverging fungal lineages, close to the Chytridiomycota. *MycoKeys* **25,** 31–50 (2017).

13. Burki, F., Okamoto, N., Pombert, J.-F. & Keeling, P. J. The evolutionary history of haptophytes and cryptophytes: phylogenomic evidence for separate origins. *Proc. R. Soc. B Biol. Sci.* **279,** 2246–2254 (2012).

14. Kang, S. *et al.* Between a Pod and a Hard Test: The Deep Evolution of Amoebae. *Mol. Biol. Evol.* **34,** 2258–2270 (2017).

15. Krabberoed, A. K. *et al.* Single cell transcriptomics, mega-phylogeny and the genetic basis of morphological innovations in Rhizaria. *bioRxiv* **34,** 64030 (2016).

16. Derelle, R., López-García, P., Timpano, H. & Moreira, D. A Phylogenomic Framework to

Study the Diversity and Evolution of Stramenopiles (=Heterokonts). *Mol. Biol. Evol.* **33,** 2890–2898 (2016).

17. Keeling, P. J. & Campo, J. del. Marine Protists Are Not Just Big Bacteria. *Curr. Biol.* **27,** R541–R549 (2017).

18. Knoll, A. H. The Multiple Origins of Complex Multicellularity. *Annu. Rev. Earth Planet. Sci.* **39,** 217–239 (2011).

19. Walker, G., Dorrell, R. G., Schlacht, A. & Dacks, J. B. Eukaryotic systematics: a user's guide for cell biologists and parasitologists. *Parasitology* **138,** 1638–63 (2011).

20. Derelle, R. *et al.* Bacterial proteins pinpoint a single eukaryotic root. *Proc. Natl. Acad. Sci.* **112,** E693–E699 (2015).

21. He, D. *et al.* An alternative root for the eukaryote tree of life. *Curr. Biol.* **24,** 465–70 (2014).

22. Ruiz-Trillo, I. *et al.* The origins of multicellularity: a multi-taxon genome initiative. *Trends Genet.* **23,** 113–118 (2007).

23. Simakov, O. *et al.* Insights into bilaterian evolution from three spiralian genomes. *Nature* **493,** 526–531 (2012).

24. Dehal, P. *et al.* The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science* **298,** 2157–67 (2002).

25. James, T. Y. *et al.* Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature* **443,** 818–822 (2006).

26. Katinka, M. D. *et al.* Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi. *Nature* **414,** 450–453 (2001).

27. Corradi, N., Haag, K. L., Pombert, J.-F., Ebert, D. & Keeling, P. J. Draft genome sequence of the Daphnia pathogen Octosporea bayeri: insights into the gene content of a large microsporidian genome and a model for host-parasite interactions. *Genome Biol.* **10,** R106 (2009).

28. Cornman, R. S. *et al.* Genomic analyses of the microsporidian Nosema ceranae, an emergent pathogen of honey bees. *PLoS Pathog.* **5,** (2009).

29. Brown, M. W. *et al.* Phylogenomics demonstrates that breviate flagellates are related to opisthokonts and apusomonads Phylogenomics demonstrates that breviate flagellates are related to opisthokonts and apusomonads. (2013).

30. Clarke, M. *et al.* Genome of *Acanthamoeba castellanii* highlights extensive lateral gene transfer and early evolution of tyrosine kinase signaling. *Genome Biol.* **14,** R11 (2013).

31. Nagington, J. *et al.* Amoebic Infection of the Eye. *Lancet* **304,** 1537–1540 (1974).

32. Culbertson, C. G. Pathogenic Acanthamoeba (Hartmanella). *Am. J. Clin. Pathol.* **35,** 195–202 (1961).

33. Torno Jr., M. S., Babapour, R., Gurevitch, A. & Witt, M. D. *Cutaneous acanthamoebiasis*

in AIDS. **42,** (2000).

34. Galarza, C. *et al.* Cutaneous acanthamebiasis infection in immunocompetent and immunocompromised patients. *Int. J. Dermatol.* **48,** 1324–1329 (2009).

35. Eichinger, L. *et al.* The genome of the social amoeba Dictyostelium discoideum. *Nature* **435,** 43–57 (2005).

36. Eichinger, L. & Walker, J. M. *Dictyostelium discoideum Protocols*. **983,** (2013).

37. Loftus, B. *et al.* The genome of the protist parasite Entamoeba histolytica. *Nature* **433,** 865–868 (2005).

38. Stanley, S. L. Amoebiasis. *Lancet* **361,** 1025–1034 (2003).

39. Niklas, K. J. The evolutionary-developmental origins of multicellularity. *Am. J. Bot.* **101,** 6–25 (2014).

40. Umen, J. G. Green Algae and the Origins of Multicellularity in the Plant Kingdom. *Cold Spring Harb. Perspect. Biol.* **6,** 1–27 (2014).

41. Price, D. C. *et al.* Cyanophora paradoxa Genome Elucidates Origin of Photosynthesis in Algae and Plants. *Science (80-.).* **335,** 843–847 (2012).

42. Jain, K. *et al.* Extreme features of the Galdieria sulphuraria organellar genomes: A consequence of polyextremophily. *Genome Biol. Evol.* **7,** 367–380 (2014).

43. Matsuzaki, M. *et al.* Genome sequence of the ultrasmall unicellular red alga Cyanidioschyzon merolae 10D. *Nature* **428,** 653–657 (2004).

44. Prochnik, S. E. *et al.* Genomic analysis of organismal complexity in the multicellular green alga Volvox carteri. *Science (80-.).* **329,** 223–226

45. Douglas, S. E. & Penny, S. L. The plastid genome of the cryptophyte alga, Guillardia theta: Complete sequence and conserved synteny groups confirm its common ancestry with red algae. *J. Mol. Evol.* **48,** 236–244 (1999).

46. Curtis, B. A. *et al.* Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. *Nature* **492,** 59–65 (2012).

47. Schwelm, A. *et al.* The Plasmodiophora brassicae genome reveals insights in its life cycle and ancestry of chitin synthases. *Sci. Rep.* **5,** 11153 (2015).

48. Glöckner, G. *et al.* The genome of the foraminiferan reticulomyxa filosa. *Curr. Biol.* **24,** 11–18 (2014).

49. Saffo, M. B., McCoy, A. M., Rieken, C. & Slamovits, C. H. Nephromyces, a beneficial apicomplexan symbiont in marine animals. *Proc. Natl. Acad. Sci.* **107,** 16190–16195

50. Woo, Y. H. *et al.* Chromerid genomes reveal the evolutionary path from photosynthetic algae to obligate intracellular parasites. *Elife* **4,** 1–41 (2015).

51. Li, L. *et al.* The Symbiodinium kawagutii genome illuminates dinoflagellate gene expression and coral symbiosis. **350,** (2015).

52. Cock, J. M. *et al.* The Ectocarpus genome and the independent evolution of multicellularity in brown algae. *Nature* **465,** 617–621 (2010).

53. Leipe, D. D. *et al.* 16S-like rDNA sequences from Developayella elegans, Labyrinthuloides haliotidis, and Proteromonas lacertae confirm that the stramenopiles are a primarily heterotrophic group. *Eur. J. Protistol.* **32,** 449–458 (1996).

54. Cavalier-Smith, T. Sagenista and bigyra, two phyla of heterotrophic heterokont chromists. *Arch. für Protistenkd.* **148,** 253–267 (1997).

55. Cavalier-Smith, T. & Chao, E. E. Y. Phylogeny and megasystematics of phagotrophic heterokonts (kingdom Chromista). *J. Mol. Evol.* **62,** 388–420 (2006).

56. Cavalier-Smith, T. & Scoble, J. M. Phylogeny of Heterokonta: Incisomonas marina, a uniciliate gliding opalozoan related to Solenicola (Nanomonadea), and evidence that Actinophryida evolved from raphidophytes. *Eur. J. Protistol.* **49,** 328–353 (2013).

57. Armbrust, E. V. *et al.* The Genome of the Diatom. *Architecture* 79–86 (2004). doi:10.1126/science.1101156

58. Silberman, J. D., Sogin, M. L., Leipe, D. D. & Clark, C. G. Human parasite finds taxonomic home. *Nature* **380,** 398–398 (1996).

59. Wawrzyniak, I. *et al. Blastocystis* , an unrecognized parasite: an overview of pathogenesis and diagnosis. *Ther. Adv. Infect. Dis.* **1,** 167–178 (2013).

60. El Safadi, D. *et al.* Children of Senegal River Basin show the highest prevalence of Blastocystissp. ever observed worldwide. *BMC Infect. Dis.* **14,** 164 (2014).

61. Stiller, J. W. *et al.* The evolution of photosynthesis in chromist algae through serial endosymbioses. *Nat. Commun.* **5,** 5764 (2014).

62. Read, B. A. *et al.* Pan genome of the phytoplankton Emiliania underpins its global distribution. *Nature* **499,** 209–213

63. Hampl, V. *et al.* Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic 'supergroups'. *Proc. Natl. Acad. Sci.* **106,** 3859–3864 (2009).

64. Cavalier-Smith, T. The excavate protozoan phyla Metamonada Grassé emend. (Anaeromonadea, Parabasalia, Carpediemonas, Eopharyngia) and Loukozoa emend. (Jakobea, Malawimonas): Their evolutionary affinities and new higher taxa. *Int. J. Syst. Evol. Microbiol.* **53,** 1741–1758 (2003).

65. Hampl, V. *et al.* Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic 'supergroups'. *Proc. Natl. Acad. Sci.* **106,** 3859–3864 (2009).

66. Karnkowska, A. *et al.* A eukaryote without a mitochondrial organelle. *Curr. Biol.* **26,** 1274–1284 (2016).

67. Zubáčová, Z. *et al.* The Mitochondrion-Like Organelle of Trimastix pyriformis Contains the Complete Glycine Cleavage System. *PLoS One* **8,** 1–9 (2013).

68. Morrison, H. G. *et al.* Genomic minimalism in the early diverging intestinal parasite Giardia lamblia. *Science (80-.).* **317,** 1921–1926 (2007).

69. Newman, L. *et al.* Global Estimates of the Prevalence and Incidence of Four Curable Sexually Transmitted Infections in 2012 Based on Systematic Review and Global Reporting. *PLoS One* **10,** 1–17 (2015).

70. Carlton, J. M. *et al.* Draft genome sequence of the sexually transmitted pathogen Trichomonas vaginalis. *Science (80-.).* **315,** 207–212 (2007).

71. Trabelsi, H. *et al.* Pathogenic free-living amoebae: Epidemiology and clinical review. *Pathol. Biol.* **60,** 399–405 (2012).

72. Brinkmann, H. & Philippe, H. The diversity of eukaryotes and the root of the eukaryotic tree. *Adv. Exp. Med. Biol.* **607,** 20–37 (2007).

73. Roger, A. J. & Simpson, A. G. B. Evolution: revisiting the root of the eukaryotic tree. *Curr. Biol.* **9,** R165–R167 (2009).

74. Burki, F. The eukaryotic tree of life from a global phylogenomic perspective. *Cold Spring Harb. Perspect. Biol.* **6,** a016147 (2014).

75. Williams, T. A. Evolution: Rooting the eukaryotic tree of life. *Curr. Biol.* **24,** R151–R152 (2014).

76. Gouy, R., Baurain, D. & Philippe, H. Rooting the tree of life: the phylogenetic jury is still out. *Philos. Trans. R. Soc. B Biol. Sci.* **370,** 20140329 (2015).

77. Derelle, R. & Lang, B. F. Rooting the eukaryotic tree with mitochondrial and bacterial proteins. *Mol. Biol. Evol.* **29,** 1277–89 (2012).

78. Dykova, I., Fiala, I., Lom, J. & Lukes, J. Perkinsiella amoebae-like enclosymbionts of Neoparamoeba spp., relatives of the kinetoplastid Ichthyobodo. *Eur. J. Protistol.* **39,** 37–52 (2003).

79. Lukeš, J., Stensvold, C. R., Jirků-Pomajbíková, K. & Wegener Parfrey, L. Are Human Intestinal Eukaryotes Beneficial or Commensals? *PLoS Pathog.* **11,** 7–12 (2015).

80. Poirier, P., Wawrzyniak, I., Vivarès, C. P., Delbac, F. & El Alaoui, H. New insights into Blastocystis spp.: A potential link with irritable bowel syndrome. *PLoS Pathog.* **8,** 1–4 (2012).

81. Khan, N. A. Pathogenesis of Acanthamoeba infections. *Microb. Pathog.* **34,** 277–285 (2003).

82. Bonifacino, J. S. & Glick, B. S. The Mechanisms of Vesicle Budding and Fusion. **116,** 153–166 (2004).

83. Tooze, S. A. & Yoshimori, T. The origin of the autophagosomal membrane. *Nat. Cell Biol.* **12,** 831–835 (2010).

84. Kim, P. K., Mullen, R. T., Schumann, U. & Lippincott-Schwartz, J. The origin and maintenance of mammalian peroxisomes involves a de novo PEX16-dependent pathway

from the ER. *J. Cell Biol.* **173,** 521–532 (2006).

85. Dacks, J. B. & Field, M. C. Evolution of the eukaryotic membrane-trafficking system: origin, tempo, mode. *J. Cell Sci.* **120,** 2977–2985 (2007).

86. Palade, G. Intracellular Aspects of the Process of Protein Synthesis. *Science (80-.).* **189,** 867–867 (1975).

87. Hammond, C., Braakman, I. & Helenius, A. Role of N-linked oligosaccharide recognition, glucose trimming, and calnexin in glycoprotein folding and quality control. *Proc. Natl. Acad. Sci.* **91,** 913–917 (1994).

88. Fagone, P. & Jackowski, S. Membrane phospholipid synthesis and endoplasmic reticulum function. *J. Lipid Res.* **50,** S311–S316 (2009).

89. Vidugiriene, J. & Menon, A. K. Early lipid intermediates in glycosyl-phosphatidylinositol anchor assembly are synthesized in the ER and located in the cytoplasmic leaflet of the ER membrane bilayer. *J. Cell Biol.* **121,** 987–996 (1993).

90. Tveit, H., Akslen, L. K. A., Fagereng, G. L., Tranulis, M. A. & Prydz, K. A Secretory Golgi bypass route to the apical surface domain of epithelial MDCK cells. *Traffic* **10,** 1685–1695 (2009).

91. Suwastika, I. N., Uemura, T., Shiina, T., H Sato, M. & Takeyasu, K. SYP71, a plant-specific Qc-SNARE protein, reveals dual localization to the plasma membrane and the endoplasmic reticulum in Arabidopsis. *Cell Struct. Funct.* **33,** 185–192 (2008).

92. Golgi, C. *Sulla fina anatomia degli organi centrali del sistema nervoso*. *Tipi di Stefano Calderini e Compagno* (1885).

93. Fabene, P. F. & Bentivoglio, M. 1898-1998: Camillo Golgi and 'the Golgi': One hundred years of terminological clones. *Brain Res. Bull.* **47,** 195–198 (1998).

94. Dalton, A. J. & Felix, M. D. Cytologic and cytochemical characteristics of the Golgi substance of epithelial cells of the epididymis - in situ, in homogenates and after isolation. *Dev. Dyn.* **94,** 171–207 (1954).

95. Preuss, D., Mulholland, J., Franzusoff, A., Segev, N. & Botstein, D. Characterization of the Saccharomyces Golgl complex through the cell cycle by immunoelectron microscopy. *Trends Cell Biol.* **2,** 326 (1992).

96. Wooding, S. & Pelham, H. R. The dynamics of golgi protein traffic visualized in living yeast cells. *Mol. Biol. Cell* **9,** 2667–80 (1998).

97. Chávez-Munguia, B., Espinosa-Cantellano, M., Castañón, G. & Martínez-Palomo, A. Ultrastructural evidence of smooth endoplasmic reticulum and Golgi-like elements in Entamoeba histolytica and Entamoeba dispar. *Arch. Med. Res.* **31,** S165–S167 (2000).

98. Stefanic, S. *et al.* Neogenesis and maturation of transient Golgi-like cisternae in a simple eukaryote. *J. Cell Sci.* **122,** 2846–2856 (2009).

99. Lee, I. *et al.* Membrane adhesion dictates Golgi stacking and cisternal morphology. *Proc. Natl. Acad. Sci.* **111,** 1849–1854 (2014).

100. Mani, S. & Thattai, M. Stacking the odds for golgi cisternal maturation. *Elife* **5,** 1–16 (2016).

101. Papanikou, E., Day, K. J., Austin, J. & Glick, B. S. COPI selectively drives maturation of the early Golgi. *Elife* **4,** 1–33 (2015).

102. Northcote, D. H. & Pickett-Heaps, J. D. A function of the Golgi apparatus in polysaccharide synthesis and transport in the root-cap cells of wheat. *Biochem. J.* **98,** 159–167 (1966).

103. Roth, J., Taatjes, D. J., Lucocq, J. M., Weinstein, J. & Paulson, J. C. Demonstration of an extensive trans -tubular network continous with the golgi apparatus stack that may function in glycosylation. *Cell* **43,** 287–295 (1985).

104. Griffiths, G. & Simons, K. The trans Golgi network: sorting at the exit site of the Golgi complex. *Science (80-.).* **234,** 438–443 (1986).

105. Holtzman, E., Novikoff, A. B. & Villaverde, H. Lysosomes and GERL in Normal and Chromatolytic Neurons of the Rat Ganglion Nodosum. *J. Cell Biol.* **33,** 419–435 (1967).

106. White, J., Keller, P. & Stelzer, E. H. Spatial partitioning of secretory cargo from Golgi resident proteins in live cells. *BMC Cell Biol.* **2,** 19 (2001).

107. Ladinsky, M. S., Wu, C. C., McIntosh, S., McIntosh, J. R. & Howell, K. E. Structure of the Golgi and Distribution of Reporter Molecules at 20C Reveals the Complexity of the Exit Compartments. *Mol. Biol. Cell* **13,** 2810–2825 (2002).

108. Gleeson, P. A., Lock, J. G., Luke, M. R. & Stow, J. L. Domains of the TGN: Coats, tethers and G proteins. *Traffic* **5,** 315–326 (2004).

109. Dell'Angelica, E. C., Mullins, C., Caplan, S. & Bonifacino, J. S. Lysosome-related organelles. **14,** 1265–1278 (2000).

110. Ngô, H. M., Yang, M. & Joiner, K. A. Are rhoptries in Apicomplexan parasites secretory granules or secretory lysosomal granules? *Mol. Microbiol.* **52,** 1531–1541 (2004).

111. Docampo, R., Jimenez, V., Lander, N., Li, Z. H. & Niyogi, S. New insights into roles of acidocalcisomes and contractile vacuole complex in osmoregulation in protists. *Int. Rev. Cell Mol. Biol.* **305,** 69–113 (2013).

112. Hille-Rehfeld, A. Mannose 6-phosphate receptors in sorting and transport of lysosomal enzymes. *Biochim. Biophys. Acta* **1241,** 177–94 (1995).

113. Jovic, M., Sharma, M., Rahajeng, J. & Caplan, S. The early endosome: a busy sorting station for proteins at the crossroads. *Histol. Histopathol.* **25,** 99–112 (2010).

114. Wong, C. O. *et al.* Lysosomal Degradation Is Required for Sustained Phagocytosis of Bacteria by Macrophages. *Cell Host Microbe* **21,** 719–730.e6 (2017).

115. Grant, B. D. & Donaldson, J. G. Pathways and mechanisms of endocytic recycling. *Nat. Rev. Mol. Cell Biol.* **10,** 597–608 (2009).

116. Choudhury, A., Sharma, D. K., Marks, D. L. & Pagano, R. E. Elevated Endosomal

Cholesterol Levels in Niemann-Pick Cells Inhibit Rab4 and Perturb Membrane Recycling. *Mol. Biol. Cell* **15,** 4500–4511 (2004).

117. Zerial, M. & McBride, H. Rab proteins as membrane organizers. *Nat. Rev. Mol. Cell Biol.* **2,** 107–117 (2001).

118. Tooze, J. & Hollinshead, M. Tubular early endosomal networks in AtT20 and other cells. *J. Cell Biol.* **115,** 635–653 (1991).

119. Mallard, F. *et al.* Direct pathway from early/recycling endosomes to the Golgi apparatus revealed through the study of Shiga toxin B-fragment transport. *J. Cell Biol.* **143,** 973–990 (1998).

120. Ghosh, R. N., Mallet, W. G., Soe, T. T., McGraw, T. E. & Maxfield, F. R. An endocytosed TGN38 chimeric protein is delivered to the TGN after trafficking through the endocytic recycling compartment in CHO cells. *J. Cell Biol.* **142,** 923–936 (1998).

121. Rink, J., Ghigo, E., Kalaidzidis, Y. & Zerial, M. Rab conversion as a mechanism of progression from early to late endosomes. *Cell* **122,** 735–749 (2005).

122. Clague, M. J., Urbé, S., Aniento, F. & Gruenberg, J. Vacuolar ATPase activity is required for endosomal carrier vesicle formation. *J. Biol. Chem.* **269,** 21–24 (1994).

123. van Weering, J. R. T., Verkade, P. & Cullen, P. J. SNX-BAR-mediated endosome tubulation is co-ordinated with endosome maturation. *Traffic* **13,** 94–107 (2012).

124. Huotari, J. & Helenius, A. Endosome maturation. *EMBO J.* **30,** 3481–3500 (2011).

125. Mellman, I., Fuchs, R. & Helenius, A. Acidification of the Endocytic and Exocytic Pathways. *Annu. Rev. Biochem.* **55,** 663–700 (1986).

126. Braulke, T. & Bonifacino, J. S. Sorting of lysosomal proteins. *Biochim. Biophys. Acta - Mol. Cell Res.* **1793,** 605–614 (2009).

127. Bright, N. A., Gratian, M. J. & Luzio, J. P. Endocytic Delivery to Lysosomes Mediated by Concurrent Fusion and Kissing Events in Living Cells. *Curr. Biol.* **15,** 360–365 (2005).

128. Luzio, J. P., Pryor, P. R. & Bright, N. A. Lysosomes: fusion and function. *Nat. Rev. Mol. Cell Biol.* **8,** 622–632 (2007).

129. Ruiz, F. A., Lea, C. R., Oldfield, E. & Docampo, R. Human platelet dense granules contain polyphosphate and are similar to acidocalcisomes of bacteria and unicellular eukaryotes. *J. Biol. Chem.* **279,** 44250–44257 (2004).

130. Seufferheld, M. *et al.* Identification of organelles in bacteria similar to acidocalcisomes of unicellular eukaryotes. *J. Biol. Chem.* **278,** 29971–29978 (2003).

131. Plattner, H. The contractile vacuole complex of protists – New cues to function and biogenesis. *Crit. Rev. Microbiol.* **41,** 218–227 (2015).

132. Klinger, C. M. *et al.* Resolving the homology—function relationship through comparative genomics of membrane-trafficking machinery and parasite cell biology. *Mol. Biochem. Parasitol.* **209,** 88–103 (2016).

133. Barlow, L. D. & Dacks, J. B. Seeing the endomembrane system for the trees: Evolutionary analysis highlights the importance of plants as models for eukaryotic membrane-trafficking. *Semin. Cell Dev. Biol.* (2017). doi:10.1016/j.semcdb.2017.09.027

134. Tooze, S. A., Weiss, U. & Huttner, W. B. Requirement for GTP hydrolysis in the formation of secretory vesicles. *Lett. to Nat.* **347,** 207–208 (1990).

135. Antonny, B. *et al.* Membrane curvature and the control of GTP hydrolysis in Arf1 during COPI vesicle formation. *Biochem. Soc. Trans.* **33,** 619–622 (2005).

136. Guo, Y., Punj, V., Sengupta, D. & Linstedt, A. D. Coat-Tether Interaction in Golgi Organization. *Mol. Biol. Cell* **19,** 2830–2843 (2008).

137. Trahey, M. & Hay, J. C. Transport vesicle uncoating: it's later than you think. *F1000 Biol. Rep.* **2,** 47 (2010).

138. Dacks, J. B. & Robinson, M. S. Outerwear through the ages: evolutionary cell biology of vesicle coats. *Curr. Opin. Cell Biol.* **47,** 108–116 (2017).

139. Devos, D. *et al.* Components of coated vesicles and nuclear pore complexes share a common molecular architecture. *PLoS Biol.* **2,** (2004).

140. Klinger, C. M., Spang, A., Dacks, J. B. & Ettema, T. J. G. Tracing the Archaeal Origins of Eukaryotic Membrane-Trafficking System Building Blocks. *Mol. Biol. Evol.* **33,** 1528–1541 (2016).

141. Field, M. C., Sali, A. & Rout, M. P. On a bender-BARs, ESCRTs, COPs, and finally getting your coat. *J. Cell Biol.* **193,** 963–972 (2011).

142. Robinson, M. S. & Bonifacino, J. S. Adaptor-related proteins. *Curr. Opin. Cell Biol.* **13,** 444–453 (2001).

143. Beck, R., Ravet, M., Wieland, F. T. & Cassel, D. The COPI system: Molecular mechanisms and function. *FEBS Lett.* **583,** 2701–2709 (2009).

144. Hirst, J. *et al.* Characterization of TSET, an ancient and widespread membrane trafficking complex. *Elife* **2014,** 1–18 (2014).

145. Kural, C. *et al.* Dynamics of Intracellular Clathrin/AP1- and Clathrin/AP3-Containing Carriers. *Cell Rep.* **2,** 1111–1119 (2012).

146. Robinson, M. S. 100-kD coated vesicle proteins: Molecular heterogeneity and intracellular distribution studied with monoclonal antibodies. *J. Cell Biol.* **104,** 887–895 (1987).

147. Hirst, J. *et al.* The fifth adaptor protein complex. *PLoS Biol* **9,** e1001170 (2011).

148. Shi, G., Faúndez, V., Roos, J., DelL'Angelica, E. C. & Kelly, R. B. Neuroendocrine synaptic vesicles are formed in vitro by both clathrin- dependent and clathrin-independent pathways. *J. Cell Biol.* **143,** 947–955 (1998).

149. Simpson, F. *et al.* A novel adaptor-related protein complex. *J. Cell Biol.* **133,** 749–760 (1996).

150. Van Damme, D. *et al.* Adaptin-like protein TPLATE and clathrin recruitment during plant

somatic cytokinesis occurs via two distinct pathways. *Proc. Natl. Acad. Sci.* **108,** 615–620 (2011).

151. Wilfling, F. *et al.* Arf1/COPI machinery acts directly on lipid droplets and enables their connection to the ER for protein targeting. *Elife* **2014,** 1–20 (2014).

152. Simon, J. P., Ivanov, I. E., Adesnik, M. & Sabatini, D. D. In vitro generation from the trans-Golgi network of coatomer-coated vesicles containing sialylated vesicular stomatitis virus-G protein. *Methods* **20,** 437–454 (2000).

153. Ahn, H.-K., Kang, Y. W., Lim, H. M., Hwang, I. & Pai, H.-S. Physiological Functions of the COPI Complex in Higher Plants. *Mol. Cells* **38,** 866–75 (2015).

154. Lee, S. Y., Yang, J. S., Hong, W., Premont, R. T. & Hsu, V. W. ARFGAP1 plays a central role in coupling COPI cargo sorting with vesicle formation. *J. Cell Biol.* **168,** 281–290 (2005).

155. Randazzo, P. A. & Kahn, R. A. GTP hydrolysis by ADP-robosylation factor is dependent on both an ADP-ribosylation factor GTAase activating protein and acid phospholipids. *J. Biol. Chem.* **269,** 10758–10763 (1994).

156. Dingt, M. *et al.* Characterization of a GTPase-activating protein that stimulates GTP hydrolysis by both ADP-ribosylation factor (ARF) and ARF-like proteins. Comparison to the ARD1 and GAP domain. *J. Biol. Chem.* **271,** 24005–24009 (1996).

157. Spang, A., Shiba, Y. & Randazzo, P. A. Arf GAPs: Gatekeepers of vesicle generation. *FEBS Lett.* **584,** 2646–2651 (2010).

158. Morinaga, N., Tsai, S. C., Moss, J. & Vaughan, M. Isolation of a brefeldin A-inhibited guanine nucleotide-exchange protein for ADP ribosylation factor (ARF) 1 and ARF3 that contains a Sec7-like domain. *Proc. Natl. Acad. Sci. U. S. A.* **93,** 12856–60 (1996).

159. Peyroche, A., Paris, S. & Jackson, C. L. Nucleotide exchange on ARF mediated by yeast Geal protein. *Nature* **384,** 479–481 (1996).

160. Casanova, J. E. Regulation of Arf activation: The Sec7 family of guanine nucleotide exchange factors. *Traffic* **8,** 1476–1485 (2007).

161. Baumgartner, F., Wiek, S., Paprotka, K., Zauner, S. & Lingelbach, K. A point mutation in an unusual Sec7 domain is linked to brefeldin A resistance in a Plasmodium falciparum line generated by drug selection. *Mol. Microbiol.* **41,** 1151–1158 (2001).

162. Rothman, J. E. The protein machinery of vesicle budding and fusion. *Protein Sci.* **5,** 185–194 (2008).

163. Zink, S., Wenzel, D., Wurm, C. A. & Schmitt, H. D. A Link between ER Tethering and COP-I Vesicle Uncoating. *Dev. Cell* **17,** 403–416 (2009).

164. Zwiewka, M. *et al.* The AP-3 adaptor complex is required for vacuolar function in Arabidopsis. *Cell Res.* **21,** 1711–1722 (2011).

165. Stepp, J. D., Huang, K. & Lemmon, S. K. The yeast adaptor protein complex, AP-3, is essential for the efficient delivery of alkaline phosphatase by the alternate pathway to the

vacuole. *J. Cell Biol.* **139,** 1761–1774 (1997).

166. Cowles, C. R., Odorizzi, G., Payne, G. S. & Emr, S. D. The AP-3 adaptor complex is essential for cargo-selective transport to the yeast vacuole. *Cell* **91,** 109–118 (1997).

167. Simmen, T., Honing, S., Icking, A., Tikkanen, R. & Hunziker, W. AP-4 binds basolateral signals and participates in basolateral sorting in epithelial MDCK cells. *Nat.Cell Biol.* **4,** 154–159 (2002).

168. Aguilar, R. C. *et al.* Signal-binding Specificity of the μ4 Subunit of the Adaptor Protein Complex AP-4. *J. Biol. Chem.* **276,** 13145–13152 (2001).

169. Hirst, J., Irving, C. & Borner, G. H. H. Adaptor Protein Complexes AP-4 and AP-5: New Players in Endosomal Trafficking and Progressive Spastic Paraplegia. *Traffic* **14,** 153–164 (2013).

170. Lewin, D. A. *et al.* Cloning , expression , and localization of a novel gamma-adaptin-like molecule. *FEBS Lett.* **435,** 263–268 (1998).

171. Takatsu, H., Sakurai, M., Shin, H. W., Murakami, K. & Nakayama, K. Identification and characterization of novel clathrin adaptor-related proteins. *J. Biol. Chem.* **273,** 24693–24700 (1998).

172. Ahle, S. & Ungewickell, E. Purification and properties of a new clathrin assembly protein. *EMBO J.* **5,** 3143–3149 (1986).

173. Berriman, M. The Genome of the African Trypanosome Trypanosoma brucei. *Science (80-.).* **309,** 416–422 (2005).

174. Field, M. C., Gabernet-Castello, C. & Dacks, J. B. in *Eukaryotic Membranes and Cytoskeleton. Advances in Experimental Medicine and Biology* **607,** 84–96 (Springer New York, 2007).

175. Barlow, L. D., Dacks, J. B. & Wideman, J. G. From all to (nearly) none: Tracing adaptin evolution in Fungi. *Cell. Logist.* **4,** e28114 (2014).

176. Kuehn, M. J., Herrmann, J. M. & Schekman, R. COPII-cargo interactions direct protein sorting into ER-derived transport vesicles. *Nature* **391,** 187–190 (1998).

177. Miller, E. A. *et al.* Multiple cargo binding sites on the COPII subunit Sec24p ensure capture of diverse membrane proteins into transport vesicles. *Cell* **114,** 497–509 (2003).

178. Wendland, B., Steece, K. E. & Emr, S. D. Yeast epsins contain an essential N-terminal ENTH domain, bind clathrin and are required for endocytosis. *EMBO J.* **18,** 4383–4393 (1999).

179. Itoh, T. Role of the ENTH Domain in Phosphatidylinositol-4,5-Bisphosphate Binding and Endocytosis. *Science (80-.).* **291,** 1047–1051 (2001).

180. Hirst, J., Motley, A., Harasaki, K., Chew, S. Y. P. & Robinson, M. S. EpsinR: an ENTH Domain-containing Protein that Interacts with AP-1. *Mol. Biol. Cell* **14,** 625–641 (2003).

181. Gabernet-Castello, C., Dacks, J. B. & Field, M. C. The single ENTH-domain protein of

Trypanosomes; endocytic functions and evolutionary relationship with Epsin. *Traffic* **10,** 894–911 (2009).

182. Carbone, R. *et al.* epslS and epslSR Are Essential Components of the Endocytic Pathway1. *Science (80-.).* 5498–5504 (1997).

183. Manna, P. T. *et al.* Lineage-specific proteins essential for endocytosis in trypanosomes. *J. Cell Sci.* **130,** 1379–1392 (2017).

184. Damke, H., Baba, T., Warnock, D. E. & Schmid, S. L. Induction of mutant dynamin specifically blocks endocytic coated vesicle formation. *J. Cell Biol.* **127,** 915–934 (1994).

185. Marks, B. *et al.* GTPase activity of dynamin and resulting conformation change are essential for endocytosis. *Nature* **410,** 231–235 (2001).

186. Elde, N. C., Morgan, G., Winey, M., Sperling, L. & Turkewitz, A. P. Elucidation of clathrin-mediated endocytosis in tetrahymena reveals an evolutionarily convergent recruitment of dynamin. *PLoS Genet.* **1,** (2005).

187. Mills, I. G. *et al.* Epsinr: An AP1/clathrin interacting protein involved in vesicle trafficking. *J. Cell Biol.* **160,** 213–222 (2003).

188. Stamnes, M. A. & Rothman, J. E. The binding of AP-1 clathrin adaptor particles to Golgi membranes requires ADP-ribosylation factor, a small GTP-binding protein. *Cell* **73,** 999–1005 (1993).

189. Price, H. P., Stark, M. & Smith, D. F. Trypanosoma brucei ARF1 plays a central role in endocytosis and Golgi-lysosome trafficking. *Mol. Biol. Cell* **18,** 864–873 (2007).

190. Seaman, M. N. J. Cargo-selective endosomal sorting for retrieval to the Golgi requires retromer. *J. Cell Biol.* **165,** 111–122 (2004).

191. Arighi, C. N., Harmell, L. M., Aguilar, R. C., Haft, C. R. & Bonifacino, J. S. Role of the mammalian retromer in sorting of the cation-independent mannose 6-phosphate receptor. *J. Cell Biol.* **165,** 123–133 (2004).

192. Collins, B. M., Skinner, C. F., Watson, P. J., Seaman, M. N. J. & Owen, D. J. Vps29 has a phosphoesterase fold that acts as a protein interaction scaffold for retromer assembly. *Nat. Struct. Mol. Biol.* **12,** 594–602 (2005).

193. Nothwehr, S. F., Ha, S. A. & Bruinsma, P. Sorting of yeast membrane proteins into an endosome-to-Golgi pathway involves direct interaction of their cytosolic domains with Vps35p. *J. Cell Biol.* **151,** 297–309 (2000).

194. Horazdovsky, B. F. *et al.* A sorting nexin-1 homologue, Vps5p, forms a complex with Vps17p and is required for recycling the vacuolar protein-sorting receptor. *Mol. Biol. Cell* **8,** 1529–41 (1997).

195. Carlton, J. *et al.* Sorting Nexin-1 Mediates Tubular Endosome-to-TGN Transport through Coincidence Sensing of High- Curvature Membranes and 3-Phosphoinositides. *Curr. Biol.* **14,** 1791–1800 (2004).

196. Koumandou, V. L. *et al.* Evolutionary reconstruction of the retromer complex and its

function in Trypanosoma brucei. *J. Cell Sci.* **124,** 1496–1509 (2011).

197. Arlt, H., Reggiori, F. & Ungermann, C. Retromer and the dynamin Vps1 cooperate in the retrieval of transmembrane proteins from vacuoles. *J. Cell Sci.* **128,** 645–55 (2015).

198. Vater, C. A., Raymond, C. K., Ekena, K., Howald-Stevenson, I. & Stevens, T. H. The VPS1 protein, a homolog of dynamin required for vacuolar protein sorting in Saccharomyces cerevisiae, is a GTPase with two functionally separable domains. *J. Cell Biol.* **119,** 773–786 (1992).

199. Schlacht, A., Mowbrey, K., Elias, M., Kahn, R. A. & Dacks, J. B. Ancient complexity, opisthokont plasticity, and discovery of the 11th subfamily of Arf GAP proteins. *Traffic* **14,** 636–649 (2013).

200. Kahn, R. A. *et al.* Consensus nomenclature for the human ArfGAP domain-containing proteins. *J. Cell Biol.* **182,** 1039–1044 (2008).

201. Kartberg, F. *et al.* ARFGAP2 and ARFGAP3 are essential for COPI coat assembly on the Golgi membrane of living cells. *J. Biol. Chem.* **285,** 36709–36720 (2010).

202. Poon, P. P. *et al.* Retrograde transport from the yeast Golgi is mediated by two ARF GAP proteins with overlapping function. *EMBO J.* **18,** 555–564 (1999).

203. Tanabe, K. *et al.* Involvement of a novel ADP-ribosylation factor GTPase-activating protein, SMAP, in membrane trafficking: Implications in cancer cell biology. *Cancer Sci.* **97,** 801–806 (2006).

204. Natsume, W. *et al.* SMAP2, a Novel ARF GTPase-activating Protein, Interacts with Clathrin and Clathrin Assembly Protein and Functions on the AP-1–positive Early Endosome/Trans-Golgi Network. *Mol. Biol. Cell* **17,** 2592–2603 (2006).

205. Chaineau, M., Danglot, L., Proux-Gillardeaux, V. & Galli, T. Role of HRB in clathrin-dependent endocytosis. *J. Biol. Chem.* **283,** 34365–34373 (2008).

206. Li, J. *et al.* An ACAP1-containing clathrin coat complex for endocytic recycling. *J. Cell Biol.* **178,** 453–464 (2007).

207. Inoue, H. & Randazzo, P. A. Arf GAPs and their interacting proteins. *Traffic* **8,** 1465–1475 (2007).

208. Venkateswarlu, K., Brandom, K. G. & Lawrence, J. L. Centaurin-α1 Is an in Vivo Phosphatidylinositol 3,4,5-Trisphosphate-dependent GTPase-activating Protein for ARF6 That Is Involved in Actin Cytoskeleton Organization. *J. Biol. Chem.* **279,** 6205–6208 (2004).

209. Randazzo, P. A., Inoue, H. & Bharti, S. Arf GAPs as regulators of the actin cytoskeleton. *Biol. Cell* **99,** 583–600 (2007).

210. Richardson, B. C., McDonold, C. M. & Fromme, C. J. The Sec7 Arf-GEF Is Recruited to the trans-Golgi Network by Positive Feedback. *Dev. Cell* **22,** 799–810 (2012).

211. Manolea, F., Claude, A., Chun, J., Rosas, J. & Melancon, P. Distinct Functions for Arf Guanine Nucleotide Exchange Factors at the Golgi Complex: GBF1 and BIGs Are

Required for Assembly and Maintenance of the Golgi Stack and trans-Golgi Network, Respectively. *Mol. Biol. Cell* **19,** 523–535 (2008).

212. Frank, S., Upender, S., Hansen, S. H. & Casanova, J. E. ARNO is a guanine nucleotide exchange factor for ADP-ribosylation factor 6. *J. Biol. Chem.* **273,** 23–27 (1998).

213. Bell, A. J. *et al.* GEF1 is a Ciliary Sec7 GEF of Tetrahymena thermophila. *Cell Motil. Cytoskelet.* **66,** 483–499 (2009).

214. Naramoto, S. *et al.* ADP-ribosylation factor machinery mediates endocytosis in plant cells. *Proc. Natl. Acad. Sci.* **107,** 21890–21895 (2010).

215. Geldner, N. *et al.* The Arabidopsis GNOM ARF-GEF mediates endosomal recycling, auxin transport, and auxin-dependent plant growth. *Cell* **112,** 219–230 (2003).

216. Piper, R. C. & Katzmann, D. J. Biogenesis and Function of Multivesicular Bodies. *Annu. Rev. Cell Dev. Biol.* **23,** 519–547 (2007).

217. Gruenberg, J. & Stenmark, H. The biogenesis of multivesicular endosomes. *Nat. Rev. Mol. Cell Biol.* **5,** 317–323 (2004).

218. Hurley, J. H. ESCRT complexes and the biogenesis of multivesicular bodies. *Curr Opin Cell Biol* **20,** 4–11 (2008).

219. Carlton, J. G. & Martin-Serrano, J. Parallels Between Cytokinesis and Retroviral Budding: A Role for the ESCRT Machinery. *Science (80-.).* **316,** 1908–1912 (2007).

220. Lindås, A.-C., Karlsson, E. a, Lindgren, M. T., Ettema, T. J. G. & Bernander, R. A unique cell division machinery in the Archaea. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 18942–6 (2008).

221. Morita, E. *et al.* Human ESCRT and ALIX proteins interact with proteins of the midbody and function in cytokinesis. *EMBO J.* **26,** 4215–4227 (2007).

222. Samson, R. Y., Obita, T., Freund, S. M., Williams, R. L. & Bell, S. D. Division in Archaea. 1710–1713 (2008).

223. Spitzer, C. *et al.* The Arabidopsis elch mutant reveals functions of an ESCRT component in cytokinesis. *Development* **133,** 4679–4689 (2006).

224. Ren, X. & Hurley, J. H. VHS domains of ESCRT-0 cooperate in high-avidity binding to polyubiquitinated cargo. *Embo J* **29,** 1045–54 (2010).

225. Prag, G. *et al.* The Vps27/Hse1 complex is a GAT domain-based scaffold for ubiquitin-dependent sorting. *Dev. Cell* **12,** 973–986 (2007).

226. Leung, K. F., Dacks, J. B. & Field, M. C. Evolution of the multivesicular body ESCRT Machinery; retention across the eukaryotic lineage. *Traffic* **9,** 11716–16198 (2008).

227. Katoh, Y. *et al.* Tollip and Tom1 form a complex and recruit ubiquitin-conjugated proteins onto early endosomes. *J. Biol. Chem.* **279,** 24435–24443 (2004).

228. Puertollano, R. Interactions of TOM1L1 with the multivesicular body sorting machinery. *J. Biol. Chem.* **280,** 9258–9264 (2005).

229. Blanc, C. *et al.* Dictyostelium Tom1 contributes to an ancestral ESCRT-0 complex. *Traffic* **10,** 161–171 (2009).

230. Herman, E. K., Walker, G., van der Giezen, M. & Dacks, J. B. Multivesicular bodies in the enigmatic amoeboflagellate Breviata anathema and the evolution of ESCRT 0. *J. Cell Sci.* **124,** 613–621 (2011).

231. Kostelansky, M. S. *et al.* Molecular Architecture and Functional Model of the Complete Yeast ESCRT-I Heterotetramer. *Cell* **129,** 485–498 (2007).

232. Langelier, C. *et al.* Human ESCRT-II Complex and Its Role in Human Immunodeficiency Virus Type 1 Release. *J. Virol.* **80,** 9465–9480 (2006).

233. Babst, M., Katzmann, D. J., Snyder, W. B., Wendland, B. & Emr, S. D. Endosome-associated complex, ESCRT-II, recruits transport machinery for protein sorting at the multivesicular body. *Dev. Cell* **3,** 283–289 (2002).

234. Slagsvold, T. *et al.* Eap45 in mammalian ESCRT-II binds ubiquitin via a phosphoinositide- interacting GLUE domain. *J. Biol. Chem.* **280,** 19600–19606 (2005).

235. Teo, H. *et al.* ESCRT-I core and ESCRT-II GLUE domain structures reveal role for GLUE in linking to ESCRT-I and membranes. *Cell* **125,** 99–111 (2006).

236. Wollert, T., Wunder, C., Lippincott-Schwartz, J. & Hurley, J. H. Membrane scission by the ESCRT-III complex. *Nature* **458,** 172–177 (2009).

237. Yang, D. & Hurley, J. H. Structural role of the Vps4-Vta1 interface in ESCRT-III recycling. *Structure* **18,** 976–984 (2010).

238. Luhtala, N. & Odorizzi, G. Bro1 coordinates deubiquitination in the multivesicular body pathway by recruiting Doa4 to endosomes. *J. Cell Biol.* **166,** 717–729 (2004).

239. Bröcker, C., Engelbrecht-Vandré, S. & Ungermann, C. Multisubunit tethering complexes and their role in membrane fusion. *Curr. Biol.* **20,** 943–952 (2010).

240. Nickerson, D. P., Brett, C. L. & Merz, A. J. Vps-C complexes: gatekeepers of endolysosomal traffic. *Curr. Opin. Cell Biol.* **21,** 543–551 (2009).

241. Meiringer, C. T. A. *et al.* The Dsl1 protein tethering complex is a resident endoplasmic reticulum complex, which interacts with five soluble NSF (N-ethylmaleimide-sensitive factor) attachment protein receptors (SNAREs): Implications for fusion and fusion regulation. *J. Biol. Chem.* **286,** 25039–25046 (2011).

242. Dubuke, M. L. & Munson, M. The Secret Life of Tethers: The Role of Tethering Factors in SNARE Complex Regulation. *Front. Cell Dev. Biol.* **4,** 1–8 (2016).

243. Fasshauer, D., Sutton, R. B., Brunger, A. T. & Jahn, R. Conserved structural features of the synaptic fusion complex: SNARE proteins reclassified as Q- and R-SNAREs. *Proc. Natl. Acad. Sci.* **95,** 15781–15786 (1998).

244. Bock, J. B., Matern, H. T., Peden, a a & Scheller, R. H. A genomic perspective on membrane compartment organization. *Nature* **409,** 839–841 (2001).

245. Duman, J. G. & Forte, J. G. What is the role of SNARE proteins in membrane fusion? *Topology* (2003).

246. Kloepper, T. H., Kienle, C. N. & Fasshauer, D. SNAREing the basis of multicellularity: Consequences of protein family expansion during evolution. *Mol. Biol. Evol.* **25,** 2055–2068 (2008).

247. Venkatesh, D. *et al.* Evolution of the endomembrane systems of trypanosomatids – conservation and specialisation. *J. Cell Sci.* **130,** 1421–1434 (2017).

248. Novick, P. & Schekman, R. Secretion and cell-surface growth are blocked in a temperature-sensitive mutant of Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci.* **76,** 1858–1862 (1979).

249. Hata, Y., Slaughter, C. A. & Südhof, T. C. Synaptic vesicle fusion complex contains unc-18 homologue bound to syntaxin. *Nature* **366,** 347–351 (1993).

250. Li, Y., Gallwitz, D. & Peng, R. Structure-based Functional Analysis Reveals a Role for the SM Protein Sly1p in Retrograde Transport to the Endoplasmic Reticulum Yujie. *Mol. Biol. Cell* **16,** 3951–3962 (2005).

251. Dascher, C. & Balch, W. E. Mammalian Sly1 regulates syntaxin 5 function in endoplasmic reticulum to Golgi transport. *J. Biol. Chem.* **271,** 15866–15869 (1996).

252. Subramanian, S., Woolford, C. A. & Jones, E. W. The Sec1/Munc18 Protein, Vps33p, Functions at the Endosome and the Vacuole of Saccharomyces cerevisiae. *Mol. Biol. Cell* **15,** 2593–2605 (2004).

253. Bryant, N. J., Piper, R. C., Gerrard, S. R. & Stevens, T. H. Traffic into the prevacuolar/endosomal compartment of Saccharomyces cerevisiae: a VPS45-dependent intracellular route and a VPS45-independent, endocytic route. *Eur. J. Cell Biol.* **76,** 43–52 (1998).

254. Koumandou, V. L. *et al.* Control systems for membrane fusion in the ancestral eukaryote; evolution of tethering complexes and SM proteins. *BMC Evol. Biol.* **7,** 29 (2007).

255. Jahn, R. & Scheller, R. H. SNAREs — engines for membrane fusion. *Nat. Rev. Mol. Cell Biol.* **7,** 631–643 (2006).

256. Littleton, J. T. *et al.* Temperature-sensitive paralytic mutations demonstrate that synaptic exocytosis requires SNARE complex assembly and disassembly. *Neuron* **21,** 401–413 (1998).

257. Oguchi, M. E., Noguchi, K. & Fukuda, M. TBC1D12 is a novel Rab11-binding protein that modulates neurite outgrowth of PC12 cells. *PLoS One* **12,** 1–17 (2017).

258. Seaman, M. N. J., Harbour, M. E., Tattersall, D., Read, E. & Bright, N. Membrane recruitment of the cargo-selective retromer subcomplex is catalysed by the small GTPase Rab7 and inhibited by the Rab-GAP TBC1D5. *J. Cell Sci.* **122,** 2371–2382 (2009).

259. Zhang, X. M., Walsh, B., Mitchell, C. A. & Rowe, T. TBC domain family, member 15 is a novel mammalian Rab GTPase-activating protein with substrate preference for Rab7.

Biochem. Biophys. Res. Commun. **335,** 154–161 (2005).

260. Davey, J. R. *et al.* TBC1D13 is a RAB35 Specific GAP that Plays an Important Role in GLUT4 Trafficking in Adipocytes. *Traffic* **13,** 1429–1441 (2012).

261. Kouranti, I., Sachse, M., Arouche, N., Goud, B. & Echard, A. Rab35 Regulates an Endocytic Recycling Pathway Essential for the Terminal Steps of Cytokinesis. *Curr. Biol.* **16,** 1719–1725 (2006).

262. Popovic, D. *et al.* Rab GTPase-activating proteins in autophagy: regulation of endocytic and autophagy pathways by direct binding to human ATG8 modifiers. *Mol. Cell. Biol.* **32,** 1733–44 (2012).

263. Elias, M., Brighouse, A., Gabernet-Castello, C., Field, M. C. & Dacks, J. B. Sculpting the endomembrane system in deep time: high resolution phylogenetics of Rab GTPases. *J. Cell Sci.* **125,** 2500–8 (2012).

264. Pereira-Leal, J. B. & Seabra, M. C. Evolution of the rab family of small GTP-binding proteins. *J. Mol. Biol.* **313,** 889–901 (2001).

265. Albert, S. & Gallwitz, D. Two New Members of a Family of Ypt/Rab GTPase Activating Proteins. *J. Biol. Chem.* **247,** 33186–33189 (1999).

266. Gabernet-Castello, C., Reilly, A. O., Dacks, J. B. & Field, M. C. Evolution of Tre-2/Bub2/Cdc16 (TBC) Rab GTPase-activating proteins. *Mol. Biol. Cell* **24,** 1574–1583 (2013).

267. Levivier, E. *et al.* uDENN, DENN, and dDENN: indissociable domains in Rab and MAP kinases signaling pathways. *Biochem. Biophys. Res. Commun.* **287,** 688–695 (2001).

268. Yoshimura, S. I., Gerondopoulos, A., Linford, A., Rigden, D. J. & Barr, F. A. Family-wide characterization of the DENN domain Rab GDP-GTP exchange factors. *J. Cell Biol.* **191,** 367–381 (2010).

269. Carney, D. S., Davies, B. A. & Horazdovsky, B. F. Vps9 domain-containing proteins: Activators of Rab5 GTPases from yeast to neurons. *Trends Cell Biol.* **16,** 27–35 (2006).

270. De Franceschi, N. *et al.* Longin and GAF Domains: Structural Evolution and Adaptation to the Subcellular Trafficking Machinery. *Traffic* **15,** 104–121 (2014).

271. Vedovato, M., Rossi, V., Dacks, J. B. & Filippini, F. Comparative analysis of plant genomes allows the definition of the 'Phytolongins': a novel non-SNARE longin domain protein family. *BMC Genomics* **10,** 510 (2009).

272. Linford, A. *et al.* Rab14 and Its Exchange Factor FAM116 Link Endocytic Recycling and Adherens Junction Stability in Migrating Cells. *Dev. Cell* **22,** 952–966 (2012).

273. Lumb, J. H., Leung, K. F., DuBois, K. N. & Field, M. C. Rab28 function in trypanosomes: interactions with retromer and ESCRT pathways. *J. Cell Sci.* **124,** 3771–3783 (2011).

274. Nookala, R. K. *et al.* Crystal structure of folliculin reveals a hidDENN function in genetically inherited renal cancer. *Open Biol.* **2,** 120071–120071 (2012).

275. Marat, A. L. & McPherson, P. S. The connecdenn family, Rab35 guanine nucleotide exchange factors interfacing with the clathrin machinery. *J. Biol. Chem.* **285,** 10627–10637 (2010).

276. Xu, D., Joglekar, A. P., Williams, A. L. & Hay, J. C. Subunit structure of a mammalian ER/Golgi SNARE complex. *J. Biol. Chem.* **275,** 39631–9 (2000).

277. Liu, Y. & Barlowe, C. Analysis of Sec22p in Endoplasmic Reticulum/Golgi Transport Reveals Cellular Redundancy in SNARE Protein Function. *Mol. Biol. Cell* **13,** 3314–3324 (2002).

278. Cai, H., Reinisch, K. & Ferro-Novick, S. Coats, Tethers, Rabs, and SNAREs Work Together to Mediate the Intracellular Destination of a Transport Vesicle. *Dev. Cell* **12,** 671–682 (2007).

279. Haas, A. K. *et al.* Analysis of GTPase-activating proteins: Rab1 and Rab43 are key Rabs required to maintain a functional Golgi complex in human cells. *J. Cell Sci.* **120,** 2997–3010 (2007).

280. Lipka, V., Kwon, C. & Panstruga, R. SNARE-Ware : The Role of SNARE-Domain Proteins in Plant Biology. *Annu. Rev. Cell Dev. Biol.* **23,** 147–74 (2007).

281. Clarke, M. *et al.* Genome of *Acanthamoeba castellanii* highlights extensive lateral gene transfer and early evolution of tyrosine kinase signaling. *Genome Biol.* **14,** R11 (2013).

282. Lewis, M. J. & Pelham, H. R. B. SNARE-mediated retrograde traffic from the Golgi complex to the endoplasmic reticulum. *Cell* **85,** 205–215 (1996).

283. Hsia, K.-C. & Hoelz, A. Crystal structure of -COP in complex with -COP provides insight into the architecture of the COPI vesicular coat. *Proc. Natl. Acad. Sci.* **107,** 11271–11276 (2010).

284. White, J. *et al.* Rab6 coordinates a novel Golgi to ER retrograde transport pathway in live cells. *J Cell Biol* **147,** 743–760 (1999).

285. Brandizzi, F. & Barlowe, C. Organization of the ER–Golgi interface for membrane traffic control. *Nat. Rev. Mol. Cell Biol.* **14,** 382–392 (2013).

286. Morozova, N. *et al.* TRAPPII subunits are required for the specificity switch of a Ypt-Rab GEF. *Nat Cell Biol.* **8,** 1263–1269 (2006).

287. Malsam, J. & Söllner, T. H. Organization of SNAREs within the Golgi stack. *Cold Spring Harb. Perspect. Biol.* **3,** 1–17 (2011).

288. Choi, C. *et al.* Organization and Assembly of the TRAPPII Complex. *Traffic* **12,** 715–725 (2011).

289. Salaun, C., James, D. J., Greaves, J. & Chamberlain, L. H. Plasma membrane targeting of exocytic SNARE proteins. *Biochim. Biophys. Acta - Mol. Cell Res.* **1693,** 81–89 (2004).

290. Schilde, C., Lutter, K., Kissmehl, R. & Plattner, H. Molecular identification of a SNAP-25-like SNARE protein in Paramecium. *Eukaryot. Cell* **7,** 1387–1402 (2008).

291. Oishi, Y. *et al.* Role of VAMP-2, VAMP-7, and VAMP-8 in constitutive exocytosis from HSY cells. *Histochem. Cell Biol.* **125,** 273–281 (2006).

292. Advani, R. J. *et al.* VAMP-7 mediates vesicular transport from endosomes to lysosomes. *J. Cell Biol.* **146,** 765–775 (1999).

293. Zheng, H. *et al.* NPSN11 is a cell plate-associated SNARE protein that interacts with the syntaxin KNOLLE. *Plant Physiol* **129,** 530–539 (2002).

294. Wiederkehr, A., De Craene, J. O., Ferro-Novick, S. & Novick, P. Functional specialization within a vesicle tethering complex: Bypass of a subset of exocyst deletion mutants by Sec1p or Sec4p. *J. Cell Biol.* **167,** 875–887 (2004).

295. TerBush, D. R., Maurice, T., Roth, D. & Novick, P. The Exocyst is a multiprotein complex required for exocytosis in Saccharomyces cerevisiae. *EMBO J.* **15,** 6483–6494 (1996).

296. Dubuke, M. L., Maniatis, S., Shaffer, S. A. & Munson, M. The exocyst subunit Sec6 interacts with assembled exocytic SNARE complexes. *J. Biol. Chem.* **290,** 28245–28256 (2015).

297. Cvrčková, F. *et al.* Evolution of the Land Plant Exocyst Complexes. *Front. Plant Sci.* **3,** 1–13 (2012).

298. Hala, M. *et al.* An Exocyst Complex Functions in Plant Cell Growth in Arabidopsis and Tobacco. *Plant Cell Online* **20,** 1330–1345 (2008).

299. Boehm, C. M. *et al.* The Trypanosome Exocyst: A Conserved Structure Revealing a New Role in Endocytosis. *PLoS Pathog.* **13,** 1–25 (2017).

300. Mallard, F. *et al.* Early/recycling endosomes-to-TGN transport involves two SNARE complexes and a Rab6 isoform. *J. Cell Biol.* **156,** 653–664 (2002).

301. Jackson, A. J., Clucas, C., Mamczur, N. J., Ferguson, D. J. & Meissner, M. Toxoplasma gondii Syntaxin 6 Is Required for Vesicular Transport Between Endosomal-Like Compartments and the Golgi Complex. *Traffic* **14,** 1166–1181 (2013).

302. Liewen, H. *et al.* Characterization of the human GARP (Golgi associated retrograde protein) complex. *Exp. Cell Res.* **306,** 24–34 (2005).

303. Schindler, C., Chen, Y., Pu, J., Guo, X. & Bonifacino, J. S. EARP is a multisubunit tethering complex involved in endocytic recycling. *Nat. Cell Biol.* **17,** 639–650 (2015).

304. Stenmark, H. Rab GTPases as coordinators of vesicle traffic. *Nat. Rev. Mol. Cell Biol.* **10,** 513–525 (2009).

305. Pinar, M. *et al.* TRAPPII regulates exocytic Golgi exit by mediating nucleotide exchange on the Ypt31 ortholog RabERAB11. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 4346–51 (2015).

306. Rojas, R. *et al.* Regulation of retromer recruitment to endosomes by sequential action of Rab5 and Rab7. *J. Cell Biol.* **183,** 513–526 (2008).

307. Chotard, L. *et al.* TBC-2 Regulates RAB-5/RAB-7-mediated Endosomal Trafficking in

Caenorhabditis elegans. *Mol. Biol. Cell* **21,** 2285–2296 (2010).

308. Otomo, A. *et al.* ALS2, a novel guanine nucleotide exchange factor for the small GTPase Rab5, is implicated in endosomal dynamics. *Hum. Mol. Genet.* **12,** 1671–1687 (2003).

309. Wurmser, A. E., Sato, T. K. & Emr, S. D. New component of the vacuolar class C-Vps complex couples nucleotide exchange on the Ypt7 GTPase to SNARE-dependent docking and fusion. *J. Cell Biol.* **151,** 551–562 (2000).

310. Nordmann, M. *et al.* The Mon1-Ccz1 complex is the GEF of the late endosomal Rab7 homolog Ypt7. *Curr. Biol.* **20,** 1654–1659 (2010).

311. Balderhaar, H. J. kleine & Ungermann, C. CORVET and HOPS tethering complexes - coordinators of endosome and lysosome fusion. *J. Cell Sci.* **126,** 1307–16 (2013).

312. Van Der Kant, R. *et al.* Characterization of the mammalian CORVET and HOPS complexes and their modular restructuring for endosome specificity. *J. Biol. Chem.* **290,** 30280–30290 (2015).

313. Lürick, A. *et al.* The Habc domain of the SNARE Vam3 interacts with the HOPS tethering complex to facilitate vacuole fusion. *J. Biol. Chem.* **290,** 5405–5413 (2015).

314. Kramer, L. & Ungermann, C. HOPS drives vacuole fusion by binding the vacuolar SNARE complex and the Vam7 PX domain via two distinct sites. *Mol. Biol. Cell* **22,** 2601–2611 (2011).

315. Collins, K. M. & Wickner, W. T. trans-SNARE complex assembly and yeast vacuole membrane fusion. *Proc. Natl. Acad. Sci.* **104,** 8755–8760 (2007).

316. Woodman, P. G. P97, a Protein Coping With Multiple Identities. *J. Cell Sci.* **116,** 4283–4290 (2003).

317. Ramanathan, H. N. & Ye, Y. The p97 ATPase associates with EEA1 to regulate the size of early endosomes. *Cell Res.* **22,** 346–359 (2012).

318. Bechtel, W. *et al.* Vps34 Deficiency Reveals the Importance of Endocytosis for Podocyte Homeostasis. *J. Am. Soc. Nephrol.* **24,** 727–743 (2013).

319. Murray, J. T., Panaretou, C., Stenmark, H., Miaczynska, M. & Backer, J. M. Role of Rab5 in the recruitment of hVps34/p150 to the early endosome. *Traffic* **3,** 416–427 (2002).

320. Jaber, N. *et al.* Vps34 regulates Rab7 and late endocytic trafficking through recruitment of the GTPase-activating protein Armus. *J. Cell Sci.* **129,** 4424–4435 (2016).

321. Adl, S. M. *et al.* The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *J. Eukaryot. Microbiol.* **52,** 399–451 (2005).

322. Grau-Bové, X., Sebé-Pedrós, A. & Ruiz-Trillo, I. The eukaryotic ancestor had a complex ubiquitin signaling system of archaeal origin. *Mol. Biol. Evol.* **32,** 726–739 (2015).

323. Eme, L., Moreira, D., Talla, E. & Brochier-Armanet, C. A complex cell division machinery was present in the last common ancestor of eukaryotes. *PLoS One* **4,** (2009).

324. Wickstead, B. & Gull, K. The evolution of the cytoskeleton. *J. Cell Biol.* **194,** 513–525

(2011).

325. Obado, S. O. *et al.* Interactome Mapping Reveals the Evolutionary History of the Nuclear Pore Complex. *PLoS Biol.* **14,** 1–30 (2016).

326. Schlacht, A., Herman, E. K., Klute, M. J., Field, M. C. & Dacks, J. B. Missing pieces of an ancient puzzle: Evolution of the eukaryotic membrane-trafficking system. *Cold Spring Harb. Perspect. Biol.* **6,** 1–14 (2014).

327. Read, B. A. *et al.* Pan genome of the phytoplankton Emiliania underpins its global distribution. *Nature* **499,** 209–213 (2012).

328. Nowrousian, M. Next-generation sequencing techniques for eukaryotic microorganisms: Sequencing-based solutions to biological problems. *Eukaryot. Cell* **9,** 1300–1310 (2010).

329. Grigoriev, I. V. *et al.* MycoCosm portal: Gearing up for 1000 fungal genomes. *Nucleic Acids Res.* **42,** 699–704 (2014).

330. Gabernet-Castello, C., Reilly, A. O., Dacks, J. B. & Field, M. C. Evolution of Tre-2/Bub2/Cdc16 (TBC) Rab GTPase-activating proteins. *Mol. Biol. Cell* **24,** 1574–1583 (2013).

331. Davies, B. A. *et al.* Vps9p CUE domain ubiquitin binding is required for efficient endocytic protein traffic. *J. Biol. Chem.* **278,** 19826–19833 (2003).

332. van Hooff, J. J., Tromer, E., van Wijk, L. M., Snel, B. & Kops, G. J. Evolutionary dynamics of the kinetochore network in eukaryotes as revealed by comparative genomics. *EMBO Rep.* e201744102 (2017). doi:10.15252/embr.201744102

333. Rosengarten, R. D. *et al.* Leaps and lulls in the developmental transcriptome of Dictyostelium discoideum. *BMC Genomics* **16,** 294 (2015).

334. Maier, T., Güell, M. & Serrano, L. Correlation of mRNA and protein in complex biological samples. *FEBS Lett.* **583,** 3966–3973 (2009).

335. Koussounadis, A., Langdon, S. P., Um, I. H., Harrison, D. J. & Smith, V. A. Relationship between differentially expressed mRNA and mRNA-protein correlations in a xenograft model system. *Sci. Rep.* **5,** 10775 (2015).

336. Wideman, J. G., Leung, K. F., Field, M. C. & Dacks, J. B. The cell biology of the endocytic system from an evolutionary perspective. *Cold Spring Harb. Perspect. Biol.* **6,** (2014).

337. Dacks, J. B. & Field, M. C. Eukaryotic cell evolution from a comparative genomic perspective: the endomembrane system. *Syst. Assoc. Spec. Vol.* **68,** 309–334 (2004).

338. Stenzel, D. J. & Boreham, P. F. L. Blastocystis hominis revisited. *Clin. Microbiol. Rev.* **9,** 563–584 (1996).

339. Vázquezdelara-Cisneros, L. G. & Arroyo-Begovich, A. Induction of Encystation of Entamoeba invadens by Removal of Glucose from the Culture Medium. *J. Parasitol.* **70,** 629–633 (1984).

340. Ali, I. K. M. *et al.* Proteomic analysis of the cyst stage of Entamoeba histolytica. *PLoS Negl. Trop. Dis.* **6,** (2012).

341. Ehrenkaufer, G. M., Haque, R., Hackney, J. A., Eichinger, D. J. & Singh, U. Identification of developmentally regulated genes in Entamoeba histolytica: Insights into mechanisms of stage conversion in a protozoan parasite. *Cell. Microbiol.* **9,** 1426–1444 (2007).

342. Chatterjee, A. *et al.* Evidence for a 'wattle and daub' model of the cyst wall of Entamoeba. *PLoS Pathog.* **5,** (2009).

343. Das, S. *et al.* The cyst wall of Entamoeba invadens contains chitosan (deacetylated chitin). *Mol. Biochem. Parasitol.* **148,** 86–92 (2006).

344. Frisardi, M. *et al.* The Most Abundant Glycoprotein of Amebic Cyst Walls (Jacob) Is a Lectin with Five Cys-Rich , Chitin-Binding Domains. **68,** 4217–4224 (2000).

345. Ghosh, S. K. *et al.* Chitinase secretion by encysting Entamoeba invadens and transfected Entamoeba histolytica trophozoites: Localization of secretory vesicles, endoplasmic reticulum, and Golgi apparatus. *Infect. Immun.* **67,** 3073–3081 (1999).

346. Billard, C. & Inouye, I. in *Coccolithophores: From Molecular Processes to Global Impact* (eds. Thierstein, H. R. & Young, J. R.) 1–29 (Springer Berlin / Heidelberg, 2004).

347. Rost, B. & Riebesell, U. in *Coccolithophores: From Molecular Processes to Global Impact* (eds. Thierstein, H. R. & Young, J. R.) 99–126 (Springer-Verlag, 2004).

348. Taylor, A. R., Russell, M. A., Harper, G. M., Collins, T. F. T. & Brownlee, C. Dynamics of formation and secretion of heterococcoliths by Coccolithus pelagicus ssp. braarudii. *Eur. J. Phycol.* **42,** 125–136 (2007).

349. Reveiller, F. L., Suh, S. J., Sullivan, K., Cabanes, P. A. & Marciano-Cabral, F. Isolation of a unique membrane protein from Naegleria fowleri. *J. Eukaryot. Microbiol.* **48,** 676–682 (2001).

350. Toney, D. M. & Marciano-Cabral, F. Alterations in protein expression and complement resistance of pathogenic Naegleria amoebae. *Infect. Immun.* **60,** 2784–2790 (1992).

351. Toney, D. M. & Marciano-Cabral, F. Membrane vesiculation of Naegleria fowleri amoebae as a mechanism for resisting complement damage. *J. Immunol.* **152,** 2952–9 (1994).

352. Fritzinger, A. E., Toney, D. M., MacLean, R. C. & Marciano-Cabral, F. Identification of a Naegleria fowleri membrane protein reactive with anti-human CD59 antibody. *Infect. Immun.* **74,** 1189–1195 (2006).

353. Shin H.J., Cho M.S., Jung S.Y., et al. Molecular cloning and characterization of a gene encoding a 13.1kDa antigenic protein of Naegleria fowleri. *J. Eukaryot. Microbiol.* **48,** 713–714 (2001).

354. Aldape, K., Huizinga, H., Bouvier, J. & McKerrow, J. Naegleria fowleri: Characterization of a secreted histolytic cysteine protease. *Experimental Pathology* **78,** 230–241 (1994).

355. Sohn, H. J., Kim, J. H., Shin, M. H., Song, K. J. & Shin, H. J. The Nf-actin gene is an

important factor for food-cup formation and cytotoxicity of pathogenic Naegleria fowleri. *Parasitol Res* **106,** 917–924

356. Hu, W. N., Band, R. N. & Kopachik, W. J. Virulence-related protein synthesis in Naegleria fowleri. *Infect. Immun.* **59,** 4278–4282 (1991).

357. Jeong, S. R. *et al.* Cloning and characterization of an immunoreactive gene encoding a calcium-binding protein from Naegleria fowleri. *Mol. Biochem. Parasitol.* **137,** 169–173 (2004).

358. Herbst, R. *et al.* Pore-forming polypeptides of the pathogenic protozoon Naegleria fowleri. *J. Biol. Chem.* **277,** 22353–22360 (2002).

359. Marciano-Cabral, F. & Cabral, G. A. The immune response to Naegleria fowleri amebae and pathogenesis of infection. *FEMS Immunol. Med. Microbiol.* **51,** 243–259 (2007).

360. Marciano-Cabral, F. M. & Fulford, D. E. Cytopathology of pathogenic and nonpathogenic Naegleria species for cultured rat neuroblastoma cells. *Appl. Environ. Microbiol.* **51,** 1133–1137 (1986).

361. Fritz-Laylin, L. K., Ginger, M. L., Walsh, C., Dawson, S. C. & Fulton, C. The Naegleria genome: a free-living microbial eukaryote lends unique insights into core eukaryotic cell biology. *Res. Microbiol.* **162,** 607–618 (2011).

362. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).

363. Camacho, C. *et al.* BLAST plus: architecture and applications. *BMC Bioinformatics* **10,** 1 (2009).

364. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.* **89,** 10915–10919 (1992).

365. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7,** e1002195 (2011).

366. Eddy, S. R. A probabilistic model of local sequence alignment that simplifies statistical significance estimation. *PLoS Comput Biol* **4,** e1000069 (2008).

367. Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14,** 755–763 (1998).

368. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).

369. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28,** 3150–3152 (2012).

370. Marchler-Bauer, A. *et al.* CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.* **39,** D225–D229 (2010).

371. Finn, R. D. *et al.* The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Res.* **44,** D279–D285 (2016).

372. Mar, J. C., Harlow, T. J. & Ragan, M. A. Bayesian and maximum likelihood phylogenetic analyses of protein sequence data under relative branch-length differences and model

violation. *BMC Evol. Biol.* **5,** 8 (2005).

373. Moretti, S. *et al.* The M-Coffee web server: a meta-method for computing multiple sequence alignments by combining alternative alignment methods. *Nucleic Acids Res.* **35,** W645–W648 (2007).

374. Notredame, C., Higgins, D. G. & Heringa, J. T-coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302,** 205–217 (2000).

375. Maddison, W. P. & Maddison, D. R. Interactive analysis of phylogeny and character evolution using the computer program MacClade. *Folia Primatol.* **53,** 190–202 (1989).

376. Maddison, W. P. & Maddison, D. R. Mesquite: a modular system for evolutionary analysis. (2015).

377. Abascal, F., Zardoya, R. & Posada, D. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21,** 2104–2105 (2005).

378. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. ProtTest3: Fast selection of best-fit models of protein evolution. *Bioinformatics* **27,** 1164–1185 (2011).

379. Jones, D. T., Taylor, W. R. & Thornton, J. M. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **8,** 275–82 (1992).

380. Kosiol, C. & Goldman, N. Different versions of the dayhoff rate matrix. *Mol. Biol. Evol.* **22,** 193–199 (2005).

381. Dayhoff, M. & Schwartz, R. A Model of Evolutionary Change in Proteins. *Atlas protein Seq. Struct.* 345–352 (1978). doi:10.1.1.145.4315

382. Whelan, S. & Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **18,** 691–699 (2001).

383. Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol Biol Evol* **25,** 1307–1320 (2008).

384. Müller, T. & Vingron, M. Modeling Amino Acid Replacement. *J. Comput. Biol.* **7,** 761–776 (2000).

385. Bozdogan, H. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika* **52,** 345–370 (1987).

386. Burnham, K. P. & Anderson, D. R. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociol. Methods Res.* **33,** 261–304 (2004).

387. Huelsenbeck, J. P., Ronquist, F., Nielsen, R. & Bollback, J. P. Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology. *Science (80-.).* **294,** 2310–2314 (2001).

388. Mau, B., A., N. M. & Larget, B. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics* **6,** 1–12 (1999).

389. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under

mixed models. *Bioinformatics* **19,** 1572–1574 (2003).

390. Ronquist, F. *et al.* MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Syst. Biol.* **61,** 539–542 (2012).

391. Lartillot, N. & Philippe, H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* **21,** 1095–1109 (2004).

392. Lartillot, N. & Philippe, H. Computing Bayes Factors Using Thermodynamic Integration. *Syst. Biol.* **55,** 195–207 (2006).

393. Lartillot, N., Brinkmann, H. & Philippe, H. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol. Biol.* **7 Suppl 1,** S4 (2007).

394. Sullivan, J. Maximum-likelihood methods for phylogeny estimation. *Methods Enzymol.* **395,** 757–779 (2005).

395. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22,** 2688–2690 (2006).

396. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30,** 1312–1313 (2014).

397. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59,** 307–321 (2010).

398. Guindon, S. & Gascuel, O. A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Syst. Biol.* **52,** 696–704 (2003).

399. Zaharia, M. *et al.* Faster and More Accurate Sequence Alignment with SNAP. 1–10 (2011).

400. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).

401. Andrews, S. FastQC: a quality control tool for high throughput sequence data. (2010).

402. Fuller, C. W. *et al.* The challenges of sequencing by synthesis. *Nat. Biotechnol.* **27,** 1013–1023 (2009).

403. Ewing, B. *et al.* Base-Calling of Automated Sequencer Traces Using. *Genome Res.* 175–185 (2005). doi:10.1101/gr.8.3.175

404. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14,** R36 (2013).

405. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–9 (2012).

406. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28,** 511–5 (2010).

407. Roberts, A., Pimentel, H., Trapnell, C. & Pachter, L. Identification of novel transcripts in annotated genomes using RNA-seq. *Bioinformatics* **27,** 2325–2329 (2011).

408. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotech* **29,** 644–652 (2011).

409. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8,** 1494–1512 (2013).

410. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* **31,** 46–53 (2013).

411. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40,** 4288–4297 (2012).

412. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).

413. Zhang, Z. H. *et al.* A comparative study of techniques for differential expression analysis on RNA-seq data. *PLoS One* **9,** (2014).

414. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19,** ii215-ii225 (2003).

415. Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* **10,** 1177–1184 (2013).

416. Stanke, M., Schöffmann, O., Morgenstern, B. & Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7,** (2006).

417. Herman, E. K. *et al.* The Mitochondrial Genome and a 60-kb Nuclear DNA Segment from Naegleria fowleri, the Causative Agent of Primary Amoebic Meningoencephalitis. *J. Eukaryot. Microbiol.* **60,** 179–91 (2013).

418. Schattner, P., Brooks, A. N. & Lowe, T. M. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.* **33,** W686–W689 (2005).

419. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29,** 2933–2935 (2013).

420. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35,** 3100–3108 (2007).

421. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23,** 1061–1067 (2007).

422. Darling, A. C. E., Mau, B., Blattner, F. R. & Perna, N. T. Mauve : Multiple Alignment of Conserved Genomic Sequence With Rearrangements Mauve : Multiple Alignment of Conserved Genomic Sequence With Rearrangements. 1394–1403 (2004).

doi:10.1101/gr.2289704

423. Darling, A. E., Mau, B. & Perna, N. T. Progressivemauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* **5,** (2010).

424. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Res.* **13,** 2178–2189 (2003).

425. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30,** 1575–1584 (2002).

426. Sonnhammer, E. L., von Heijne, G. & Krogh, A. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Sixth Int. Conf. Intell. Syst. Mol. Biol.* **6,** 175–182 (1998).

427. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. . Predicting transmembrane protein topology with a hidden markov model: application to complete genomes11Edited by F. Cohen. *J. Mol. Biol.* **305,** 567–580 (2001).

428. Fulton, C. Axenic cultivation of Naegleria gruberi: Requirement for methionine. *Exp. Cell Res.* **88,** 365–370 (1974).

429. Silberman, J. D., Sogin, M. L., Leipe, D. D. & Clark, C. G. Human parasite finds taxonomic home. *Nature* **380,** 398–398 (1996).

430. Scanlan, P. D. & Stensvold, C. R. Blastocystis: Getting to grips with our guileful guest. *Trends Parasitol.* **29,** 523–529 (2013).

431. Roberts, T., Stark, D., Harkness, J. & Ellis, J. Update on the pathogenic potential and treatment options for Blastocystis sp. *Gut Pathog.* **6,** 17 (2014).

432. Parkar, U. *et al.* Molecular characterization of Blastocystis isolates from zoo animals and their animal-keepers. *Vet. Parasitol.* **169,** 8–17 (2010).

433. Alfellani, M. A. *et al.* Variable geographic distribution of Blastocystis subtypes and its potential implications. *Acta Trop.* **126,** 11–18 (2013).

434. Roberts, T., Stark, D., Harkness, J. & Ellis, J. Subtype distribution of Blastocystis isolates identified in a Sydney population and pathogenic potential of Blastocystis. *Eur. J. Clin. Microbiol. Infect. Dis.* **32,** 335–343 (2013).

435. Kaneda, Y. *et al.* Ribodemes of Blastocystis Hominis isolated in Japan. *Am. J. Trop. Med. Hyg.* **65,** 393–396 (2001).

436. Özyurt, M. *et al.* Molecular epidemiology of Blastocystis infections in Turkey. *Parasitol. Int.* **57,** 300–306 (2008).

437. Böhm-Gloning, B., Knobloch, J. & Walderich, B. Five subgroups of Blastocystis hominis from symptomatic and asymptomatic patients revealed by restriction site analysis of PCR-amplified 16S-like rDNA. *Trop. Med. Int. Health* **2,** 771–778 (1997).

438. Vogelberg, C. *et al.* Blastocystis sp. subtype 2 detection during recurrence of gastrointestinal and urticarial symptoms. *Parasitol. Int.* **59,** 469–471 (2010).

439. Dogruman-Al, F., Dagci, H., Yoshikawa, H., Kurt, Ö. & Demirel, M. A possible link between subtype 2 and asymptomatic infections of Blastocystis hominis. *Parasitol. Res.* **103,** 685–689 (2008).

440. Yoshikawa, H. *et al.* Polymerase chain reaction-based genotype classification among human Blastocystis hominis populations isolated from different countries. *Parasitol. Res.* **92,** 22–29 (2004).

441. Yan, Y. *et al.* Genetic variability of Blastocystis hominis isolates in China. *Parasitol. Res.* **99,** 597–601 (2006).

442. Domínguez-Márquez, M. V., Guna, R., Muñoz, C., Gómez-Muñoz, M. T. & Borrás, R. High prevalence of subtype 4 among isolates of Blastocystis hominis from symptomatic patients of a health district of Valencia (Spain). *Parasitol. Res.* **105,** 949–955 (2009).

443. Stensvold, C. R., Christiansen, D. B., Olsen, K. E. P. & Nielsen, H. V. Blastocystis sp. subtype 4 is common in Danish Blastocystis-positive patients presenting with acute diarrhea. *Am. J. Trop. Med. Hyg.* **84,** 883–885 (2011).

444. Fouad, S. A., Basyoni, M. M. A., Fahmy, R. A. & Kobaisi, M. H. The pathogenic role of different Blastocystis hominis genotypes isolated from patients with irritable bowel syndrome. *Arab J. Gastroenterol.* **12,** 194–200 (2011).

445. Pérez-Brocal, V., Shahar-Golan, R. & Clark, C. G. A linear molecule with two large inverted repeats: The mitochondrial genome of the stramenopile Proteromonas lacertae. *Genome Biol. Evol.* **2,** 257–266 (2010).

446. Maia, J., Gómez-Díaz, E. & Harris, D. Apicomplexa primers amplify Proteromonas (Stramenopiles, Slopalinida, Proteromonadidae) in tissue and blood samples from lizards. *Acta Parasitol.* **57,** 337–341 (2012).

447. Noël, C. & Dufernez, F. Molecular phylogenies of Blastocystis isolates from different hosts: implications for genetic diversity, identification of species, and zoonosis. *J. Clin. Microbiol.* **43,** 348 (2005).

448. Frank, W. *Diseases of Amphibians and Reptiles*. (Plenum Press, 1984).

449. Boenigk, J. & Arndt, H. Particle handling during interception feeding by four species of heterotrophic nanoflagellates. *J. Eukaryot. Microbiol.* **47,** 350–358 (2000).

450. Brugerolle, G. & Bardele, C. F. Cortical cytoskeleton of the flagellate Proteromonas lacertae: Interrelation between microtubules, membrane and somatonemes. *Protoplasma* **142,** 46–54 (1988).

451. Sym, S. D. & Maneveldt, G. W. in *eLS* 1–8 (John Wiley & Sons, Inc., 2011). doi:10.1002/9780470015902.a0001960.pub2

452. Dunn, L. A., Boreham, P. F. L. & Stenzel, D. J. Ultrastructural variation of Blastocystis hominis stocks in culture. *Int. J. Parasitol.* **19,** 43–56 (1989).

453. Chandrashekar, V. Detection and Enumeration of the Commonest Stool Parasites Seen in a Tertiary Care Center in South India. **2013,** (2013).

454. Yoshikawa, H. *et al.* Fecal-oral transmission of the cyst form of Blastocystis hominis in rats. *Parasitol. Res.* **94,** 391–396 (2004).

455. Makiuchi, T. & Nozaki, T. Highly divergent mitochondrion-related organelles in anaerobic parasitic protozoa. *Biochimie* **100,** 3–17 (2014).

456. Nasirudeen, A. M. A. & Tan, K. S. W. Isolation and characterization of the mitochondrion-like organelle from Blastocystis hominis. *J. Microbiol. Methods* **58,** 101–109 (2004).

457. Gray, M. W. Mitochondrial evolution. *Cold Spring Harb. Perspect. Biol.* **4,** (2012).

458. Nakamura, Y. *et al.* Phylogenetic position of Blastocystis hominis that contains cytochrome-free mitochondria, inferred from the protein phylogeny of elongation factor 1α. *Mol. Biochem. Parasitol.* **77,** 241–245 (1996).

459. Jackson, A. P. *et al.* Kinetoplastid Phylogenomics Reveals the Evolutionary Innovations Associated with the Origins of Parasitism. *Curr. Biol.* **26,** 161–172 (2016).

460. Saito-Nakano, Y., Loftus, B. J., Hall, N. & Nozaki, T. The diversity of Rab GTPases in Entamoeba histolytica. *Exp. Parasitol.* **110,** 244–252 (2005).

461. Mirza, H. & Tan, K. S. W. Blastocystis exhibits inter- and intra-subtype variation in cysteine protease activity. *Parasitol. Res.* **104,** 355–361 (2009).

462. Yoshikawa, H. & Hayakawa, A. Morphological changes in the central vacuole of Blastocystis hominis during in vitro culture. **131,** 63–68 (1996).

463. Burki, F. *et al.* Untangling the early diversification of eukaryotes: a phylogenomic study of the evolutionary origins of Centrohelida, Haptophyta and Cryptista. *Proc. R. Soc. B Biol. Sci.* **283,** 20152802 (2016).

464. Wawrzyniak, I. *et al.* Draft genome sequence of the intestinal parasite Blastocystis subtype 4-isolate WR1. *Genomics Data* **4,** 22–23 (2015).

465. Denoeud, F. *et al.* Genome sequence of the stramenopile Blastocystis, a human anaerobic parasite. *Genome Biol.* **12,** R29 (2011).

466. Gentekaki, E. *et al.* Extreme genome diversity in the hyper- prevalent parasitic eukaryote Blastocystis. *PLoS Biol.* **15,** e2003769 (2017).

467. Miller, E., Antonny, B., Hamamoto, S. & Schekman, R. Cargo selection into COPII vesicles is driven by the Sec24p subunit. *EMBO J.* **21,** 6105–6113 (2002).

468. Schlacht, A. & Dacks, J. B. Unexpected ancient paralogs and an evolutionary model for the COPII coat complex. *Genome Biol. Evol.* **7,** 1098–1109 (2015).

469. Manna, P. T., Gadelha, C., Puttick, A. E. & Field, M. C. ENTH and ANTH domain proteins participate in AP2-independent clathrin-mediated endocytosis. *J. Cell Sci.* **128,** 2130–2142 (2015).

470. Brodsky, F. M. *et al.* Clathrin light chains: arrays of protein motifs that regulate coated-vesicle dynamics. *Trends Biochem. Sci.* **16,** 208–213 (1991).

471. Manna, P. T., Gadelha, C., Puttick, A. E. & Field, M. C. ENTH and ANTH domain proteins participate in AP2-independent clathrin-mediated endocytosis. *J. Cell Sci.* **128,** 2130–2142 (2015).

472. Jahn, R. & Scheller, R. H. SNAREs — engines for membrane fusion. *Nat. Rev. Mol. Cell Biol.* **7,** 631–643 (2006).

473. Duszenko, M. *et al.* Autophagy in protists. *Autophagy* **7,** 127–158 (2011).

474. Abraham, R. T. & Wiederrecht, G. J. Immunopharmacology of rapamycin. *Annu. Rev. Immunol.* 483–510 (1996).

475. Stack, J. H., Herman, P. K., Schu, P. V & Emr2, S. D. A membrane-associated complex containing the Vpsl 5 protein kinase and the Vps34 Pi 3-kinase is essential for protein sorting to the yeast lysosome-like vacuole. *EMBO J.* **12,** 2195–2204 (1993).

476. Seaman, M. N. J., Marcusson, E. G., Cereghino, J. L. & Emr, S. D. Endosome to Golgi retrieval of the vacuolar protein sorting receptor, Vps10p, requires the function of the VPS29, VPS30, and VPS35 gene products. *J. Cell Biol.* **137,** 79–92 (1997).

477. Van Den Hazel, H. B., Kielland-Brandt, M. C. & Winther, J. R. Review: biosynthesis and function of yeast vacuolar proteases. *Yeast* **12,** 1–16 (1996).

478. Schmid, E. M. *et al.* Role of the AP2 β-appendage hub in recruiting partners for clathrin-coated vesicle assembly. *PLoS Biol.* **4,** 1532–1548 (2006).

479. Fyfe, I., Schuh, A. L., Edwardson, J. M. & Audhya, A. Association of the endosomal sorting complex ESCRT-II with the Vps20 subunit of ESCRT-III generates a curvature-sensitive complex capable of nucleating ESCRT-III filaments. *J. Biol. Chem.* **286,** 34262–34270 (2011).

480. Klinger, C. M., Klute, M. J. & Dacks, J. B. Comparative genomic analysis of multi-subunit tethering complexes demonstrates an ancient pan-eukaryotic complement and sculpting in apicomplexa. *PLoS One* **8,** e76278 (2013).

481. Opperdoes, F. R. Localization of the initial steps in alkoxyphospholipid biosynthesis in glycosomes (microbodies) of Trypanosoma brucei. *FEBS Lett.* **169,** 35–39 (1984).

482. Žárský, V. & Tachezy, J. Evolutionary loss of peroxisomes – not limited to parasites. *Biol. Direct* **10,** 74 (2015).

483. VanRheenen, S. M., Reilly, B. a, Chamberlain, S. J. & Waters, M. G. Dsl1p, an essential protein required for membrane traffic at the endoplasmic reticulum/Golgi interface in yeast. *Traffic* **2,** 212–31 (2001).

484. Cosson, P. & Letourneur, F. Coatomer interaction with di-lysine endoplasmic reticulum retention motifs. *Science* **263,** 1629–1631 (1994).

485. Suckling, R. J. *et al.* Structural basis for the binding of tryptophan-based motifs by δ -COP. 1–6 (2015). doi:10.1073/pnas.1506186112

486. Yin, J., Ye, A. J. J. & Tan, K. S. W. Autophagy is involved in starvation response and cell death in Blastocystis. *Microbiology* **156,** 665–677 (2010).

487. Delahaye, J. L. *et al.* Caenorhabditis elegans HOPS and CCZ-1 mediate trafficking to lysosome-related organelles independently of RAB-7 and SAND-1. **25,** 1073–1096 (2014).

488. Huang, G. *et al.* Adaptor protein-3 (AP-3) complex mediates the biogenesis of acidocalcisomes and is essential for growth and virulence of. **3,** (2011).

489. Morlon-Guyot, J., Pastore, S., Berry, L., Lebrun, M. & Daher, W. Toxoplasma gondii Vps11 , a subunit of HOPS and CORVET tethering complexes, is essential for the biogenesis of secretory organelles. *Cell. Microbiol.* **17,** 1157–1178 (2015).

490. WHO, PAHO & UNESCO. *Report: a consultation with experts on amoebiasis.* (1997).

491. Walsh, J. A. Problems in Recognition and Diagnosis of Amebiasis : Estimation of the Global Magnitude of Morbidity and Mortality. *Rev. Infect. Dis.* **8,** 228–238 (1986).

492. Ximénez, C., Morán, P., Rojas, L., Valadez, A. & Gómez, A. Reassessment of the epidemiology of amebiasis: State of the art. *Infect. Genet. Evol.* **9,** 1023–1032 (2009).

493. Ali, V. & Nozaki, T. Current therapeutics, their problems, and sulfur-containing-amino-acid metabolism as a novel target against infections by 'amitochondriate' protozoan parasites. *Clin. Microbiol. Rev.* **20,** 164–187 (2007).

494. Ishida, H. *et al.* Fulminant amoebic colitis with perforation successfully treated by staged surgery: A case report. *J. Gastroenterol.* **38,** 92–96 (2003).

495. Gupta, S. S., Singh, O., Shukla, S. & Raj, M. K. Acute fulminant necrotizing amoebic colitis: a rare and fatal complication of amoebiasis: a case report. *Cases J.* **2,** 6557 (2009).

496. Kenneth, J. R. & Ray, C. G. *Sherris Medical Microbiology*. *Vasa* (2004). doi:10.1036/0838585299

497. Nozaki, T. & Bhattacharya, A. *Amebiasis*. (2015). doi:10.1007/978-4-431-55200-0

498. Frisardi, M. *et al.* The Most Abundant Glycoprotein of Amebic Cyst Walls (Jacob) Is a Lectin with Five Cys-Rich , Chitin-Binding Domains. *Infect. Immun.* **68,** 4217–4224 (2000).

499. Vega, H. De *et al.* Cloning and expression of chitinases of Entamoeba. *Mol. Biochem. Parasitol.* **85,** 139–147 (1997).

500. De Cádiz, A. E., Jeelani, G., Nakada-Tsukui, K., Caler, E. & Nozaki, T. Transcriptome analysis of encystation in Entamoeba invadens. *PLoS One* **8,** e74840 (2013).

501. Herman, E. *et al.* Membrane Trafficking Modulation during Entamoeba Encystation. *Sci. Rep.* **7,** 12854 (2017).

502. Li, B. & Dewey, C. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12,** 323 (2011).

503. Nakada-Tsukui, K., Saito-Nakano, Y., Ali, V. & Nozaki, T. A Retromerlike Complex Is a Novel Rab7 Effector That Is Involved in the Transport of the Virulence Factor Cysteine Protease in the Enteric Protozoan Parasite Entamoeba histolytica. *Mol. Biol. Cell* **16,**

5294–5303 (2005).

504. Schlüter, A. *et al.* The evolutionary origin of peroxisomes: An ER-peroxisome connection. *Mol. Biol. Evol.* **23,** 838–845 (2006).

505. Ju, X. C. *et al.* The hominoid-specific gene TBC1D3 promotes generation of basal neural progenitors and induces cortical folding in mice. *Elife* **5,** 1–25 (2016).

506. Makioka, A., Kumagai, M., Hiranuka, K., Kobayashi, S. & Takeuchi, T. Different structure and mRNA expression of Entamoeba invadens chitinases in the encystation and excystation. *Parasitol. Res.* **109,** 417–423 (2011).

507. Nakano, A. & Muramatsu, M. A Novel GTP-binding Protein, Sarlp, Is Involved in Transport from the Endoplasmic Reticulum to the Golgi Apparatus. *J. Cell Biol.* **109,** 2677–2691 (1989).

508. Welter, B. H. & Temesvari, L. a. Overexpression of a mutant form of EhRabA, a unique rab GTPase of Entamoeba Histolytica, alters endoplasmic reticulum morphology and localization of the Gal/GalNAc adherence lectin. *Eukaryot. Cell* **8,** 1014–1026 (2009).

509. Zahraoui, A., Touchot, N., Chardin, P. & Tavitian, A. The human rab genes encode a family of GTP-binding proteins related to yeast YPT1 and SEC4 products involved in secretion. *J. Biol. Chem.* **264,** 12394–12401 (1989).

510. Gotthardt, D. *et al.* Proteomics Fingerprinting of Phagosome Maturation and Evidence for the Role of a G during Uptake. *Mol. Cell. Proteomics* **5,** 2228–2243 (2006).

511. Bosch, D. E. & Siderovski, D. P. G protein signaling in the parasite Entamoeba histolytica. *Exp. Mol. Med.* **45,** e15 (2013).

512. Romero-Diaz, M. *et al.* Structural and functional analysis of the Entamoeba histolytica EhrabB gene promoter. *BMC Mol Biol* **8,** 82 (2007).

513. Hendrick, H. M. *et al.* Phosphorylation of Eukaryotic Initiation Factor-2α during Stress and Encystation in Entamoeba Species. *PLoS Pathog.* **12,** 1–21 (2016).

514. Shock, J. L., Fischer, K. F. & DeRisi, J. L. Whole-genome analysis of mRNA decay in Plasmodium falciparum reveals a global lengthening of mRNA half-life during the intra-erythrocytic development cycle. *Genome Biol.* **8,** R134 (2007).

515. Luna-Nácar, M. *et al.* Proteomic study of Entamoeba histolytica trophozoites, cysts, and cyst-like structures. *PLoS One* **11,** 1–24 (2016).

516. Duran, J. M., Anjard, C., Stefan, C., Loomis, W. F. & Malhotra, V. Unconventional secretion of Acb1 is mediated by autophagosomes. *J. Cell Biol.* **188,** 527–536 (2010).

517. Lidell, M. E., Moncada, D. M., Chadee, K. & Hansson, G. C. Entamoeba histolytica cysteine proteases cleave the MUC2 mucin in its C-terminal domain and dissolve the protective colonic mucus gel. *Proc. Natl. Acad. Sci.* **103,** 9298–9303 (2006).

518. Mitra, B. N., Saito-Nakano, Y., Nakada-Tsukui, K., Sato, D. & Nozaki, T. Rab11B small GTPase regulates secretion of cysteine proteases in the enteric protozoan parasite Entamoeba histolytica. *Cell. Microbiol.* **9,** 2112–2125 (2007).

519. Urbé, S.; Huber, L. A.; Zerial, M.; Tooze, S. A. and Parton, R. G. Rab 11, a small GTPase associated with both constitutive and regulatory secretory pathwyas in PC12 cells. *FEBS lett.* **334,** 175–182 (1993).

520. Chen, W.; Feng, Y.; Chen, D. and Wandinger-Ness, A. Rab11 is required for trans-golgi network-to-plasma membrane transport and a preferential target for GDP dissociation inhibitor. *Mol. Biol. Cell* **9,** 3241–3257 (1998).

521. Zolov, S. N. & Lupashin, V. V. Cog3p depletion blocks vesicle-mediated Golgi retrograde trafficking in HeLa cells. *J. Cell Biol.* **168,** 747–759 (2005).

522. Conibear, E. & Stevens, T. H. Vps52p, Vps53p, and Vps54p form a novel multisubunit complex required for protein sorting at the yeast late Golgi. *Mol. Biol. Cell* **11,** 305–323 (2000).

523. Cai, H., Zhang, Y., Pypaert, M., Walker, L. & Ferro-Novick, S. Mutants in trs 120 disrupt traffic from the early endosome to the late Golgi. *J. Cell Biol.* **171,** 823–833 (2005).

524. Rojas, A. M., Fuentes, G., Rausell, A. & Valencia, A. The Ras protein superfamily: Evolutionary tree and role of conserved amino acids. *J. Cell Biol.* **196,** 189–201 (2012).

525. McGugan, G. C. & Temesvari, L. a. Characterization of a Rab11-like GTPase, EhRab11, of Entamoeba histolytica. *Mol. Biochem. Parasitol.* **129,** 137–146 (2003).

526. Temesvari, L. A., Harris, E. N., Stanley Jr., S. L. & Cardelli, J. A. Early and late endosomal compartments of Entamoeba histolytica are enriched in cysteine proteases, acid phosphatase and several Ras-related Rab GTPases. *Mol. Biochem. Parasitol.* **103,** 225–241 (1999).

527. Saito-Nakano, Y., Yasuda, T., Nakada-Tsukui, K., Leippe, M. & Nozaki, T. Rab5-associated vacuoles play a unique role in phagocytosis of the enteric protozoan parasite Entamoeba histolytica. *J. Biol. Chem.* (2004). doi:10.1074/jbc.M403987200

528. Okada, M. & Nozaki, T. New insights into molecular mechanisms of phagocytosis in Entamoeba histolytica by proteomic analysis. *Archives of Medical Research* (2006). doi:10.1016/j.arcmed.2005.10.003

529. Lasko, P. mRNA localization and translational control in Drosophila oogenesis. *Cold Spring Harb. Perspect. Biol.* **4,** 1–15 (2012).

530. Petri Jr., W. A. & Singh, U. Diagnosis and Management of Amebiasis. *Clin. Infect. Dis.* **29,** 1117–1125 (1990).

531. Byers, T. J., Kim, B. G., King, L. E. & Hugo, E. R. Molecular Aspects of the Cell Cycle and Encystment of Acanthamoeba. *Rev. Infect. Dis.* **13,** S378-84 (1991).

532. Benítez, L. & Gutiérrez, J. C. Encystment-specific mRNA is accumulated in the resting cysts of the ciliate Colpoda inflata. *Biochem. Mol. Biol. Int.* **41,** 1137–1141 (1997).

533. Ehrenkaufer, G. M. *et al.* The genome and transcriptome of the enteric parasite Entamoeba invadens, a model for encystation. *Genome Biol.* **14,** R77 (2013).

534. Poulton, A. J., Adey, T. R., Balch, W. M. & Holligan, P. M. Relating coccolithophore

calcification rates to phytoplankton community dynamics: Regional differences and implications for carbon export. *Deep. Res. Part II Top. Stud. Oceanogr.* **54,** 538–557 (2007).

535. Holligan, P. M. *et al.* A biogeochemical study of the coccolithophore, Emiliania huxleyi, in the North Atlantic. *Global Biogeochem. Cycles* **7,** 879–900 (1993).

536. Tyrrell, T., Holligan, P. M. & Mobley, C. D. Optical impacts of oceanic coccolithophore blooms. *J. Geophys. Res.* **104,** 3223 (1999).

537. Charlson, R. J., Lovelock, J. E., Andreae, M. O. & Warren, S. G. Oceanic phytoplankton, atmospheric sulfur, cloud albedo and climate. *Nature* **326,** 655–661 (1987).

538. Meier, K. J. S., Beaufort, L., Heussner, S. & Ziveri, P. The role of ocean acidification in Emiliania huxleyi coccolith thinning in the Mediterranean Sea. *Biogeosciences* **11,** 2857–2869 (2014).

539. Fichtinger-Schepman, A. M. J., Kamerling, J. P., Versluis, C. & Vliegenthart, J. F. G. Structural analysis of acidic oligosaccharides derived from the methylated, acidic polysaccharide associated with coccoliths of Emiliania huxleyi (lohmann) kamptner. *Carbohydr. Res.* **86,** 215–225 (1980).

540. Brownlee, C., Wheeler, G. L. & Taylor, A. R. Coccolithophore biomineralization: New questions, new answers. *Semin. Cell Dev. Biol.* **46,** 11–16 (2015).

541. Sviben, S. *et al.* A vacuole-like compartment concentrates a disordered calcium phase in a key coccolithophorid alga. *Nat. Commun.* **7,** 11228 (2016).

542. Young, J. R. & Ziveri, P. Calculation of coccolith volume and its use in calibration of carbonate flux estimates. *Deep. Res. Part II Top. Stud. Oceanogr.* **47,** 1679–1700 (2000).

543. Edvardsen, B. *et al.* Phylogenetic reconstructions of the Haptophyta inferred from 18S ribosomal DNA sequences and available morphological data. *Phycologia* **39,** 19–35 (2000).

544. Fujiwara, S., Tsuzuki, M., Kawachi, M., Minaka, N. & Inouye, I. Molecular Phylogeny of the Haptophyta based on the rbc L gene and sequence variation in the spacer region of the RUBISCO operon. *J. Phycol.* **129,** 121–129 (2001).

545. Bendif, E. M. & Young, J. On the Ultrastructure of Gephyrocapsa oceanica (Haptophyta) Life Stages. *Cryptogam. Algol.* **35,** 379–388 (2014).

546. Laguna, R., Romo, J., Read, B. a & Wahlund, M. Induction of Phase Variation Events in the Life Cycle of the Marine Coccolithophorid Emiliania huxleyi Induction of Phase Variation Events in the Life Cycle of the Marine Coccolithophorid Emiliania huxleyi. **67,** 3824–3831 (2001).

547. Paasche, E. A review of the coccolithophorid Emiliania huxleyi. *Phycologia* **40,** 503–529 (2002).

548. Liu, H., Aris-Brosou, S., Probert, I. & De Vargas, C. A time line of the environmental genetics of the haptophytes. *Mol. Biol. Evol.* **27,** 161–176 (2010).

549. Hovde, B. T. *et al.* Genome Sequence and Transcriptome Analyses of Chrysochromulina tobin: Metabolic Tools for Enhanced Algal Fitness in the Prominent Order Prymnesiales (Haptophyceae). *PLoS Genet.* **11,** 1–31 (2015).

550. Green, J. C. & Jennigs, D. H. A physical and chemical investigation of the scales produced by the golgi apparatus within and found on the surface of the cells of Chrysochromulina chiton parke et manton. *J. Exp. Bot.* **18,** 359–370 (1967).

551. Kawachi, M., Inouye, I., Maeda, O. & Chihara, M. The haptonema as a food-capturing device: observations on Chrysochromulina hirta (Prymnesiophyceae). *Phycologia* **30,** 563–573 (1991).

552. Lee, L. J. Y., Klute, M. J., Herman, E. K., Read, B. & Dacks, J. B. Losses, Expansions, and Novel Subunit Discovery of Adaptor Protein Complexes in Haptophyte Algae. *Protist* **166,** 585–597 (2015).

553. Bendif, E. M. *et al.* Genetic delineation between and within the widespread coccolithophore morpho-species Emiliania huxleyi and Gephyrocapsa oceanica (Haptophyta). *J. Phycol.* **50,** 140–148 (2014).

554. Aridor, M., Weissman, J., Bannykh, S., Nuoffer, C. & Balch, W. E. during Export from the ER. *J. Cell Biol.* **141,** 61–70 (1998).

555. Yoshihisa, T., Barlowe, C. & Schekman, R. Requirement for a GTPase-activating protein in vesicle budding from the endoplasmic reticulum. *Science (80-.).* **259,** 1466–1468 (1993).

556. Deloche, O., Yeung, B. G., Payne, G. S. & Schekman, R. Vps10p transport from the trans-Golgi network to the endosome is mediated by clathrin-coated vesicles. *Mol. Biol. Cell* **12,** 475–85 (2001).

557. Lottridge, J. M., Flannery, A. R., Vincelli, J. L. & Stevens, T. H. Vta1p and Vps46p regulate the membrane association and ATPase activity of Vps4p at the yeast multivesicular body. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 6202–7 (2006).

558. Agromayor, M. & Martin-Serrano, J. Interaction of AMSH with ESCRT-III and deubiquitination of endosomal cargo. *J. Biol. Chem.* **281,** 23083–23091 (2006).

559. Kalinowska, K. *et al. Arabidopsis* ALIX is required for the endosomal localization of the deubiquitinating enzyme AMSH3. *Proc. Natl. Acad. Sci.* **112,** E5543–E5551 (2015).

560. Shahriari, M. *et al.* The AAA-type ATPase AtSKD1 contributes to vacuolar maintenance of Arabidopsis thaliana. *Plant J.* **64,** 71–85 (2010).

561. Boman, A. L., Zhang, C. J., Zhu, X. & Kahn, R. A. A family of ADP-ribosylation factor effectors that can alter membrane transport through the trans-Golgi. *Mol. Biol. Cell* **11,** 1241–55 (2000).

562. Dell'Angelica, E. C. *et al.* GGAs: A family of ADP ribosylation factor-binding proteins related to adaptors and associated with the Golgi complex. *J. Cell Biol.* **149,** 81–93 (2000).

563. Hirst, J. *et al.* A family of proteins with γ-adaptin and VHS domains that facilitate trafficking between the trans-golgi network and the vacuole/lysosome. *J. Cell Biol.* **149,** 67–79 (2000).

564. Hirst, J. *et al.* Distinct and overlapping roles for AP-1 and GGAs revealed by the 'knocksideways' system. *Curr. Biol.* **22,** 1711–1716 (2012).

565. Dascher, C., Matteson, J. & Balch, W. E. Syntaxin 5 Regulates Endoplasmic Reticulum to Golgi Transport. *J. Biol. Chem.* **269,** 29363–29366 (1994).

566. Lupashin, V. V, Pokrovskaya, I. D., McNew, J. A. & Waters, M. G. Characterization of a novel yeast SNARE protein implicated in Golgi retrograde traffic. *Mol. Biol. Cell* **8,** 2659–76 (1997).

567. Amessou, M. *et al.* Syntaxin 16 and syntaxin 5 are required for efficient retrograde transport of several exogenous and endogenous cargo proteins. *J. Cell Sci.* **120,** 1457–1468 (2007).

568. Chaineau, M., Danglot, L. & Galli, T. Multiple roles of the vesicular-SNARE TI-VAMP in post-Golgi and endosomal trafficking. *FEBS Lett.* **583,** 3817–3826 (2009).

569. Hsu, S. C. *et al.* The mammalian brain rsec6/8 complex. *Neuron* **17,** 1209–1219 (1996).

570. Curwin, A. J., Fairn, G. D. & McMaster, C. R. Phospholipid transfer protein Sec14 is required for trafficking from endosomes and regulates distinct trans-golgi export pathways. *J. Biol. Chem.* **284,** 7364–7375 (2009).

571. Coda, L. *et al.* Eps15R is a tyrosine kinase substrate with characteristics of a docking protein possibly involved in coated pits-mediated internalization. *J. Biol. Chem.* **273,** 3003–3012 (1998).

572. Hayashi, K. & Shirao, T. Change in the shape of dendritic spines caused by overexpression of drebrin in cultured cortical neurons. *J. Neurosci.* **19,** 3918–3925 (1999).

573. Cremona, O. *et al.* Essential role of phosphoinositide metabolism in synaptic vesicle recycling. *Cell* **99,** 179–188 (1999).

574. Lombardi, D. *et al.* Rab9 functions in transport between late endosomes and the trans Golgi network. *EMBO J.* **12,** 677–82 (1993).

575. Carruthers, V. B. & Sibley, L. D. Sequential protein secretion from three distinct organelles of Toxoplasma gondii accompanies invasion of human fibroblasts. *Eur. J. Cell Biol.* **73,** 114–123 (1997).

576. Sawada, K. & Shiraiwa, Y. Alkenone and alkenoic acid compositions of the membrane fractions of Emiliania huxleyi. *Phytochemistry* **65,** 1299–1307 (2004).

577. Eltgroth, M. L., Watwood, R. L. & Wolfe, G. V. Production and cellular localization of neutral long-chain lipids in the haptophyte algae Isochrysis galbana and Emiliania huxleyi. *J. Phycol.* **41,** 1000–1009 (2005).

578. Sutton, R. B., Fasshauer, D., Jahn, R. & Brunger, A. T. Crystal structure of a SNARE complex involved in synaptic exocytosis at 2.4 A resolution. *Nature* **395,** 347–353 (1998).

579. Carter, R. F. Description of a Naegleria sp. isolated from two cases of primary amoebic meningo-encephalitis, and of the experimental pathological changes induced by it. *J. Pathol.* **100,** 217–244 (1970).

580. Capewell, L. G. *et al.* Diagnosis, clinical course, and treatment of primary amoebic meningoencephalitis in the United States, 1937-2013. *J. Pediatric Infect. Dis. Soc.* **4,** e68–e75 (2015).

581. Heggie, T. W. Swimming with death: Naegleria fowleri infections in recreational waters. *Travel Med. Infect. Dis.* **8,** 201–206 (2010).

582. Das, S. R. Pathogenicity of Flagellate Stage Naegleria aerobia and its Bearing on the Epidemiology of Exogenous Amoebiasis. *Ann. Soc. Belg. Med. Trop. (1920).* **54,** 327–332 (1974).

583. Barnett, N. D. P., Kaplan, A. M., Hopkin, R. J., Saubolle, M. A. & Rudinsky, M. F. Primary amoebic meningoencephalitis with Naegleria fowleri: Clinical review. *Pediatr. Neurol.* **15,** 230–234 (1996).

584. Ma, P. *et al.* Naegleria and Acanthamoeba Infections : Review. *Rev. Infect. Dis.* **12,** 490–513 (2016).

585. Yoder, J. S. *et al.* Primary amebic meningoencephalitis deaths associated with sinus irrigation using contaminated tap water. *Clin. Infect. Dis.* **55,** 79–85 (2012).

586. Siddiqui, R. & Khan, N. A. Primary Amoebic Meningoencephalitis Caused by Naegleria fowleri: An Old Enemy Presenting New Challenges. *PLoS Negl. Trop. Dis.* **8,** (2014).

587. Fowler, M. & Carter, R. Acute Pyogenic Meningitis Probably Due to Acanthamoeba sp.: a Preliminary Report. *Br. Med. J.* **2,** 740–742 (1965).

588. Červa1, L. & Novak, K. Amoebic Meningoencephalitis: Sixteen Fatalities. *Science (80-.).* **160,** 92 (1968).

589. Butt, C. G., Baro, C. & Knorr, R. W. Naegleria (sp.) Identified in Amebic Encephalitis. *Am. J. Clin. Pathol.* **50,** 568–574 (1968).

590. Callicott, J. H. *et al.* Meningoencephalitis Due to Pathogenic Free-Living Amoebae. *Jama* **206,** 579 (1968).

591. Patras, D. & Andujar, J. J. Meningoencephalitis due to Hartmannella (Acanthamoeba). *Am. J. Clin. Pathol.* **45,** 226–33 (1966).

592. De Jonckheere, J. F. Origin and evolution of the worldwide distributed pathogenic amoeboflagellate Naegleria fowleri. *Infect. Genet. Evol.* **11,** 1520–1528 (2011).

593. Griffin, J. L. Temperature Tolerance of Pathogenic and Nonpathogenic Free-Living Amoebas. *Science (80-.).* **178,** 869–870 (1972).

594. Chang, S. L. Resistance of pathogenic Naegleria to some common physical and chemical agents. *Appl. Environ. Microbiol.* **35,** 368–375 (1978).

595. Marciano-Cabral, F. Biology of Naegleria spp. *Microbiol. Rev.* **52,** 114–133 (1988).

596. Izumiyama, S. *et al.* Occurrence and Distribution of Naegleria Species in Thermal Waters in Japan. *J. Eukaryot. Microbiol.* **50,** 514–515 (2003).

597. Sheehan, K. B., Ferris, M. J. & Benson, J. M. Detection of Naegleria sp. in a Thermal, Acidic Stream in Yellowstone National Park. *J. Eukaryot. Microbiol.* **50,** 263–265 (2003).

598. John, D. T. in *Parasitic Protozoa* (eds. Kreier, J. P. & Baker, J. R.) 143–246 (Academic Press, 1993).

599. Kyle, D. E. & Noblet, G. P. Seasonal Distribution of Thermotolerant Free-Living Amoebae. II. Lake Issaqueena. *J. Protozool.* **34,** 10–15 (1987).

600. Dingle, A. D. & Fulton, C. Development of the Flagellar Apparatus of Naegleria. *J. Cell Biol.* **31,** 43–54 (1966).

601. Daft, B. M. *et al.* Seasonal meningoencephalitis in Holstein cattle caused by Naegleria fowleri. *J. Vet. Diagn. Invest.* **17,** 605–609 (2005).

602. John, D. T. & Hoppe, K. L. Susceptibility of Wild Mammals to Infection with Naegleria fowleri. *J. Parasitol.* **76,** 865–868 (1990).

603. Scaglia, M., Gatti, S., Cevini, C., Bernuzzi, A. M. & Martinez, A. J. Naegleria australiensis Experimental Study in Mice. *Exp. Parasitol.* **299,** 294–299 (1989).

604. Scaglia, M. *et al.* Isolation and identification of pathogenic Naegleria australiensis (Amoebida, Vahlkampfiidae) from a spa in northern Italy. *Appl. Environ. Microbiol.* **46,** 1282–1285 (1983).

605. De Jonckheere, J. F. Molecular definition and the ubiquity of species in the genus Naegleria. *Protist* **155,** 89–103 (2004).

606. De Jonckheere, J. F. What do we know by now about the genus Naegleria? *Exp. Parasitol.* **145,** S2–S9 (2014).

607. Whiteman, L. Y. & Marciano-Cabral, F. Susceptibility of pathogenic and nonpathogenic Naegleria spp. to complement-mediated lysis. *Infect. Immun.* **55,** 2442–2447 (1987).

608. Stevens, A. R., De Jonckheere, J. & Willaert, E. NAEGLERIA LOVANIENSIS NEW SPECIES : ISOLATION AND IDENTIFICATION OF SIX THERMOPHILIC STRAINS OF A NEW SPECIES FOUND IN ASSOCIATION WITH NAEGLERIA FOWLERI. *Int. J. Parasitol.* **10,** 51–64 (1980).

609. Serrano-Luna, J., Cervantes-Sandoval, I., Tsutsumi, V. & Shibayama, M. A Biochemical Comparison of Proteases from Pathogenic Naegleria fowleri and Non-Pathogenic Naegleria gruberi. *J. Eukaryot. Microbiol.* **54,** 411–417 (2007).

610. Fritz-Laylin, L. K. *et al.* The genome of Naegleria gruberi illuminates early eukaryotic versatility. *Cell* **140,** 631–642 (2010).

611. Baverstock, P. R., Illana, S., Christy, P. E., Robinson, B. S. & Johnson, A. M. srRNA evolution and phylogenetic relationships of the genus Naegleria (Protista: Rhizopoda). *Mol. Biol. Evol.* **6,** 243–257 (1989).

612. Mowbrey, K. & Dacks, J. B. Evolution and diversity of the Golgi body. *FEBS Lett.* **583,** 3738–3745 (2009).

613. Zysset-Burri, D. C. *et al.* Genome-wide identification of pathogenicity factors of the free-living amoeba Naegleria fowleri. *BMC Genomics* **15,** 496 (2014).

614. Maruyama, S. & Nozaki, H. Sequence and Intranuclear Location of the Extrachromosomal rDNA Plasmid of the Amoebo-Flagellate Naegleria gruberi. *J. Eukaryot. Microbiol.* **54,** 333–337 (2007).

615. Burger, G., Gray, M. W. & Franz Lang, B. Mitochondrial genomes: anything goes. *Trends Genet.* **19,** 709–716 (2003).

616. Lang, B. F. *et al.* An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. *Nature* **387,** 493–497 (1997).

617. Pernin, P., Ataya, A. & Cariou, M. L. Genetic structure of natural populations of the free-living amoeba, Naegleria lovaniensis. Evidence for sexual reproduction. *Heredity (Edinb).* **68,** 173–181 (1992).

618. Rowbotham, T. J. Preliminary report on the pathogenicity of Legionella pneumophila for freshwater and soil amoebae. *J. Clin. Pathol.* **33,** 1179–1183 (1980).

619. Inglis, T. J. J. *et al.* Interaction between Burkholderia pseudomallei and Acanthamoeba species results in coiling phagocytosis, endamebic bacterial survival, and escape. *Infect. Immun.* **68,** 1681–1686 (2000).

620. Ohno, H., Fournier, M. C., Poy, G. & Bonifacino, J. S. Structural determinants of interaction of tyrosine-based sorting signals with the adaptor medium chains. *J. Biol. Chem.* **271,** 29009–29015 (1996).

621. Hirst, J. & Robinson, M. S. Clathrin and adaptors. *Biochim. Biophys. Acta - Mol. Cell Res.* **1404,** 173–193 (1998).

622. Hirst, J., Bright, N. A., Rous, B. & Robinson, M. S. Characterization of a fourth adaptor-related protein complex. *Mol. Biol. Cell* **10,** 2787–802 (1999).

623. Ogata, M. *et al.* Autophagy is activated for cell survival after endoplasmic reticulum stress. *Mol. Cell. Biol.* **26,** 9220–31 (2006).

624. Goldshmidt, H. *et al.* Persistent ER stress induces the spliced leader RNA silencing pathway (SLS), leading to programmed cell death in Trypanosoma brucei. *PLoS Pathog.* **6,** e1000731 (2010).

625. Määttänen, P., Gehring, K., Bergeron, J. J. M. & Thomas, D. Y. Protein quality control in the ER: the recognition of misfolded proteins. *Semin. Cell Dev. Biol.* **21,** 500–11 (2010).

626. Axe, E. L. *et al.* Autophagosome formation from membrane compartments enriched in phosphatidylinositol 3-phosphate and dynamically connected to the endoplasmic reticulum. *J. Cell Biol.* **182,** 685–701 (2008).

627. Chew, L. H. & Yip, C. K. Structural biology of the macroautophagy machinery. *Front. Biol. (Beijing).* **9,** 18–34 (2014).

628. Meinecke, M. *et al.* The peroxisomal importomer constitutes a large and highly dynamic pore. *Nat. Cell Biol.* **12,** 273–277 (2010).

629. Gonzalez, N. H. *et al.* A single peroxisomal targeting signal mediates matrix protein import in diatoms. *PLoS One* **6,** (2011).

630. Matsuura, A., Tsukada, M., Wada, Y. & Ohsumi, Y. Apg1p, a novel protein kinase required for the autophagic process in *Saccharomyces cerevisiae*. *Gene* **192,** 245–250 (1997).

631. Warnecke, D. *et al.* Cloning and functional expression of UGT genes encoding sterol glucosyltransferases from Saccharomyces cerevisiae, Candida albicans, Pichia pastoris, and Dictyostelium discoideum. *J. Biol. Chem.* **274,** 13048–13059 (1999).

632. Sousa, M. & Parodi, A. J. The molecular basis for the recognition of misfolded glycoproteins by the UDP-GIc:glycoprotein glucosyltransferase. *EMBO Rep.* **14,** 4196–4203 (1995).

633. Christianson, J. C., Shaler, T. A., Tyler, R. E. & Kopito, R. R. OS-9 and GRP94 deliver mutant alpha1-antitrypsin to the Hrd1-SEL1L ubiquitin ligase complex for ERAD. *Nat. Cell Biol.* **10,** 272–82 (2008).

634. Spork, S. *et al.* An unusual ERAD-like complex is targeted to the apicoplast of Plasmodium falciparum. *Eukaryot. Cell* **8,** 1134–1145 (2009).

635. Hempel, F., Bullmann, L., Lau, J., Zauner, S. & Maier, U. G. ERAD-derived preprotein transport across the second outermost plastid membrane of diatoms. *Mol. Biol. Evol.* **26,** 1781–90 (2009).

636. Horn, S. C. *et al.* Usa1 Functions as a Scaffold of the HRD-Ubiquitin Ligase. *Mol. Cell* **36,** 782–793 (2009).

637. Zattas, D. & Hochstrasser, M. Ubiquitin-dependent Protein Degradation at the Yeast Endoplasmic Reticulum and Nuclear Envelope. *Crit Rev Biochem Mol Biol* **50,** 1–17 (2015).

638. Szathmary, R., Bielmann, R., Nita-Lazar, M., Burda, P. & Jakob, C. A. Yos9 protein is essential for degradation of misfolded glycoproteins and may function as lectin in ERAD. *Mol. Cell* **19,** 765–775 (2005).

639. Oda, Y. *et al.* Derlin-2 and Derlin-3 are regulated by the mammalian unfolded protein response and are required for ER-associated degradation. *J. Cell Biol.* **172,** 383–393 (2006).

640. Richly, H. *et al.* A series of ubiquitin binding factors connects CDC48/p97 to substrate multiubiquitylation and proteasomal targeting. *Cell* **120,** 73–84 (2005).

641. Suzuki, T., Park, H., Hollingsworth, N. M., Sternglanz, R. & Lennarz, W. J. PNG1, a yeast gene encoding a highly conserved peptide:N-glycanase. *J. Cell Biol.* **149,** 1039–1052 (2000).

642. Spycher, C. *et al.* An ER-directed transcriptional response to unfolded protein stress in the

absence of conserved sensor-transducer proteins in Giardia lamblia. *Mol. Microbiol.* **88,** 754–771 (2013).

643. Chow, C., Cloutier, S., Dumas, C., Chou, M.-N. & Papadopoulou, B. Promastigote to amastigote differentiation of Leishmania is markedly delayed in the absence of PERK eIF2alpha kinase-dependent eIF2alpha phosphorylation. *Cell. Microbiol.* **13,** 1059–77 (2011).

644. Joyce, B. R., Tampaki, Z., Kim, K., Wek, R. C. & Sullivan, W. J. The unfolded protein response in the protozoan parasite Toxoplasma gondii features translational and transcriptional control. *Eukaryot. Cell* **12,** 979–89 (2013).

645. Brown, M. S. & Goldstein, J. L. The SREBP Pathway: Regulation of Cholesterol Metabolism by Proteolysis of a Membrane-Bound Transcription Factor. *Cell* **89,** 331–340 (1997).

646. Kawahara, T., Yanagi, H., Yura, T. & Mori, K. Unconventional Splicing of HAC1/ERN4 mRNA Required for the Unfolded Protein Response: Sequence-Specific and Non-Sequential Cleavage of the Splice Sites. *J. Biol. Chem.* **273,** 1802–1807 (1998).

647. Liu, J.-X. & Howell, S. H. bZIP28 and NF-Y transcription factors are activated by ER stress and assemble into a transcriptional complex to regulate stress response genes in Arabidopsis. *Plant Cell* **22,** 782–96 (2010).

648. Nagashima, Y. *et al.* Arabidopsis IRE1 catalyses unconventional splicing of bZIP60 mRNA to produce the active transcription factor. *Sci. Rep.* **1,** 1–10 (2011).

649. Yoshida, H., Matsui, T., Yamamoto, A., Okada, T. & Mori, K. XBP1 mRNA Is Induced by ATF6 and Spliced by IRE1 in Response to ER Stress to Produce a Highly Active Transcription Factor. *Cell* **107,** 881–891 (2001).

650. Shen, J., Chen, X., Hendershot, L. & Prywes, R. ER Stress Regulation of ATF6 Localization by Dissociation of BiP/GRP78 Binding and Unmasking of Golgi Localization Signals. *Dev. Cell* **3,** 99–111 (2002).

651. Jamerson, M., da Rocha-Azevedo, B., Cabral, G. A. & Marciano-Cabral, F. Pathogenic Naegleria fowleri and non-pathogenic Naegleria lovaniensis exhibit differential adhesion to, and invasion of, extracellular matrix proteins. *Microbiology* **158,** 791–803 (2012).

652. Sebé-Pedrós, A. & Ruiz-Trillo, I. Integrin-mediated adhesion complex. *Commun. Integr. Biol.* **3,** 475–477 (2010).

653. Dorsam, R. T. & Gutkind, J. S. G-protein-coupled receptors and cancer. *Nat. Rev. Cancer* **7,** 79–94 (2007).

654. Emery, S. J. *et al.* Induction of virulence factors in Giardia duodenalis independent of host attachment. *Sci. Rep.* **6,** 20765 (2016).

655. Beckmann, J., Schubert, R., Chiquet-Ehrismann, R. & Müller, D. J. Deciphering teneurin domains that facilitate cellular recognition, cell-cell adhesion, and neurite outgrowth using atomic force microscopy-based single-cell force spectroscopy. *Nano Lett.* **13,** 2937–2946 (2013).

656. Xiao, D. *et al.* Ephrin/Eph receptor expression in brain of adult nonhuman primates: Implications for neuroadaptation. *Brain Res.* **1067,** 67–77 (2006).

657. Froquet, R. *et al.* TM9/Phg1 and SadA proteins control surface expression and stability of SibA adhesion molecules in Dictyostelium. *Mol. Biol. Cell* **23,** 679–686 (2012).

658. Benghezal, M. *et al.* Synergistic Control of Cellular Adhesion by Transmembrane 9 Proteins. *Mol. Biol. Cell* **14,** 2890–2899 (2003).

659. Caffrey, C. R. & Steverding, D. Kinetoplastid papain-like cysteine peptidases. *Mol. Biochem. Parasitol.* **167,** 12–19 (2009).

660. Nikolskaia, O. V *et al.* Blood-brain barrier traversal by African trypanosomes requires calcium signaling induced by parasite cysteine protease. *J. Clin. Invest.* **116,** 2739–47 (2006).

661. Dou, Z., Carruthers, V. B., Robinson, M. W. & Dalton, J. P. in *Cysteine Proteases of Pathogenic Organisms* **712,** 49–61 (Springer US, 2011).

662. Law, R. H. P. *et al.* Cloning and Expression of the Major Secreted Cathepsin B-Like Protein from Juvenile Fasciola hepatica and Analysis of Immunogenicity following Liver Fluke Infection. *Infect. Immun.* **71,** 6921–6932 (2003).

663. Zhang, Z. *et al.* Structure of the yeast vacuolar ATPase. *J. Biol. Chem.* **283,** 35983–35995 (2008).

664. Kane, P. M. Disassembly and Reassembly of the Yeast Vacoular H+ ATPase in vivo. *The Journal of Biological Chemistry,* **270,** 17025–17032 (1995).

665. Sumner, J. P. *et al.* Regulation of plasma membrane V-ATPase activity by dissociation of peripheral subunits. *Journal of Biological Chemistry* **270,** 5649–5653 (1995).

666. Maruzs, T. *et al.* Retromer Ensures the Degradation of Autophagic Cargo by Maintaining Lysosome Function in Drosophila. *Traffic* **16,** 1088–1107 (2015).

667. Gomez, T. S. & Billadeau, D. D. A FAM21-Containing WASH Complex Regulates Retromer-Dependent Sorting. *Dev. Cell* **17,** 699–711 (2009).

668. King, J. S. *et al.* WASH is required for lysosomal recycling and efficient autophagic and phagocytic digestion. *Mol. Biol. Cell* **24,** 2714–2726 (2013).

669. Shang, N. *et al.* Squalene Synthase As a Target for Chagas Disease Therapeutics. *PLoS Pathog.* **10,** (2014).

670. Van Klinken, B. J., Dekker, J., Büller, H. a & Einerhand, a W. Mucin gene structure and expression: protection vs. adhesion. *Am. J. Physiol.* **269,** 613–627 (1995).

671. Rowland, A. F., Larance, M., Hughes, W. E. & James, D. E. Identification of RhoGAP22 as an Akt-dependent regulator of cell motility in response to insulin. *Mol. Cell. Biol.* **31,** 4789–800 (2011).

672. Baek, S. H. *et al.* Requirement for Pak3 in Rac1-induced organization of actin and myosin during Drosophila larval wound healing. *FEBS Lett.* **586,** 772–777 (2012).

673. Cloutier, P., Lavallée-Adam, M., Faubert, D., Blanchette, M. & Coulombe, B. A Newly Uncovered Group of Distantly Related Lysine Methyltransferases Preferentially Interact with Molecular Chaperones to Regulate Their Activity. *PLoS Genet.* **9,** e1003210 (2013).

674. Tamura, N. *et al.* Atg18 phosphoregulation controls organellar dynamics by modulating its phosphoinositidebinding activity. *J. Cell Biol.* **202,** 685–698 (2013).

675. Kolter, T. & Sandhoff, K. Principles of Lysosomal Membrane Digestion: Stimulation of Sphingolipid Degradation by Sphingolipid Activator Proteins and Anionic Lysosomal Lipids. *Annu. Rev. Cell Dev. Biol.* **21,** 81–103 (2005).

676. Gutsmann, T. *et al.* Interaction of amoebapores and NK-lysin with symmetric phospholipid and asymmetric lipopolysaccharide/phospholipid bilayers. *Biochemistry* **42,** 9804–9812 (2003).

677. Lee, J. Y. *et al.* Hemolytic activity and developmental expression of pore-forming peptide, clonorin. *Biochem. Biophys. Res. Commun.* **296,** 1238–1244 (2002).

678. Leippe, M., Ebel, S., Schoenberger, O. L., Horstmann, R. D. & Müller-Eberhard, H. J. Pore-forming peptide of pathogenic Entamoeba histolytica. *Proc. Natl. Acad. Sci. U. S. A.* **88,** 7659–63 (1991).

679. Espino, A. M. & Hillyer, G. V. Molecular cloning of a member of the Fasciola hepatica saposin-like protein family. *J. Parasitol.* **89,** 545–552 (2003).

680. Willis, C. *et al.* Insights into the membrane interactions of the saposin-like proteins Na-SLP-1 and Ac-SLP-1 from human and dog hookworm. *PLoS One* **6,** (2011).

681. Dyková, I., Pecková, H., Fiala, I. & Dvořáková, H. *Filamoeba sinensis* sp. n., a second species of the genus *Filamoeba* Page, 1967, isolated from gills of *Carassius gibelio* (Bloch, 1782). *Acta Protozool.* **44,** 75–80 (2005).

682. Runeberg-Roos, P. & Saarma, M. Phytepsin, a barley vacuolar aspartic proteinase, is highly expressed during autolysis of developing tracheary elements and sieve cells. *Plant J.* **15,** 139–145 (1998).

683. Runeberg-Roos, P., Tormakangas, K. & Ostman, A. Primary structure of a barley-grain aspartic proteinase: A plant aspartic proteinase resembling mammalian cathepsin D. *Eur. J. Biochem.* **202,** 1021–1027 (1991).

684. Kaneko, T. *et al.* Assembly Pathway of the Mammalian Proteasome Base Subcomplex Is Mediated by Multiple Specific Chaperones. *Cell* **137,** 914–925 (2009).

685. Speroni, J., Federico, M. B., Mansilla, S. F., Soria, G. & Gottifredi, V. Kinase-independent function of checkpoint kinase 1 (Chk1) in the replication of damaged DNA. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 1–6 (2012).

686. Smith, E. F. & Lefebvre, P. A. PF20 gene product contains WD repeats and localizes to the intermicrotubule bridges in Chlamydomonas flagella. *Mol. Biol. Cell* **8,** 455–67 (1997).

687. Fritz-Laylin, L. K. & Cande, W. Z. Ancestral centriole and flagella proteins identified by

analysis of Naegleria differentiation. *J. Cell Sci.* **123,** 4024–4031 (2010).

688. Dailey, H. A. *et al.* The Escherichia coli protein YfeX functions as a porphyrinogen oxidase, not a heme dechelatase. *MBio* **2,** 1–8 (2011).

689. French, C. E., Bell, J. M. L. & Ward, F. B. Diversity and distribution of hemerythrin-like proteins in prokaryotes. *FEMS Microbiol. Lett.* **279,** 131–145 (2008).

690. Sajid, M. & McKerrow, J. H. Cysteine proteases of parasitic organisms. *Mol. Biochem. Parasitol.* **120,** 1–21 (2002).

691. Rawlings, N. D., Barrett, A. J. & Finn, R. Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res.* **44,** D343–D350 (2016).

692. Zavalova, L. L. *et al.* Destabilase from the medicinal leech is a representative of a novel family of lysozymes. *Biochim. Biophys. Acta - Protein Struct. Mol. Enzymol.* **1478,** 69–77 (2000).

693. Zavalova, L. *et al.* Genes from the medicinal leech (Hirudo medicinalis) coding for unusual enzymes that specifically cleave endo-ε(γ-Glu)-Lys isopeptide bonds and help to dissolve blood clots. *Mol. Gen. Genet.* **253,** 20–25 (1996).

694. Benes, P., Vetvicka, V. & Fusek, M. Cathepsin D - Many Functions of One Aspartic Protease. *Critcical Rev. Oncol Hematol.* **68,** 12–28 (2008).

695. Somanna, A., Mundodi, V. & Gedamu, L. Functional Analysis of Cathepsin B-like Cysteine Proteases from Leishmania donovani Complex: Evidence for the Activation of Latent Transforming Growth Factor β. *J. Biol. Chem.* **277,** 25305–25312 (2002).

696. Kissoon-Singh, V., Mortimer, L., Chadee, K., Robinson, M. W. & Dalton, J. P. in *Cysteine Proteases of Pathogenic Organisms* **712,** 62–83 (Springer US, 2011).

697. De Jesus, J. B. *et al.* Cysteine peptidase expression in *Trichomonas vaginalis* isolates displaying high- and low-virulence phenotypes. *J. Proteome Res.* **8,** 1555–1564 (2009).

698. Kong, Y., Chung, Y. B., Cho, S. Y. & Kang, S. Y. Cleavage of immunoglobulin G by excretory-secretory cathepsin S-like protease of Spirometra mansoni plerocercoid. *Parasitology* **109,** 611–621 (1994).

699. Buck, M. R., Karustis, D. G., Day, N. A., Honntt, K. V & Sloane, B. F. Degradation of extracellular-matrix proteins by human cathepsin B from normal and tumour tissues. *Biochem. J* **282,** 273–278 (1992).

700. Goulet, B. *et al.* A cathepsin L isoform that is devoid of a signal peptide localizes to the nucleus in S phase and processes the CDP/Cux transcription factor. *Mol. Cell* **14,** 207–219 (2004).

701. Zhang, X. & Wang, Y. GRASPs in Golgi Structure and Function. *Front. Cell Dev. Biol.* **3,** 1–8 (2016).

702. Hehl, A. B. & Marti, M. Secretory protein trafficking in Giardia intestinalis. *Mol. Microbiol.* **53,** 19–28 (2004).

703. Marti, M. *et al.* An Ancestral Secretory Apparatus in the Protozoan Parasite Giardia intestinalis. *J. Biol. Chem.* **278,** 24837–24848 (2003).

704. Gabaldón, T. *et al.* Comparative genomics of emerging pathogens in the Candida glabrata clade. *BMC Genomics* **14,** 623 (2013).

705. Angrist, M. Less is more: Compact genomes pay dividends. *Genome Res.* **8,** 683–685 (1998).

706. Eme, L. *et al.* Lateral Gene Transfer in the Adaptation of the Anaerobic Parasite Blastocystis to the Gut Article Lateral Gene Transfer in the Adaptation of the Anaerobic Parasite Blastocystis to the Gut. *Curr. Biol.* **27,** 807–820 (2017).

707. Uemura, T. *et al.* Systematic Analysis of SNARE Molecules in Arabidopsis: Dissection of the post-Golgi Network in Plant Cells. *Cell Struct. Funct.* **29,** 49–65 (2004).

708. Song, K. J. *et al.* Naegleria fowleri: functional expression of the Nfa1 protein in transfected Naegleria gruberi by promoter modification. *Exp Parasitol* **112,** 115–120 (2006).

709. Lee, Y.-J. *et al.* Production of Nfa1-specific monoclonal antibodies that influences the in vitro cytotoxicity of Naegleria fowleri trophozoites on microglial cells. *Parasitol. Res.* **101,** 1191–1196 (2007).

710. Somlata, Nakada-Tsukui, K. & Nozaki, T. AGC family kinase 1 participates in trogocytosis but not in phagocytosis in Entamoeba histolytica. *Nat. Commun.* **8,** 101 (2017).

711. Beznoussenko, G. V. *et al.* Analogs of the Golgi complex in microsporidia: structure and avesicular mechanisms of function. *J. Cell Sci.* **120,** 1288–1298 (2007).

712. Manna, P. T., Kelly, S. & Field, M. C. Adaptin evolution in kinetoplastids and emergence of the variant surface glycoprotein coat in African trypanosomatids. *Mol. Phylogenet. Evol.* **67,** 123–128 (2013).

713. Horii, M. *et al.* CHMP7, a novel ESCRT-III-related protein, associates with CHMP4b and functions in the endosomal sorting pathway. *Biochem. J.* **400,** 23–32 (2006).

714. Tsui, M. M. & Banfield, D. K. Yeast Golgi SNARE interactions are promiscuous. *J. Cell Sci.* **113 (Pt 1,** 145–152 (2000).

715. Hsu, S. C., Hazuka, C. D., Foletti, D. L. & Scheller, R. H. Targeting vesicles to specific sites on the plasma membrane: The role of the sec6/8 complex. *Trends Cell Biol.* **9,** 150–153 (1999).

716. Ravindran, S. & Boothroyd, J. C. Secretion of proteins into host cells by Apicomplexan parasites. *Traffic* **9,** 647–656 (2008).

717. Boothroyd, J. C. & Dubremetz, J.-F. Kiss and spit: the dual roles of Toxoplasma rhoptries. *Nat. Rev. Microbiol.* **6,** 79–88 (2008).

718. Luzio, J. P., Hackmann, Y., Dieckmann, N. M. G. & Griffiths, G. M. The Biogenesis of Lysosomes and Lysosome-Related Organelles. *Cold Spring Harb. Perspect. Biol.* 1–17

(2014).

719. Theos, A. C. *et al.* Functions of Adaptor Protein (AP)-3 and AP-1 in Tyrosinase Sorting from Endosomes to Melanosomes. *4Mol Biol Cell* **16,** 5536–5572 (2005).

720. Venugopal, K. *et al.* Dual role of the Toxoplasma gondii clathrin adaptor AP1 in the sorting of rhoptry and microneme proteins and in parasite division. *PLoS Pathog.* **13,** 1–38 (2017).

721. Sparvoli, D. *et al.* An evolutionary switch in the specificity of an endosomal CORVET tether underlies formation of regulated secretory vesicles in the ciliate Tetrahymena thermophila. *bioRxiv* 1–49 (2017).

722. Lal, K., Field, M. C., Carlton, J. M., Warwicker, J. & Hirt, R. P. Identification of a very large Rab GTPase family in the parasitic protozoan Trichomonas vaginalis. *Mol. Biochem. Parasitol.* **143,** 226–235 (2005).

723. Dacks, J. B. & Doolittle, W. F. Novel syntaxin gene sequences from Giardia, Trypanosoma and algae: implications for the ancient evolution of the eukaryotic endomembrane system. *J. Cell Sci.* **115,** 1635–1642 (2002).

724. Sanderfoot, A. Increases in the Number of SNARE Genes Parallels the Rise of Multicellularity among the Green Plants. *Plant Physiol.* **144,** 6–17 (2007).

725. Katinka, M. D. *et al.* Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi. *Nature* **414,** 450–453 (2001).

726. Franzen, O. *et al.* Draft Genome Sequencing of Giardia intestinalis Assemblage B Isolate GS: Is Human Giardiasis Caused by Two Different Species? *PLoS Pathog* **5,** e1000560 (2009).

727. Okada, M. *et al.* Proteomic Analysis of Phagocytosis in the Enteric Protozoan Parasite Entamoeba histolytica. *Eukaryot. Cell* **4,** 827–831 (2005).

728. Frazier, M. N. *et al.* Molecular Basis for the Interaction Between AP4 β4 and its Accessory Protein, Tepsin. *Traffic* **17,** 400–415 (2016).

729. Archuleta, T. L. *et al.* Structure and evolution of ENTH and VHS/ENTH-like domains in tepsin. *Traffic* **18,** 590–603 (2017).

Appendix 1: Chapter 2 Supplementary Material

Supplementary Table S2.1. Trimmomatic parameters and surviving read statistics

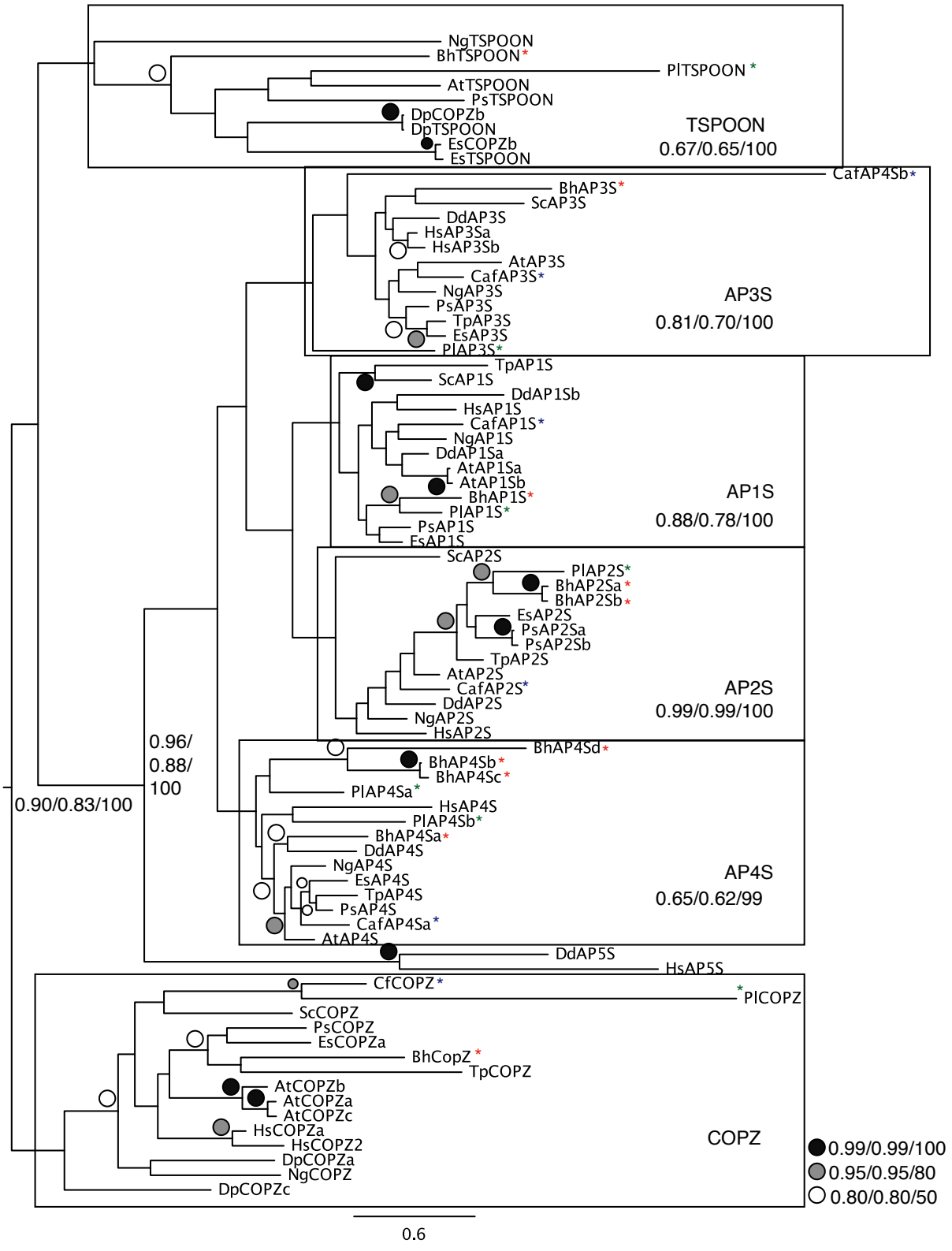
Sample name	Input parameters
4-2-LEEMP2	ILLUMINACLIP:Trimmomatic-0.32/adapters/TruSeq3-PE.fa:2:30:10 HEADCROP:15 SLIDINGWINDOW:4:20 MINLEN:50
5-2-LEEMP3	ILLUMINACLIP:Trimmomatic-0.32/adapters/TruSeq3-PE.fa:2:30:10 HEADCROP:15 SLIDINGWINDOW:4:20 MINLEN:50
LeeMP4-2_S42	ILLUMINACLIP:Trimmomatic-0.32/adapters/TruSeq3-PE.fa:2:30:10 HEADCROP:15 SLIDINGWINDOW:4:20 CROP:240 MINLEN:50
EXTRA_4-2-LeeMP2-S1_S41	ILLUMINACLIP:Trimmomatic-0.32/adapters/TruSeq3-PE.fa:2:30:10 HEADCROP:15 SLIDINGWINDOW:4:20 CROP:220 MINLEN:50
7-2-LEEA2	ILLUMINACLIP:Trimmomatic-0.32/adapters/TruSeq3-PE.fa:2:30:10 HEADCROP:15 SLIDINGWINDOW:4:20 MINLEN:50
8-2-LEEA3	ILLUMINACLIP:Trimmomatic-0.32/adapters/TruSeq3-PE.fa:2:30:10 HEADCROP:15 SLIDINGWINDOW:4:20 MINLEN:50
6-2-LEEA1	ILLUMINACLIP:Trimmomatic-0.32/adapters/TruSeq3-PE.fa:2:30:10 HEADCROP:15 SLIDINGWINDOW:4:20 MINLEN:50

Sample name	Output statistics
4-2-LEEMP2	Input Read Pairs: 3042215 Both Surviving: 2674783 (87.92%) Forward Only Surviving: 147576 (4.85%) Reverse Only Surviving: 23601 (0.78%) Dropped: 196255 (6.45%)
5-2-LEEMP3	Input Read Pairs: 2516909 Both Surviving: 2282388 (90.68%) Forward Only Surviving: 161973 (6.44%) Reverse Only Surviving: 22489 (0.89%) Dropped: 50059 (1.99%)
LeeMP4-2_S42	Input Read Pairs: 3560875 Both Surviving: 3132339 (87.97%) Forward Only Surviving: 232927 (6.54%) Reverse Only Surviving: 52803 (1.48%) Dropped: 142806 (4.01%)
EXTRA_4-2-LeeMP2-S1_S41	Input Read Pairs: 2031845 Both Surviving: 1808084 (88.99%) Forward Only Surviving: 82967 (4.08%) Reverse Only Surviving: 23353 (1.15%) Dropped: 117441 (5.78%)
7-2-LEEA2	Input Read Pairs: 3363687 Both Surviving: 3085953 (91.74%) Forward Only Surviving: 170421 (5.07%) Reverse Only Surviving: 39262 (1.17%) Dropped: 68051 (2.02%)
8-2-LEEA3	Input Read Pairs: 3234500 Both Surviving: 2866026 (88.61%) Forward Only Surviving: 203606 (6.29%) Reverse Only Surviving: 33449 (1.03%) Dropped: 131419 (4.06%)
6-2-LEEA1	Input Read Pairs: 2377575 Both Surviving: 2001555 (84.18%) Forward Only Surviving: 235542 (9.91%) Reverse Only Surviving: 30685 (1.29%) Dropped: 109793 (4.62%)

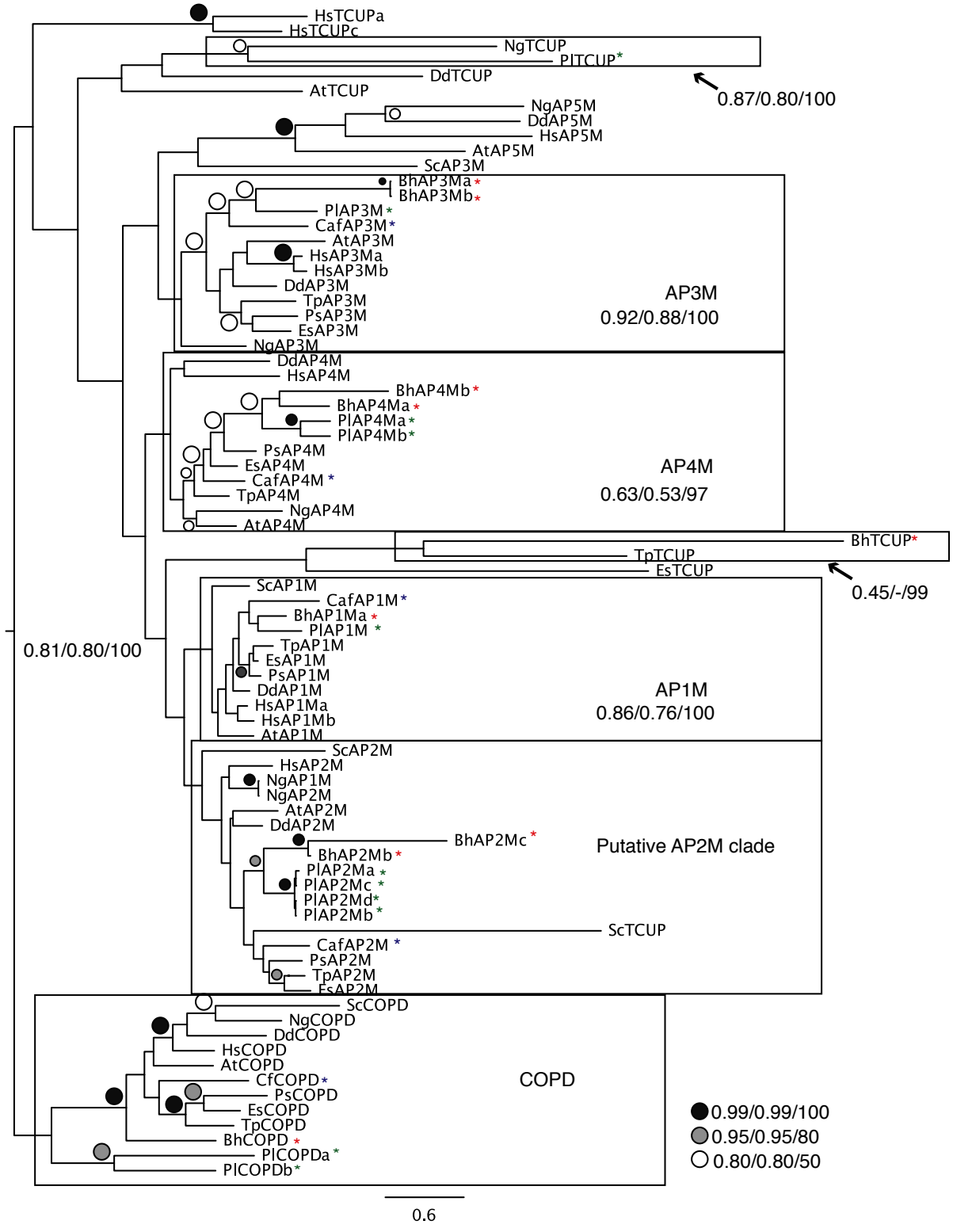
Appendix 2: Chapter 3 Supplementary Material

Supplementary Figures S3.1-3.15. Phylogenetic analyses used to classify paralogous families of membrane trafficking machinery in *Blastocystis* sp., *P. lacertae*, and *C. roenbergensis*. Node values indicating statistical support are listed as MrBAYES/Phylobayes/RAxML (posterior probability/posterior probability/bootstrap) for all trees except Supplementary Figure S3.6 (Vps24 family tree), which has MrBAYES/PhyML/RAxML values (posterior probability/bootstrap/bootstrap). The best Bayesian topology is shown with support values or symbols shown on the nodes. Important clades are boxed or shaded. Circles rather than node values indicate that the node has the minimum support shown in the legend. For SNARE and TBC family trees (Supplementary Figure S3.11-S3.14), support values are only given for major clades for legibility. Dashes instead of node values indicate clades that were not recovered using that method. *Blastocystis* sp. sequences are marked with a red asterisk, *P. lacterae* sequences are marked with a green asterisk, and *C. roenbergensis* sequences are marked with a blue asterisk. Circle size differences are for legibility only; they are not related to phylogeny quality. In the TBC tree (Supplementary Figure S3.14), sequence names are colour-coded with the above scheme for readability.

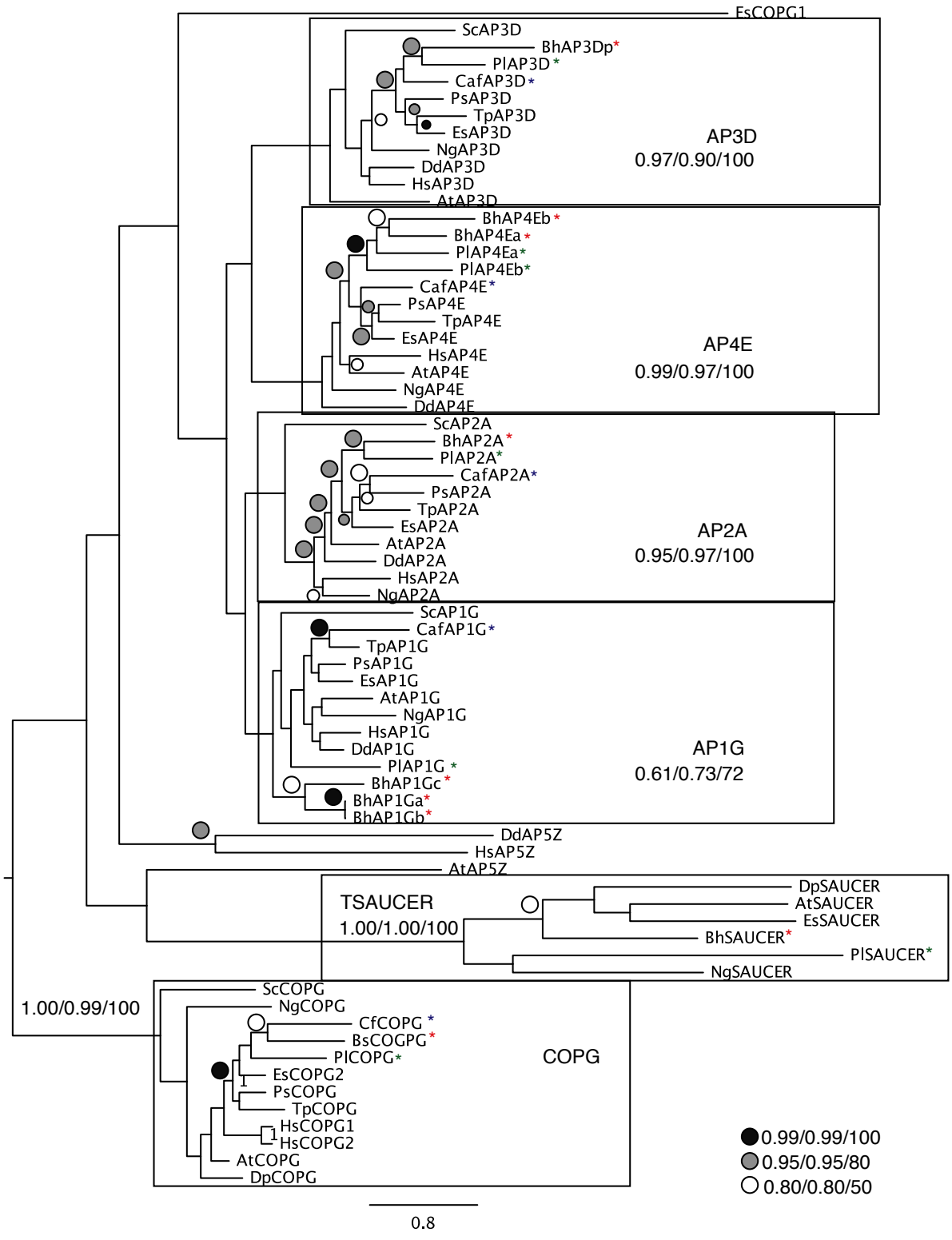
Consensus phylogenies were used to classify the following *Blastocystis* sp., *P. lacertae*, and *C. roenbergensis* sequences: S3.1, TSPOON; S3.2, TCUP sequences; S3.3, TSAUCER; S3.4, TTRAY1 and TTRAY2; S3.5, adaptin γ/α/δ/ε/ζ subunits; S3.6, adaptin β subunits; S3.7, adaptin μ subunits; S3.8, adaptin σ subunits; S3.9, Vps60, Vps20, Vps32; S3.10, Vps2, Vps24, Vps46; S3.11, Qa SNAREs; S3.12, Qb SNAREs; S3.13, Qc SNAREs; S3.14 TBC RabGAPs; S3.15, ArfGAPs. Accessions for sequences found in the trees and tree metadata are found in Online Appendix Table 3.3



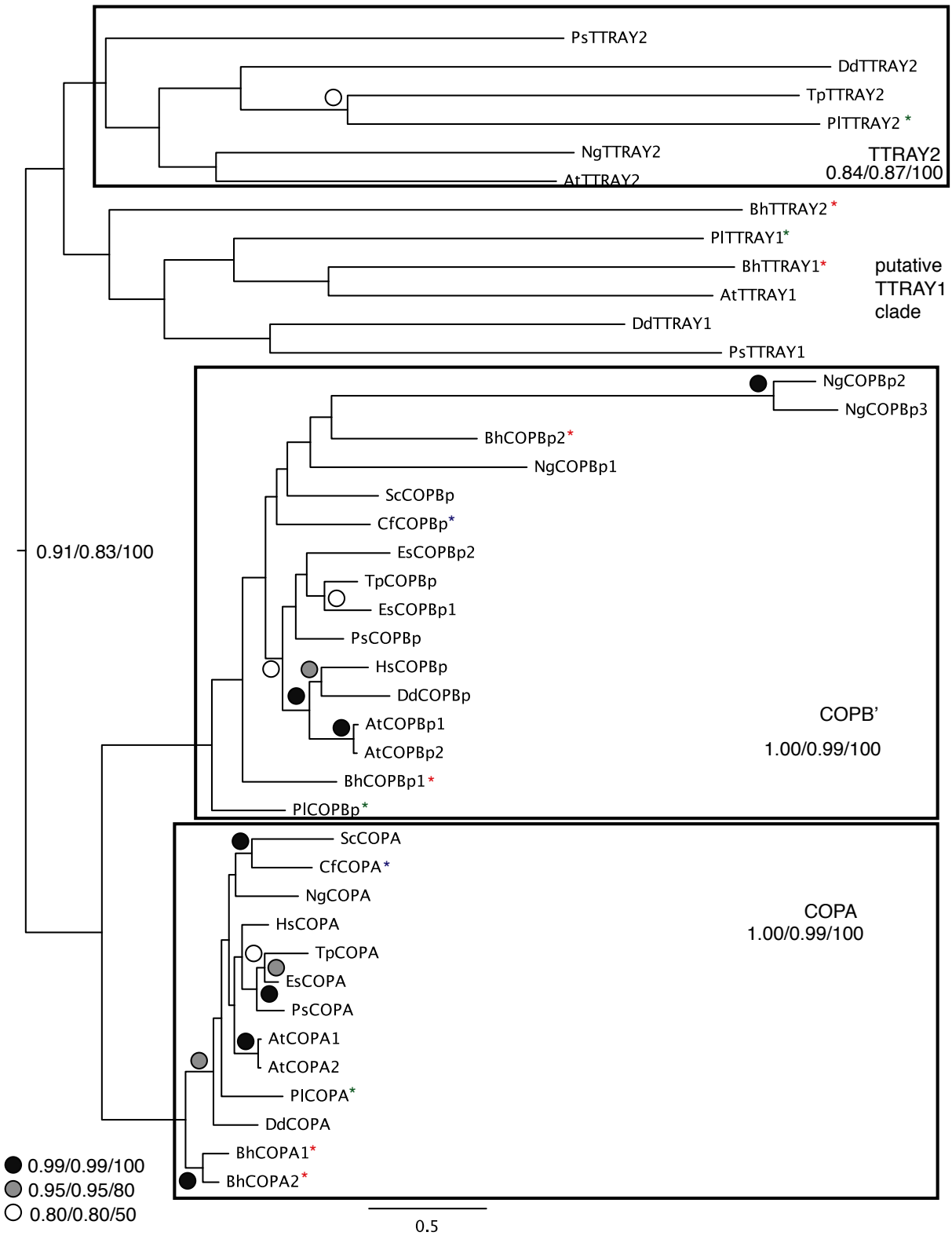
Supplementary Figure S3.1



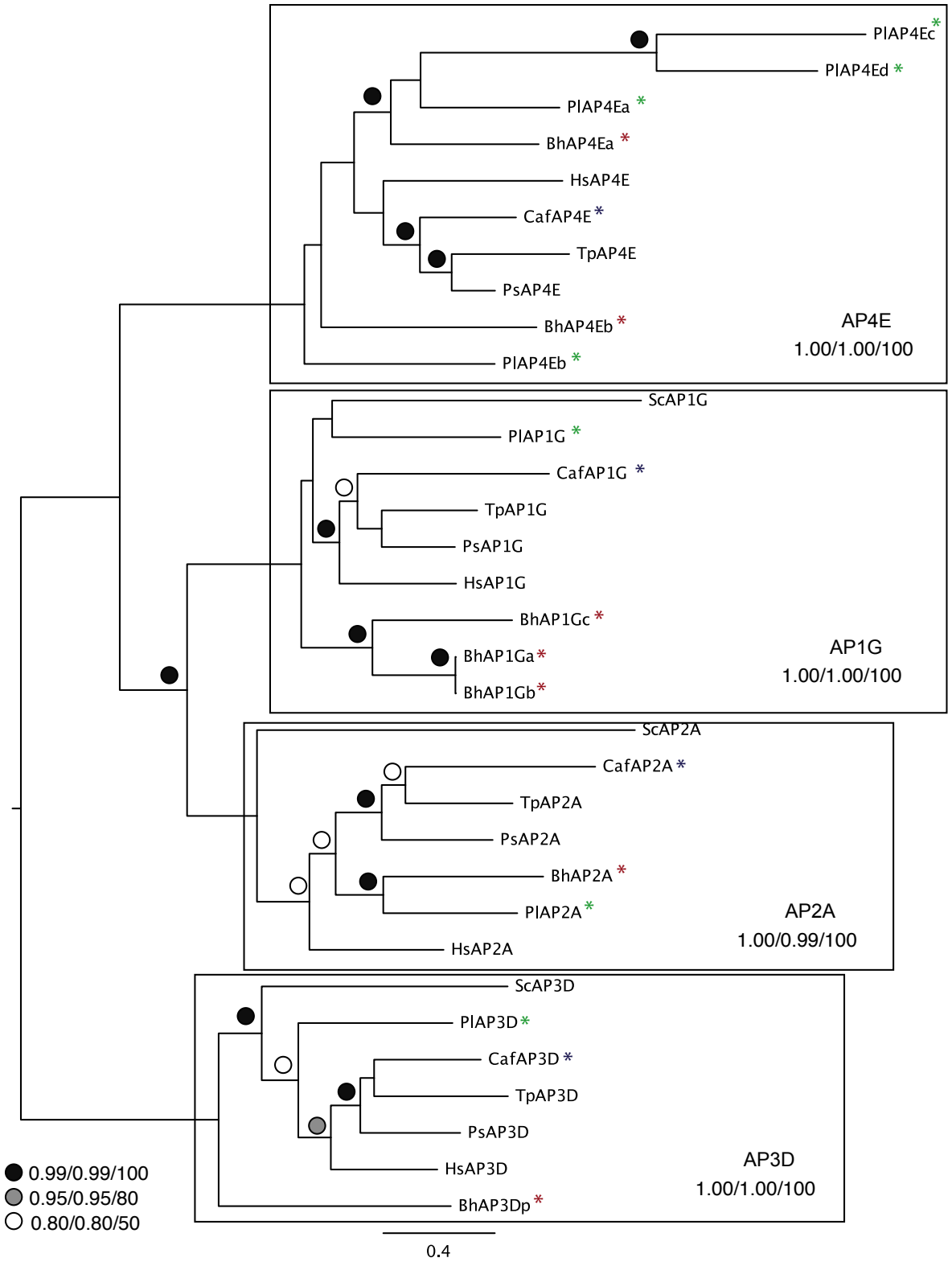
Supplementary Figure S3.2



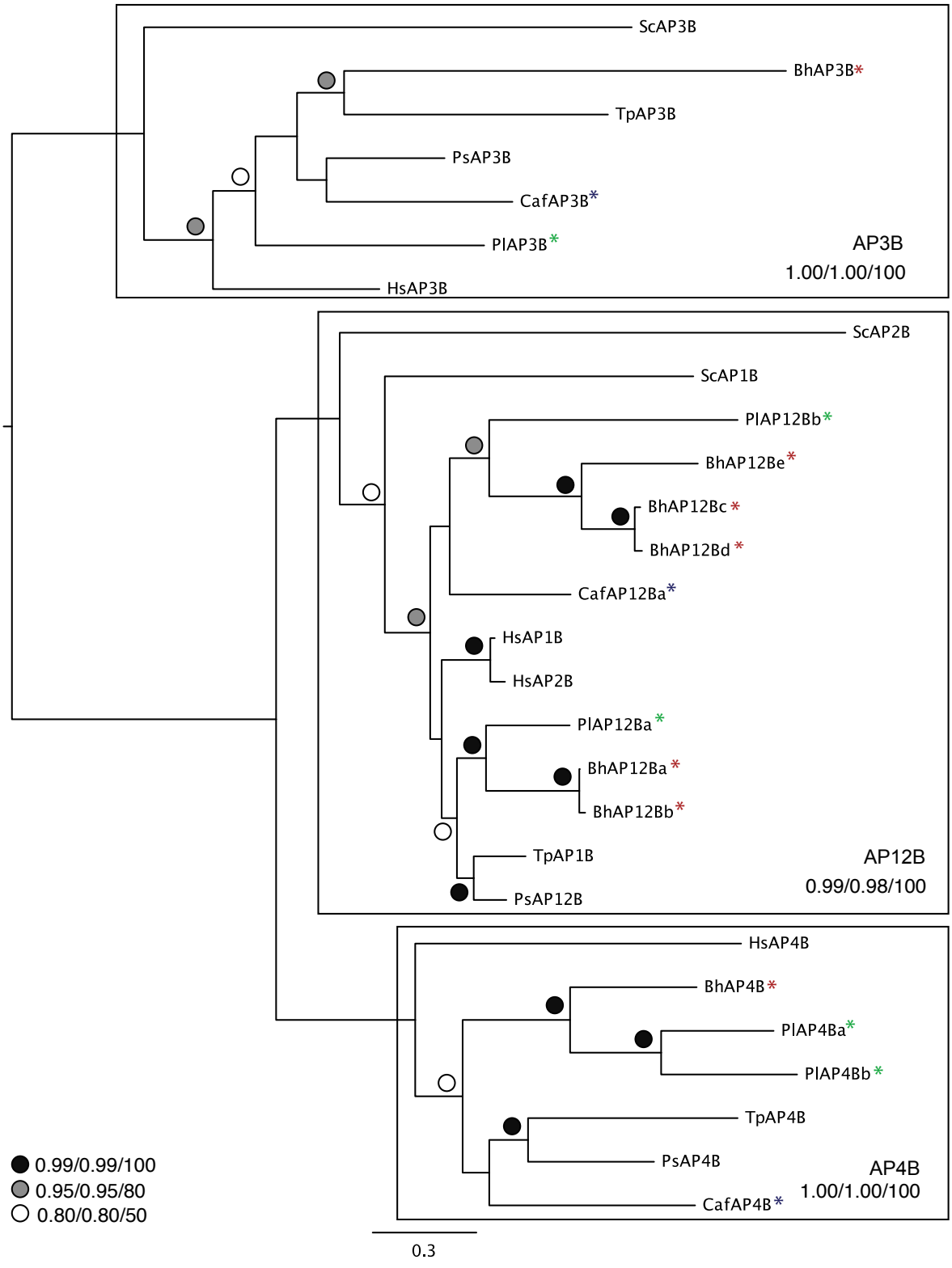
Supplementary Figure S3.3



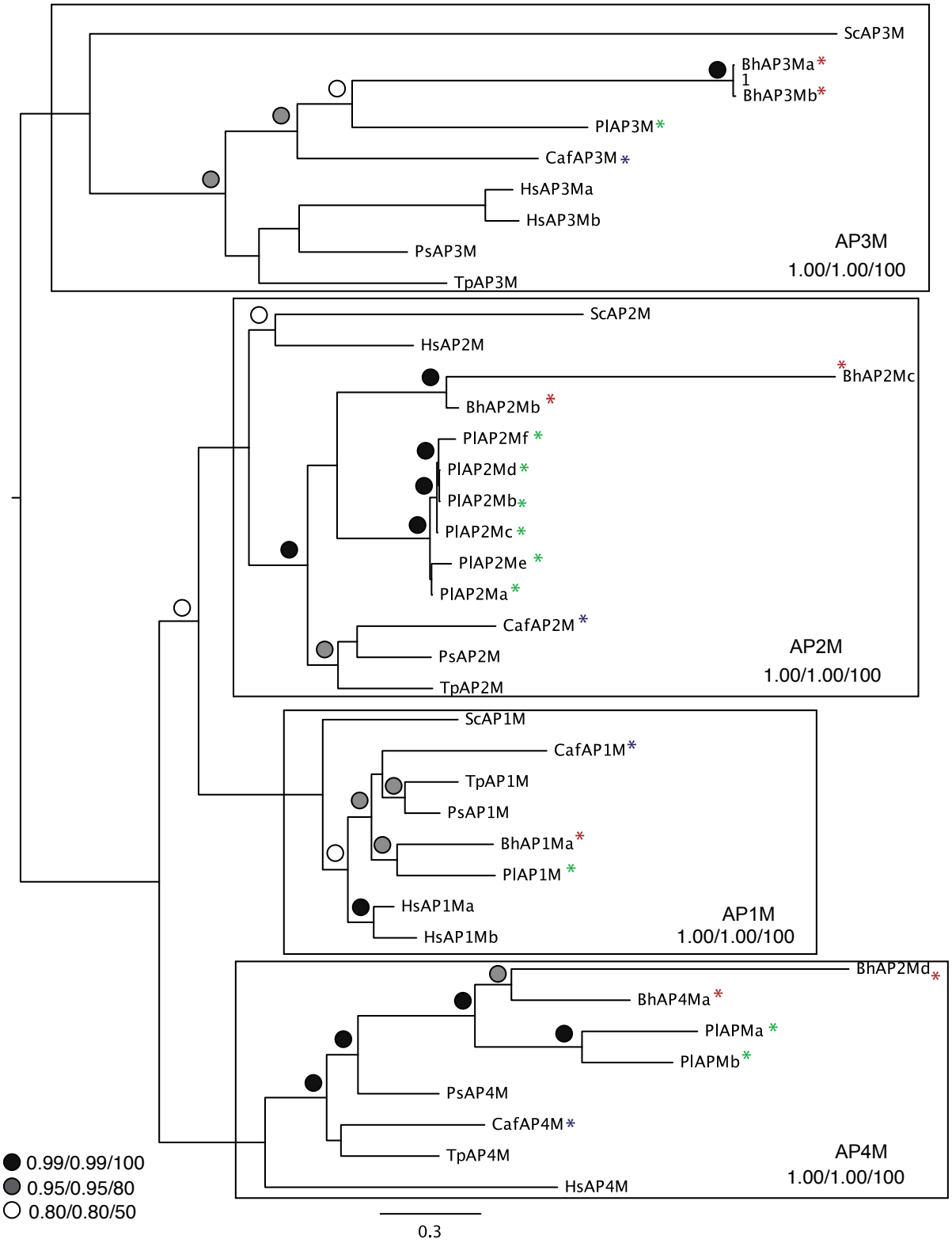
Supplementary Figure S3.4



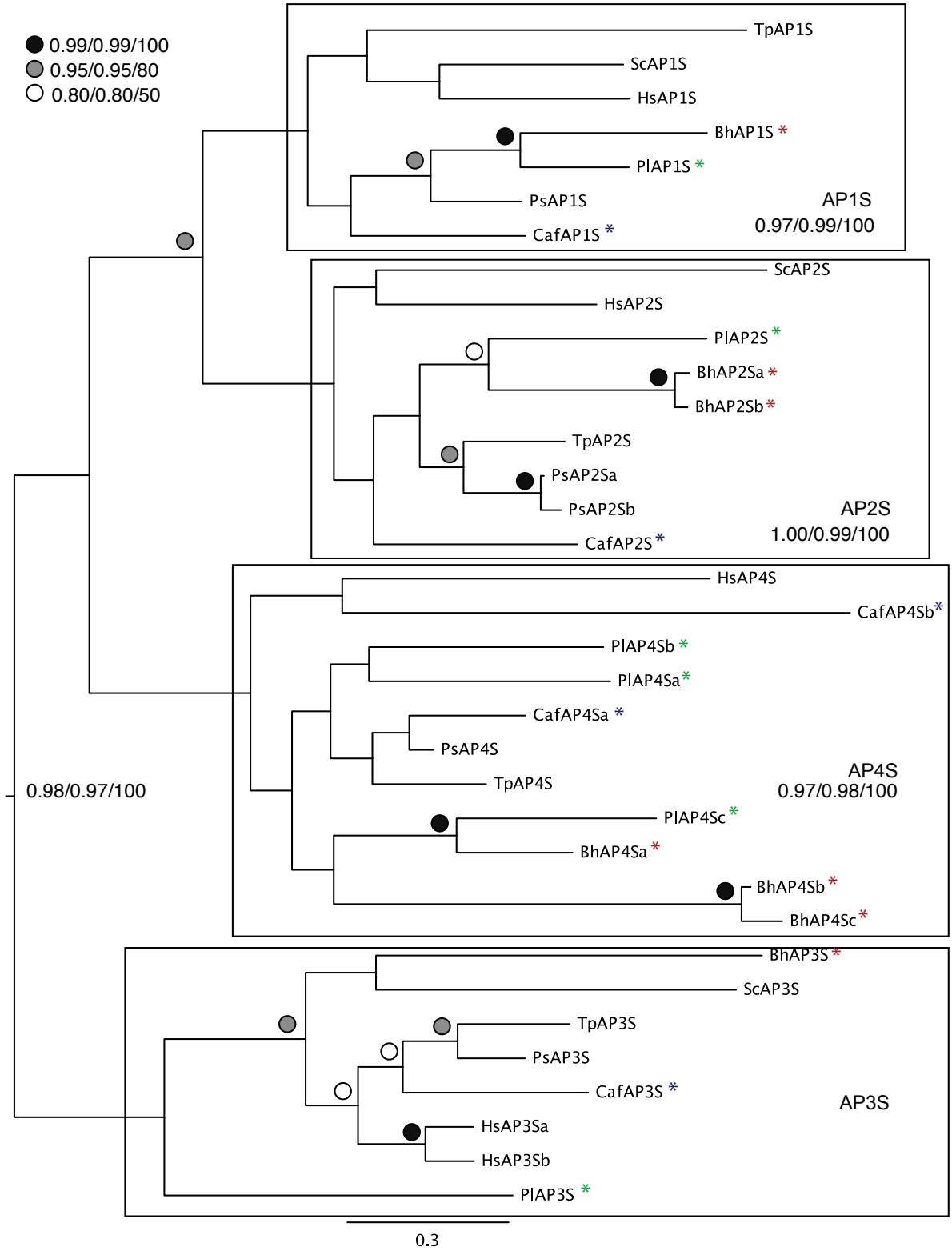
Supplementary Figure S3.5



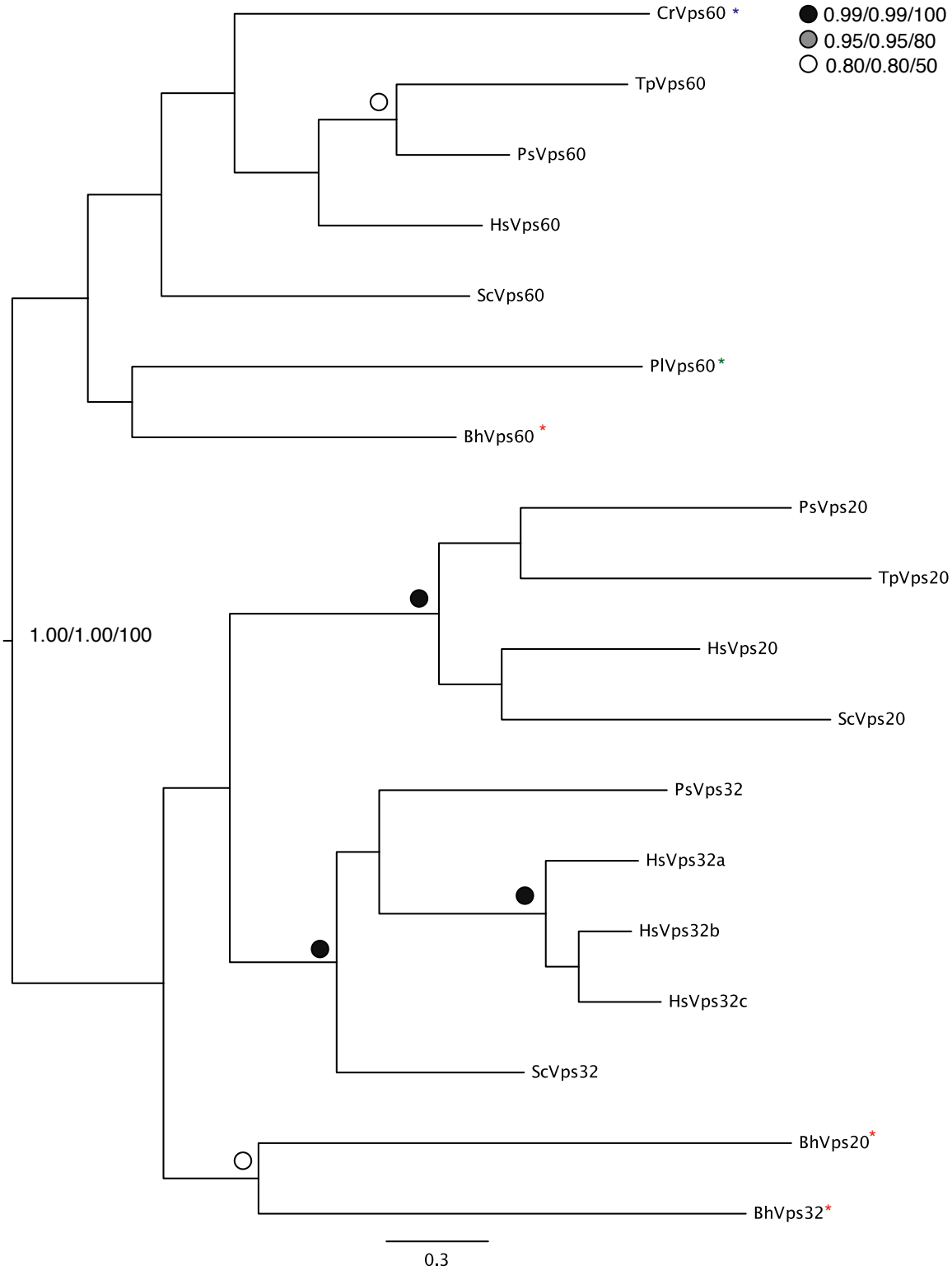
Supplementary Figure S3.6



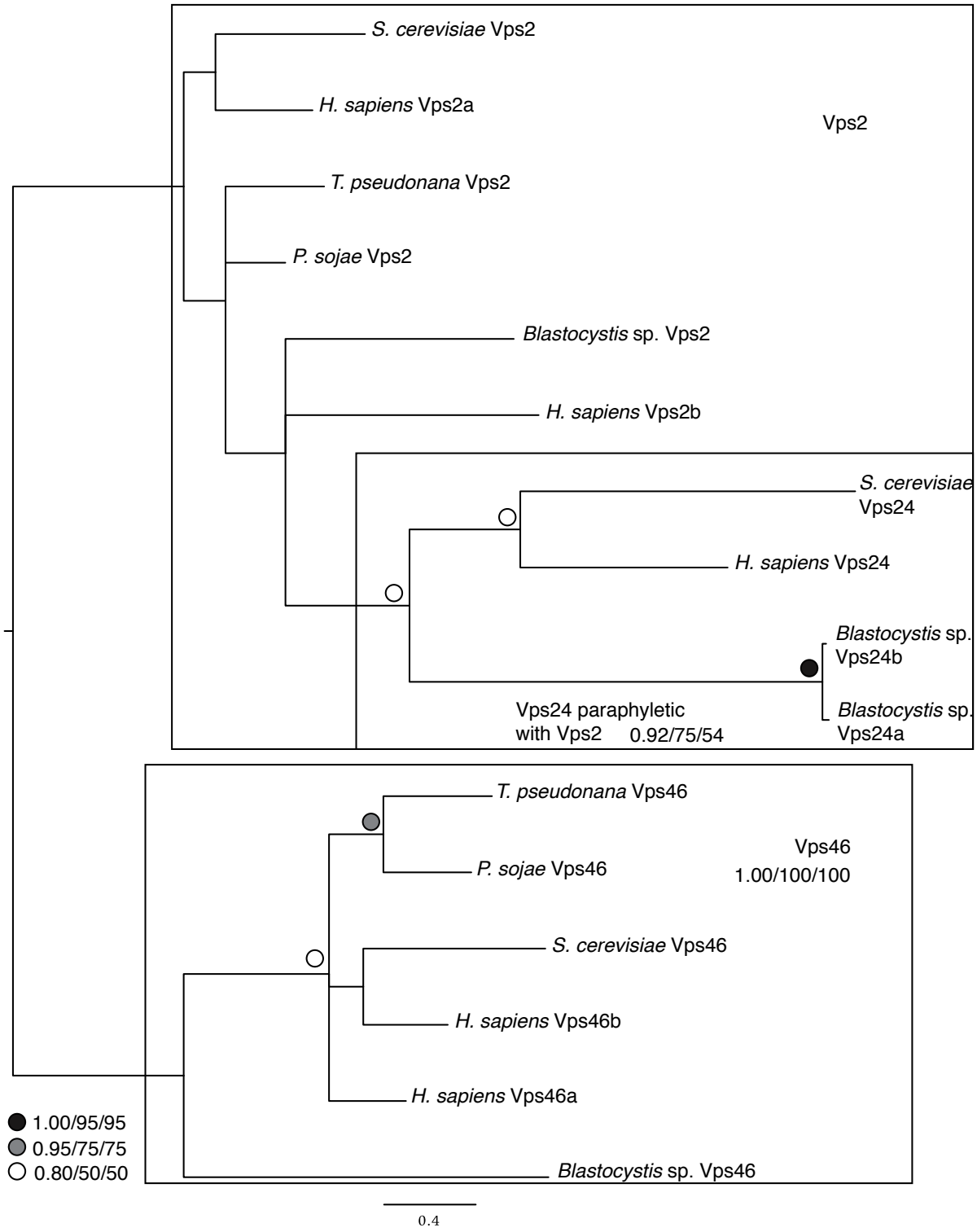
Supplementary Figure S3.7



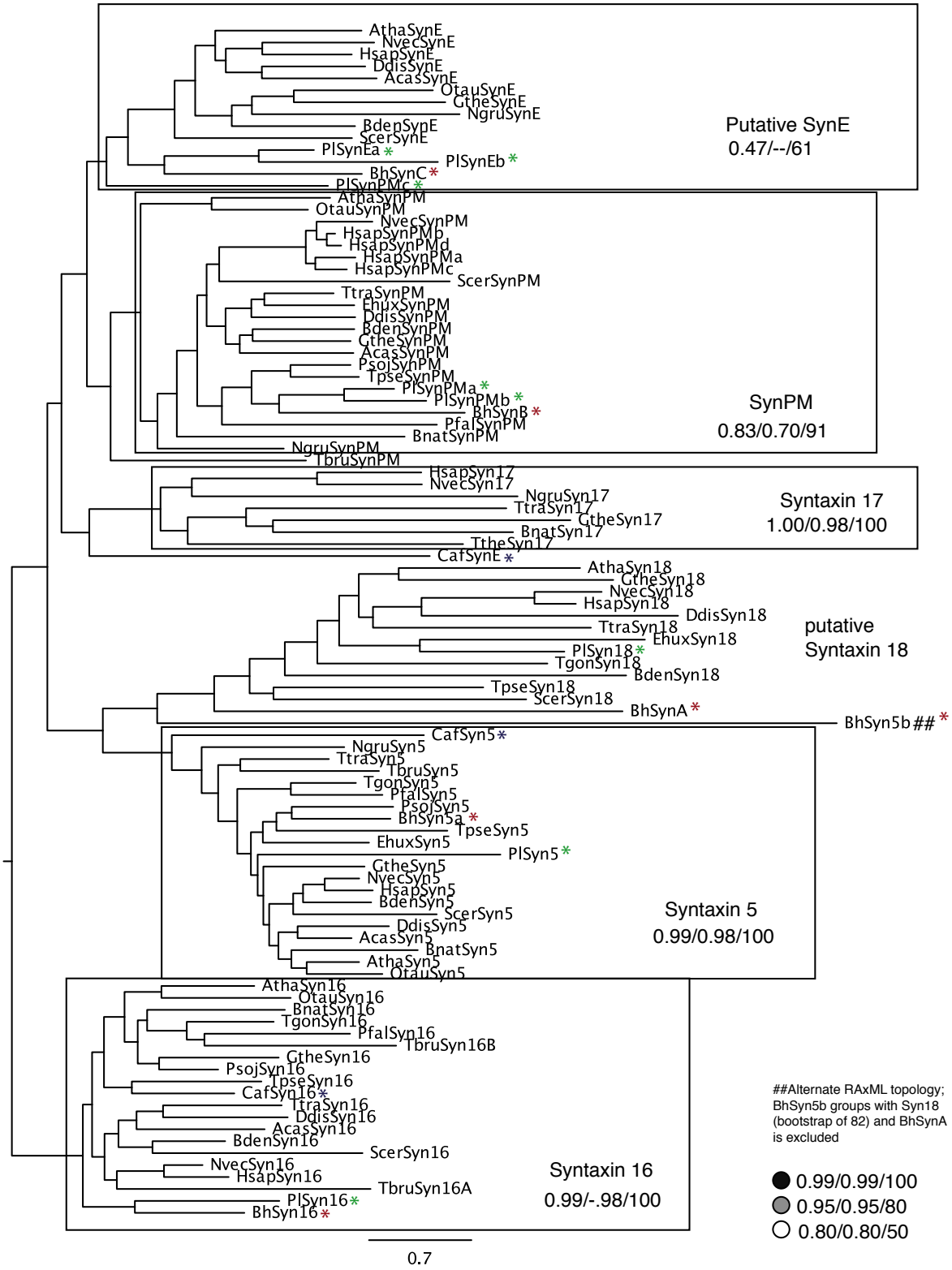
Supplementary Figure S3.8



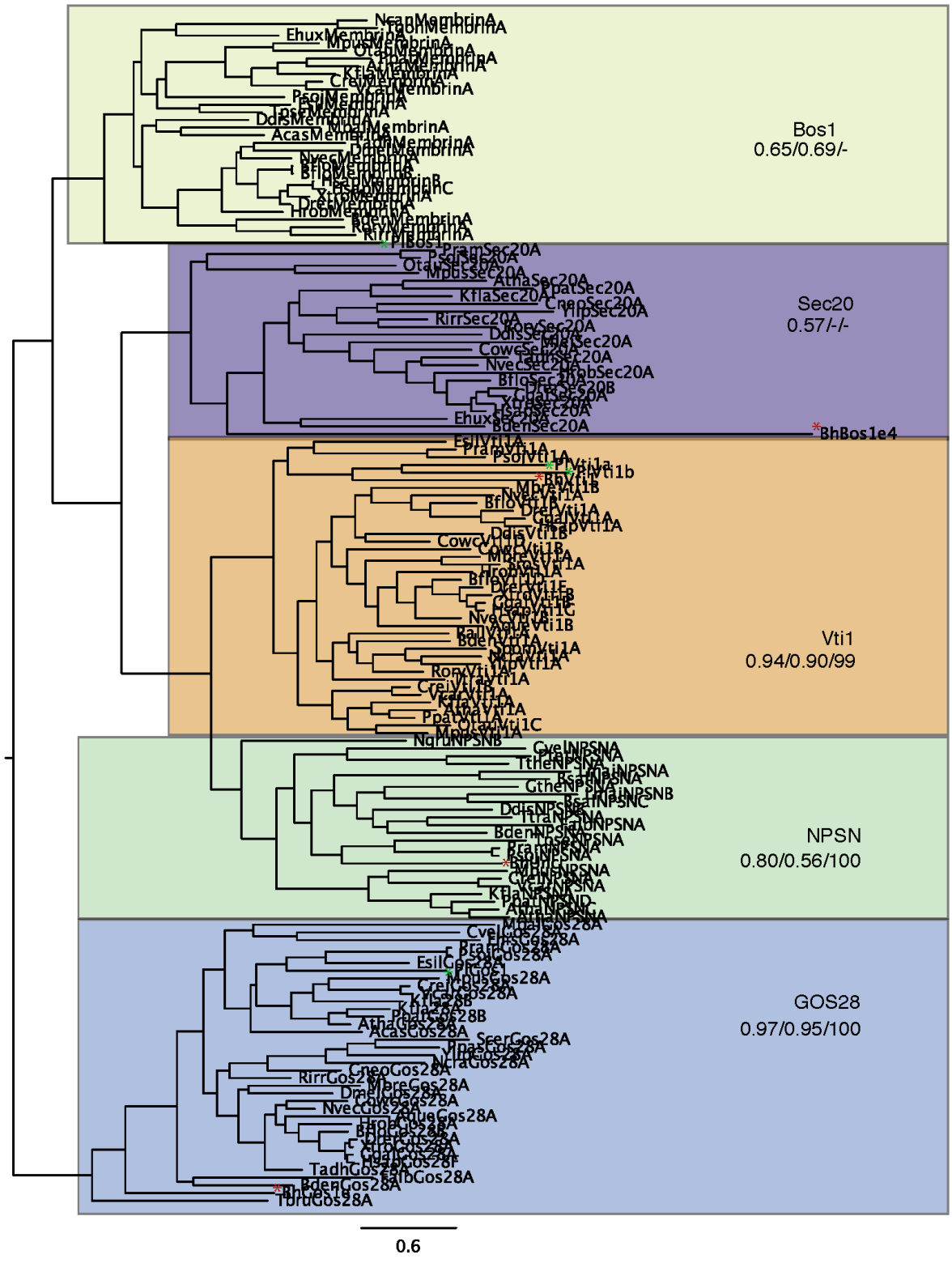
Supplementary Figure S3.9



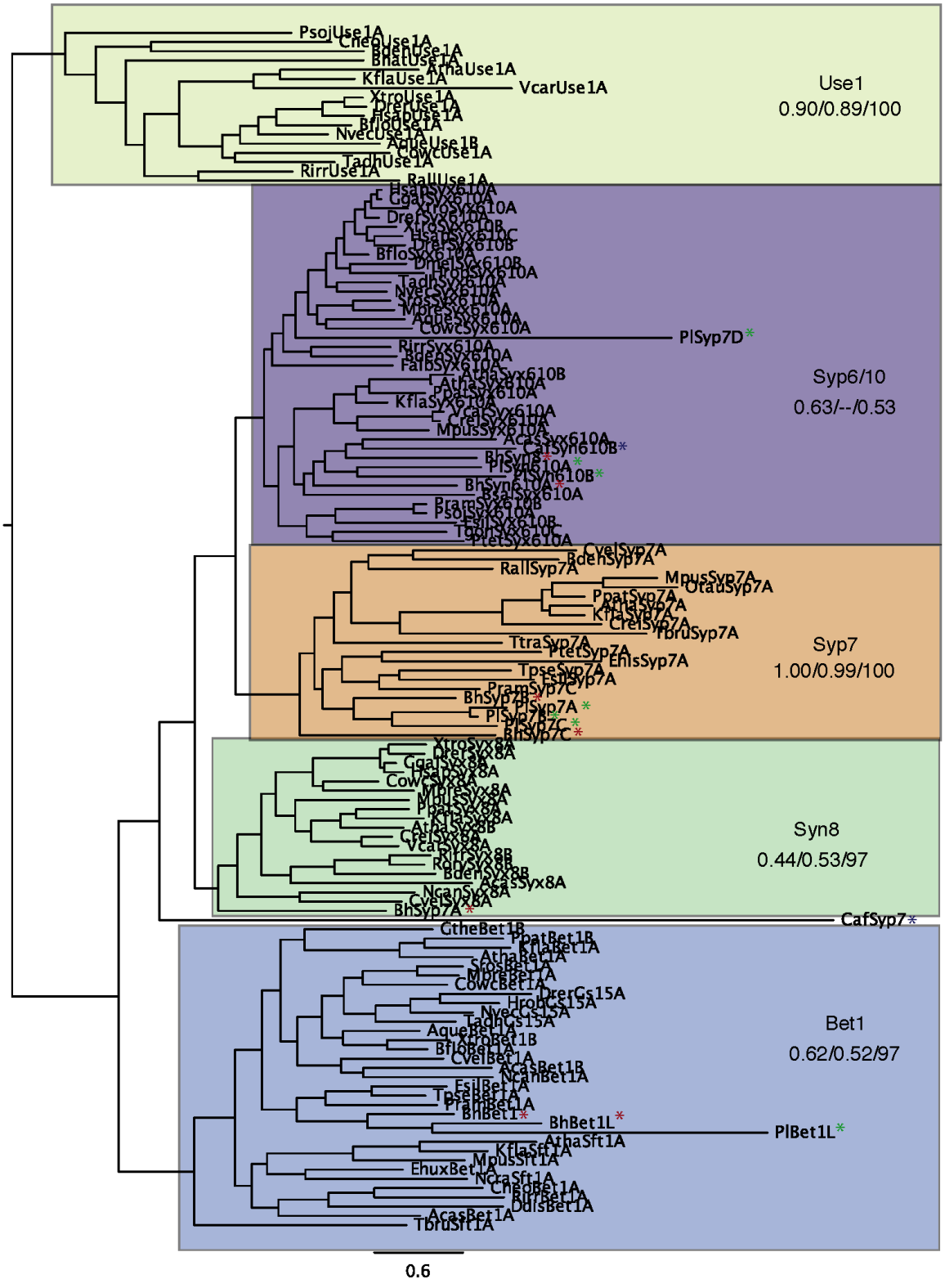
Supplementary Figure S3.10



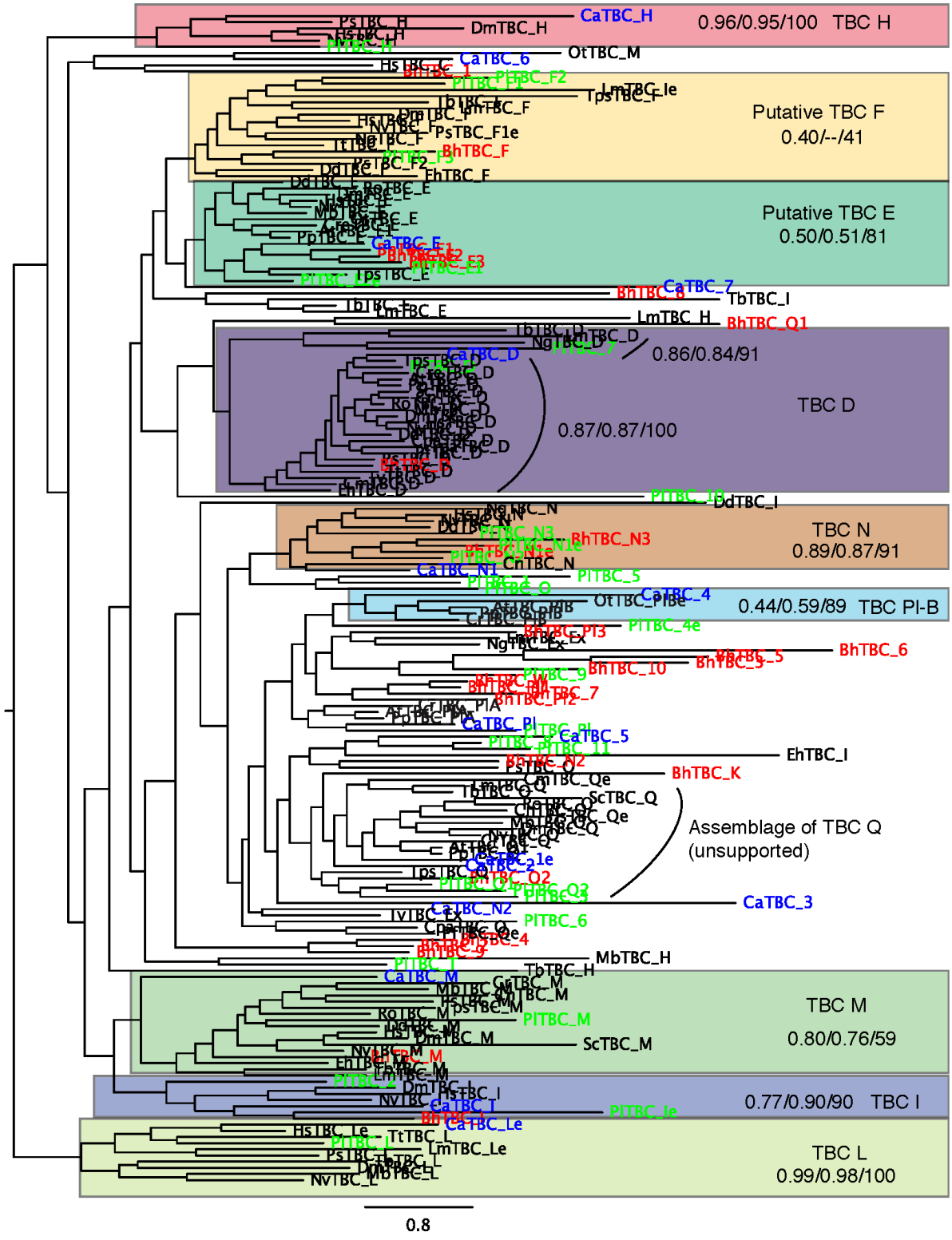
Supplementary Figure S3.11



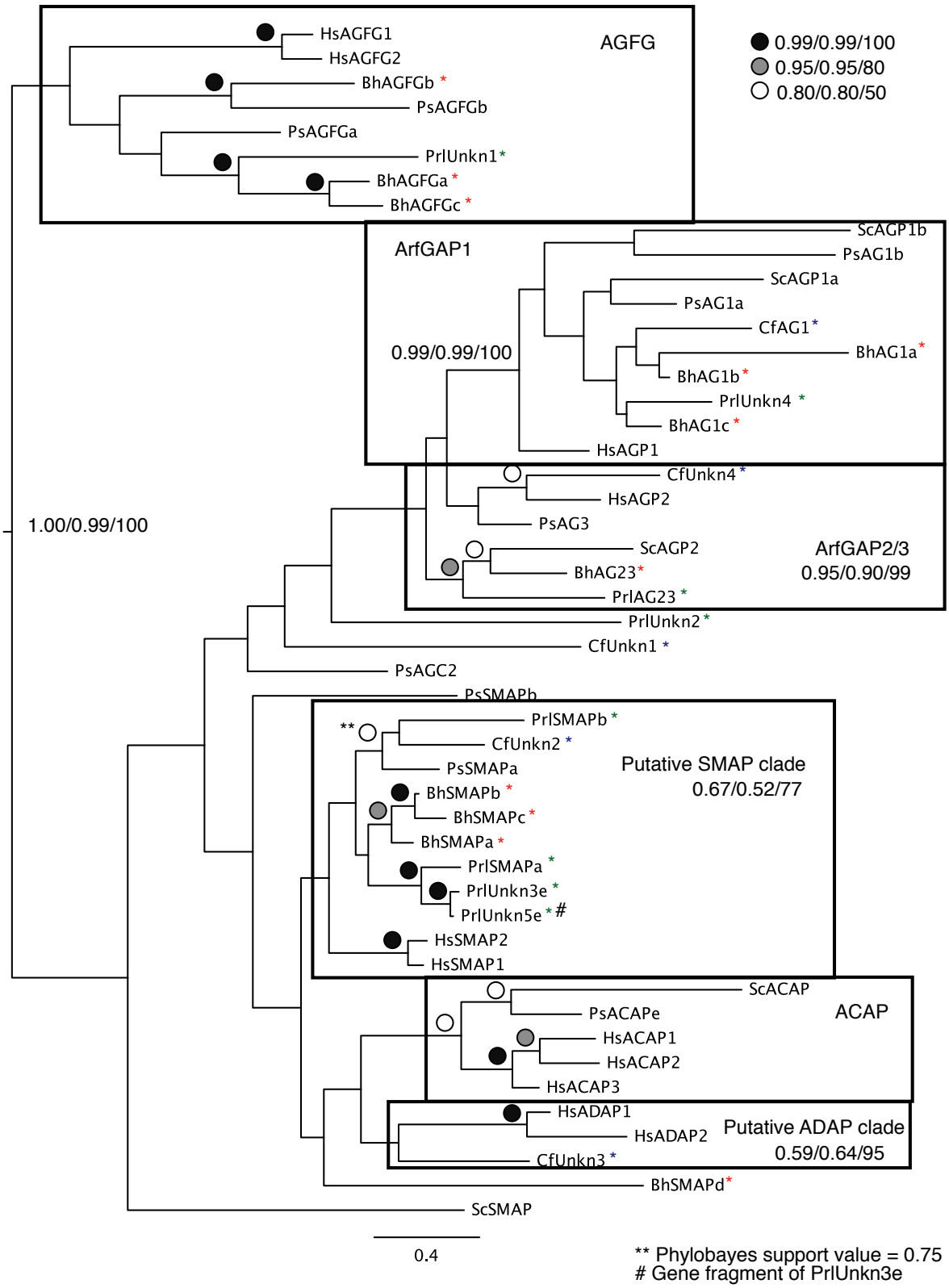
Supplementary Figure S3.12



Supplementary Figure S3.13



Supplementary Figure S3.14

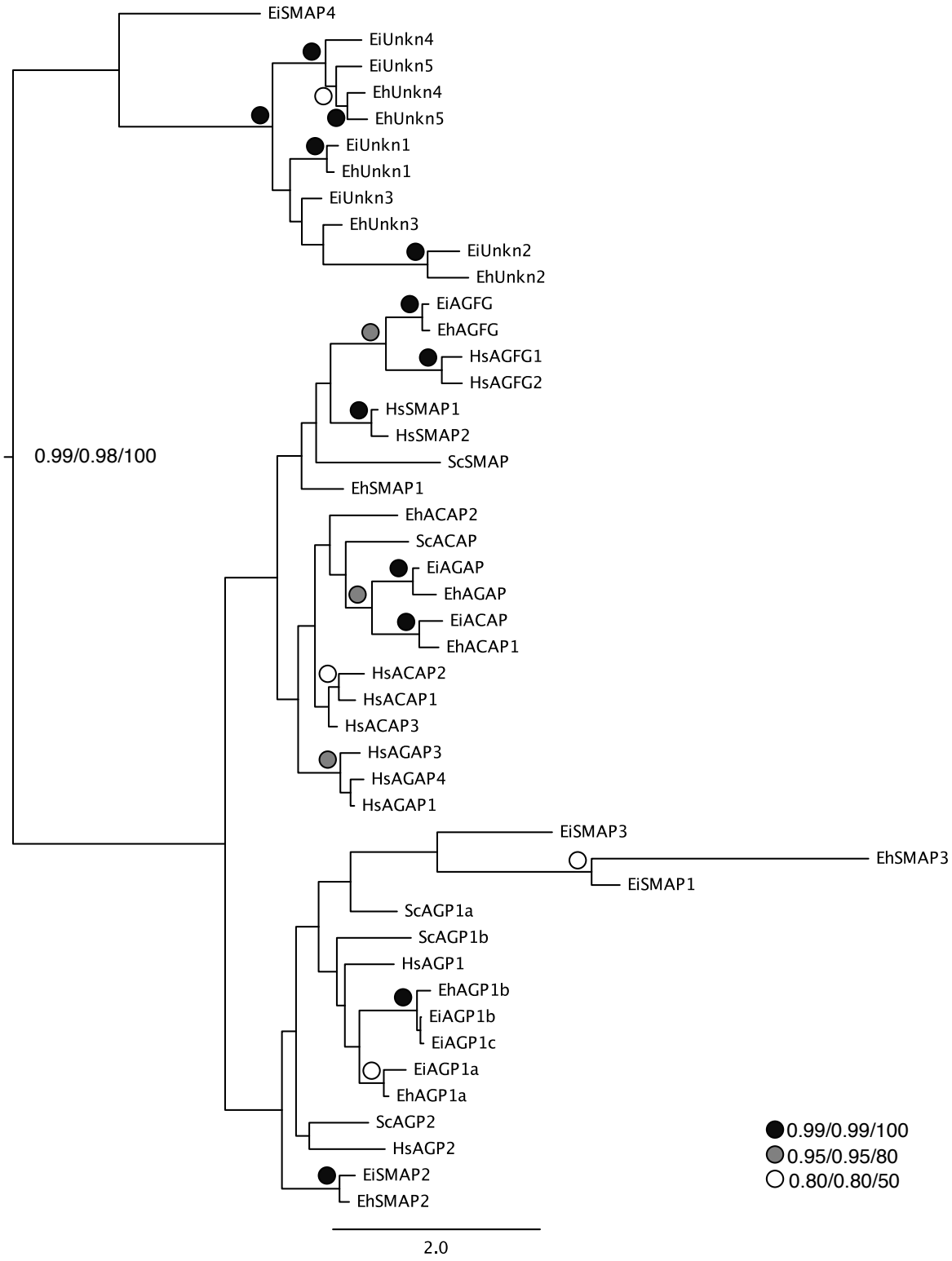


Supplementary Figure S3.15

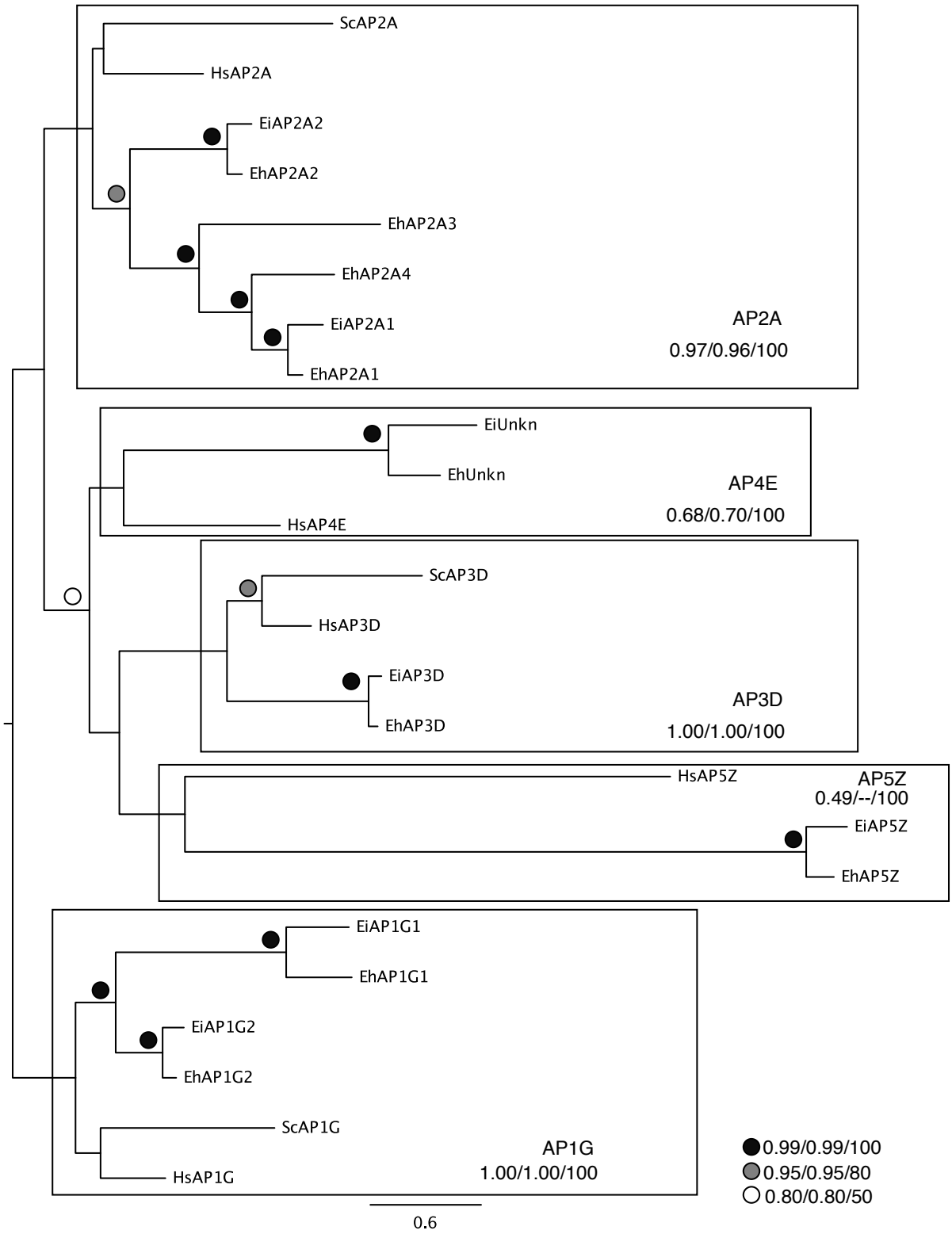
Appendix 3: Chapter 4 Supplementary Material

Supplementary Figures S4.1-S4.4. **Phylogenetic analyses used to classify paralogous families of membrane trafficking machinery in *Entamoeba*.** Node values indicating statistical support are listed as MrBAYES/Phylobayes/RAxML (posterior probability/posterior probability/bootstrap), or as circles indicating the minimum level of support for that node. Nodes without values do not have significant support in one or more phylogeny. The best Bayesian topology (MrBAYES) is shown with node values from all three methods. Tree metadata are found in Online Appendix 4.4.

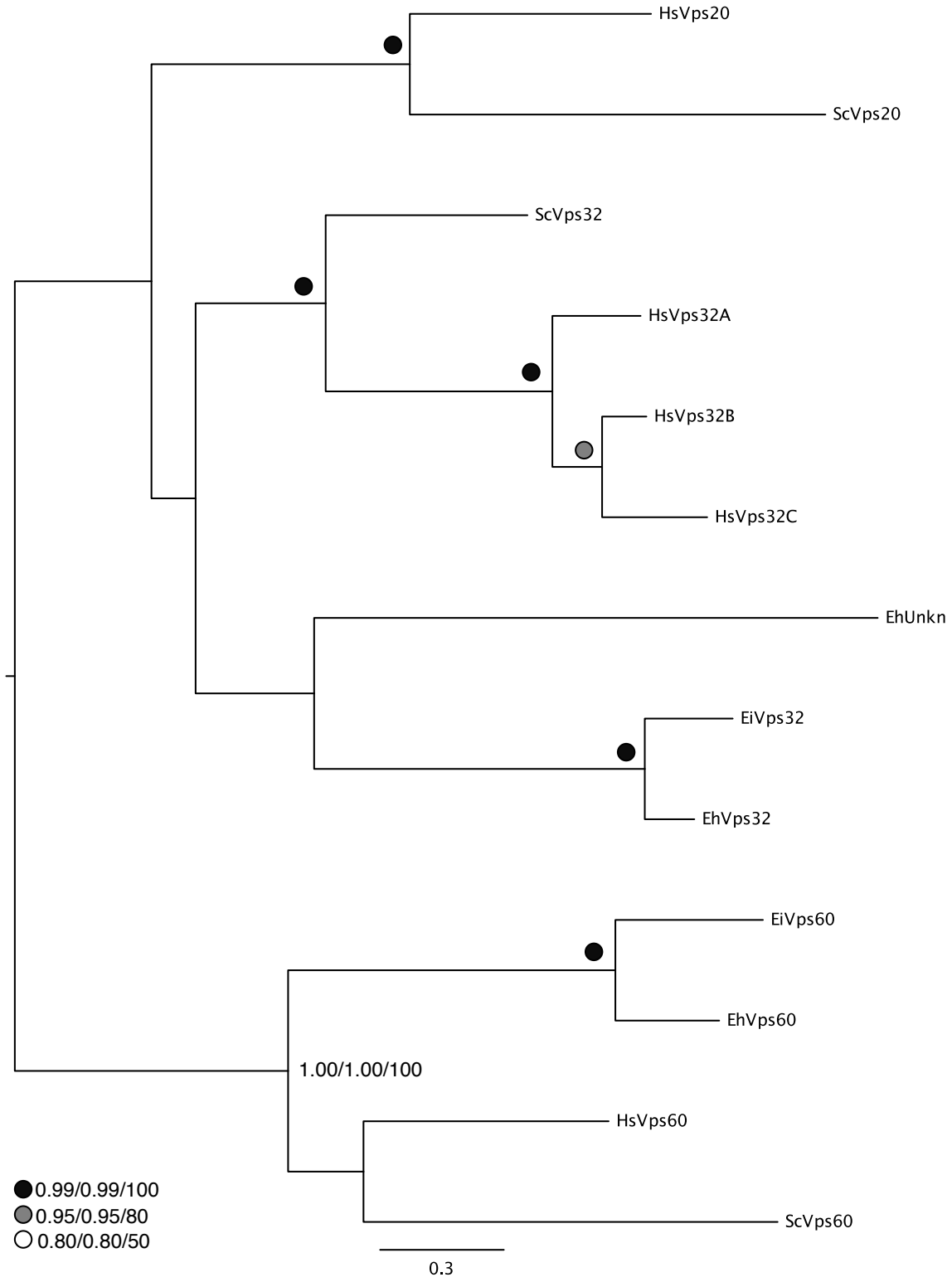
Consensus phylogenies were used to classify the following *Entamoeba* sp. sequences: S4.1, ArfGAPs; S4.2, adaptin γ/α/δ/ε/ζ subunits; S4.3, Vps60, Vps20, Vps32; S4.4, Qc SNAREs.



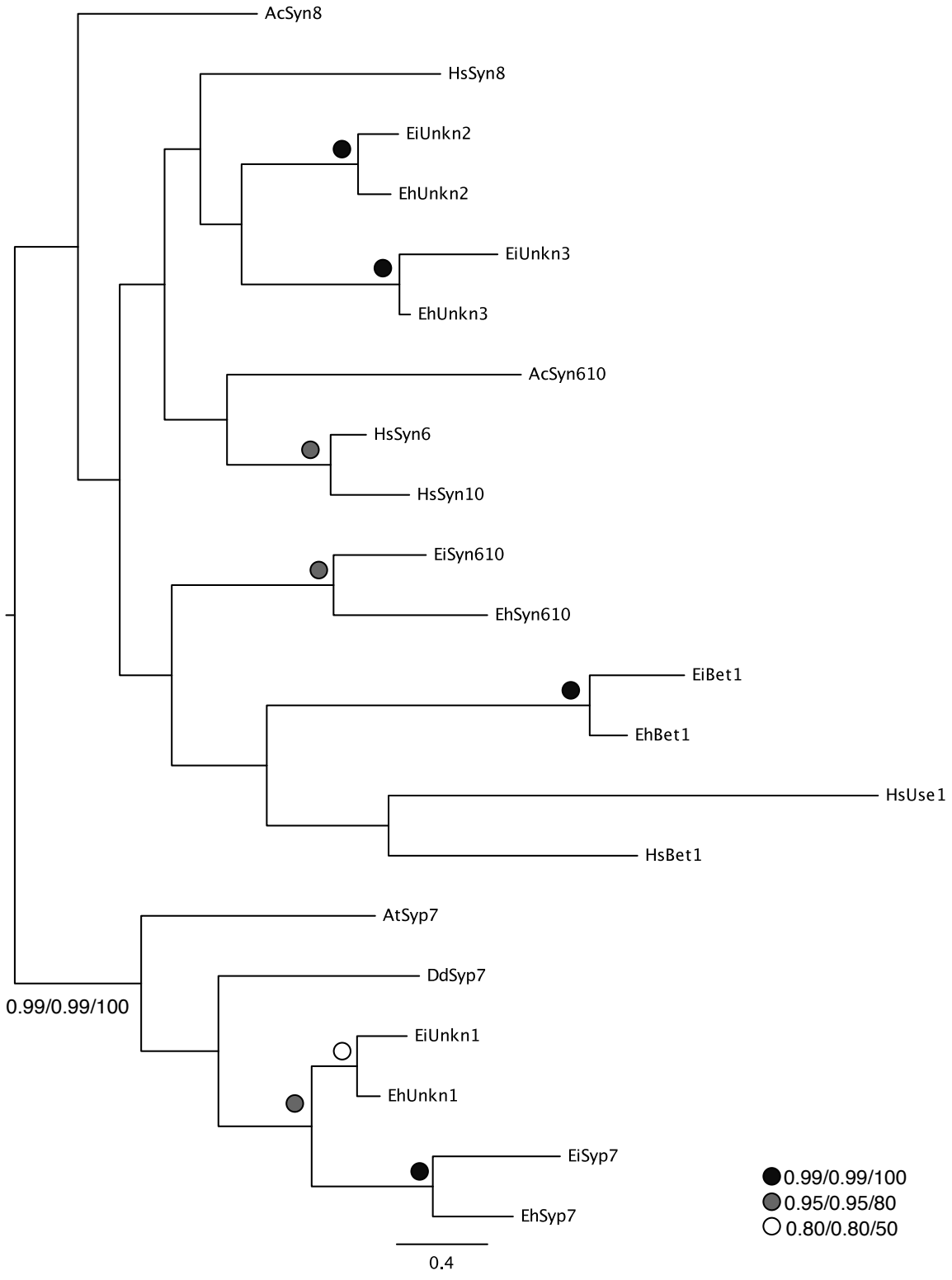
Supplementary Figure S4.1



Supplementary Figure S4.2



Supplementary Figure S4.3

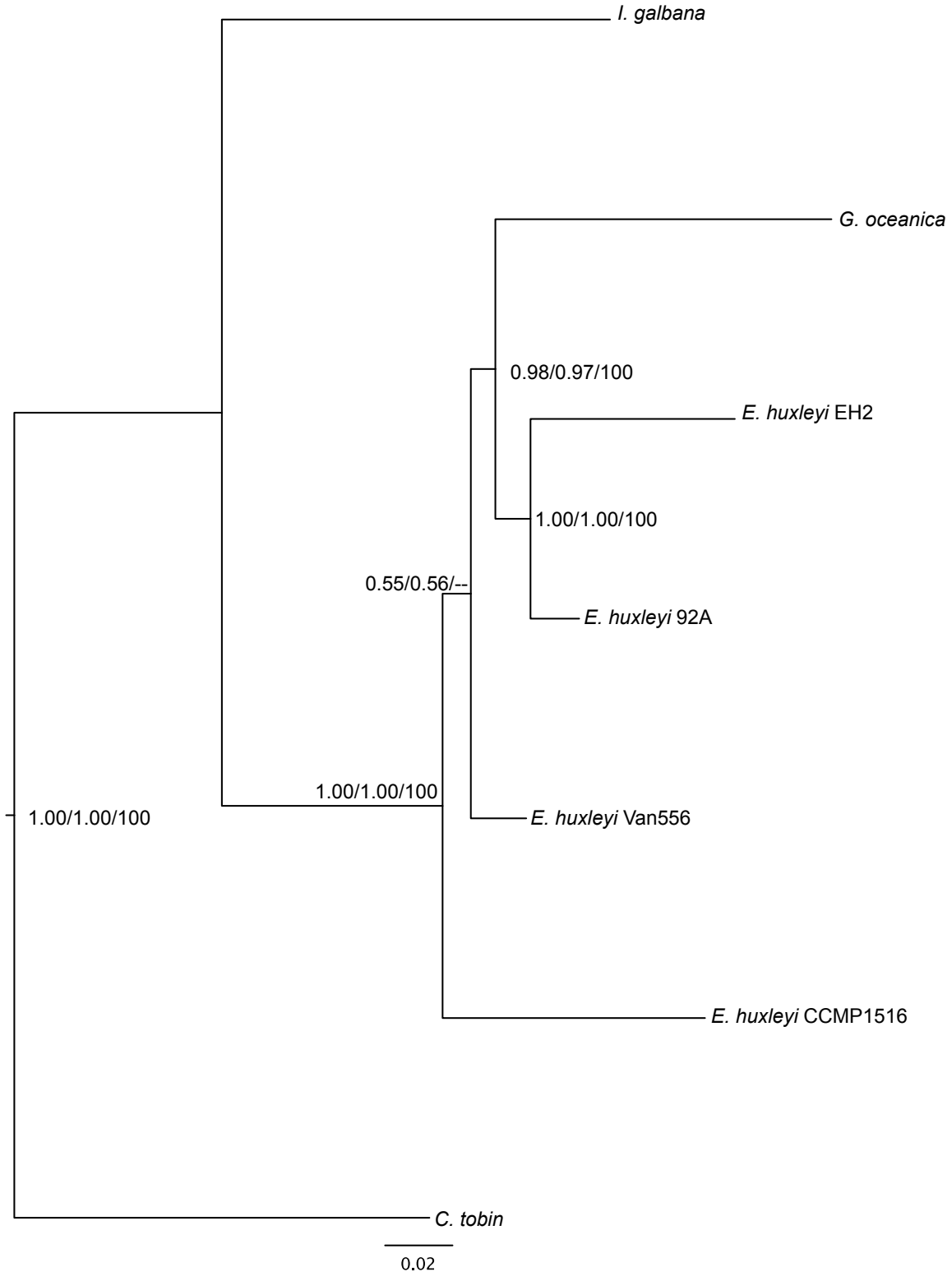


Supplementary Figure S4.4

Online Appendix Figures 4.1-4.10. Phylogenetic analysis of Arf and Rab GTPase proteins and their GAP and GEF regulators. This figure compilation contains MrBAYES, Phylobayes, and RAxML trees used to determine orthology between *E. invadens* and *E. histolytica* sequences. All node values are given for each tree: MrBAYES and Phylobayes values are posterior probabilities and RAxML values are bootstrap values. Tree metadata are found in Online Appendix Table 4.4. Online Appendix Figures 4.1-4.2 were used to classify *Entamoeba* sp. TBC RabGAP sequences, 4.3-4.5 were used to classify ArfGEF sequences, 4.6-4.8 were used to classify DENN domain-containing RabGEF sequences, and 4.9-4.10 were used to classify ArfGAP sequences.

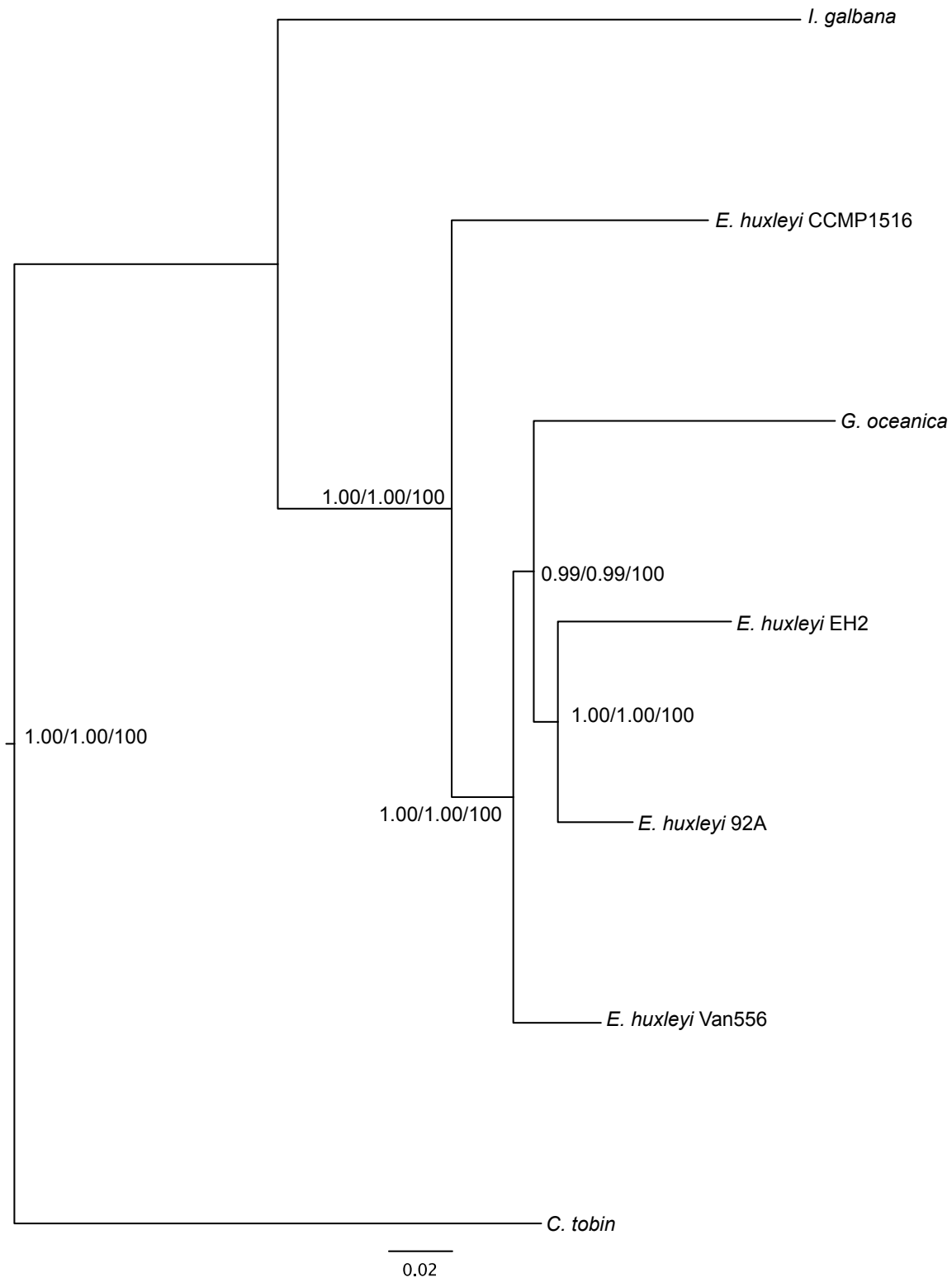
Appendix 4: Chapter 5 Supplementary Material

Supplementary Figure S5.1. Concatenated gene phylogeny of six genes to infer haptophyte evolution. This phylogeny is based on the dataset used to generate the tree in Figure 5.1, with three genes removed based on their individual gene trees. Node values indicating statistical support are listed as MrBAYES/Phylobayes/RAxML (posterior probability/posterior probability/bootstrap), and the tree is rooted on the *C. tobin* sequence, as it is the known outgroup to the Isochrysidales. Node values are shown on the best Bayesian topology. *G. oceanica* groups within a larger clade of *E. huxleyi* strains, although the internal topology of this clade is not well-supported.



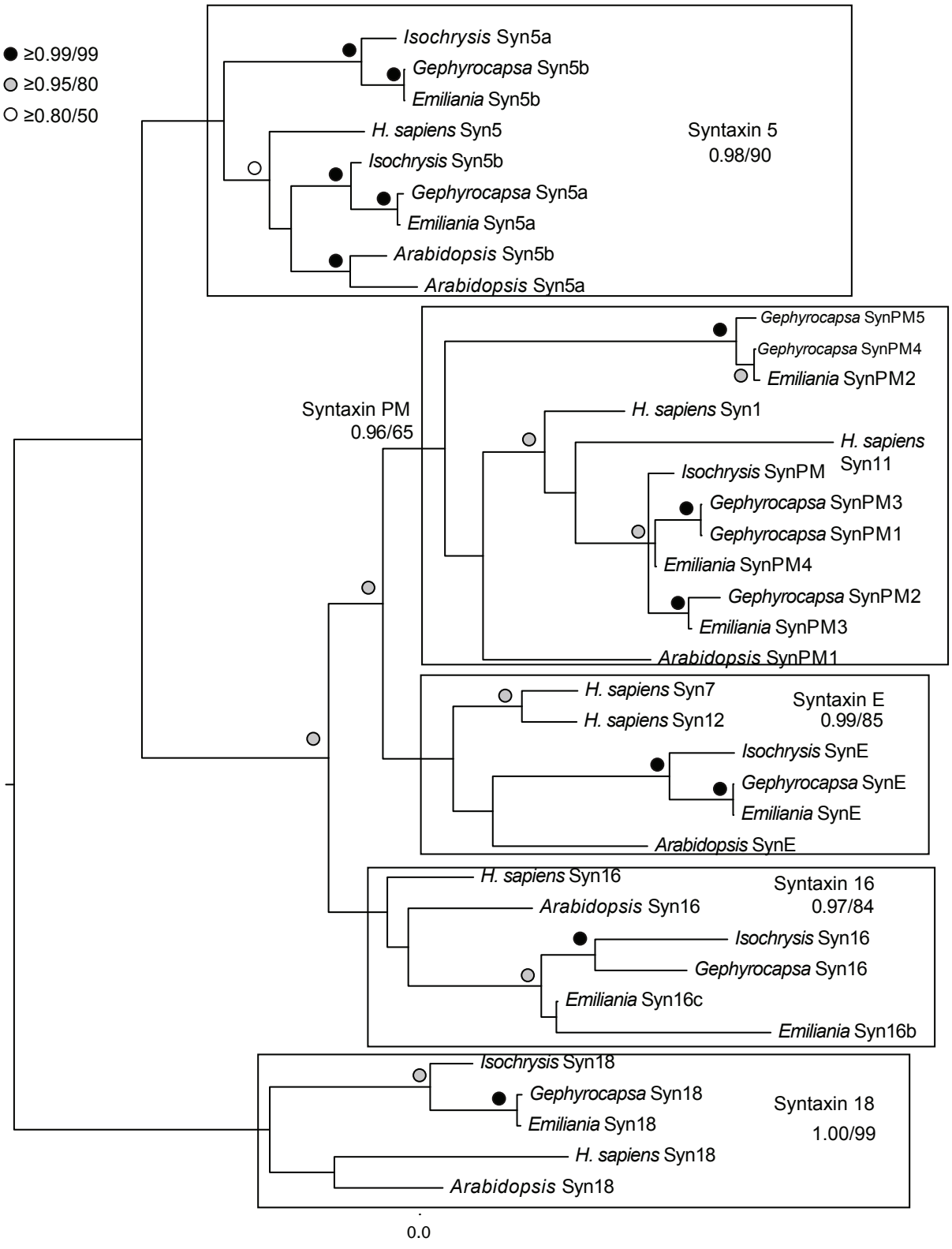
Supplementary Figure S5.1

Supplementary Figure S5.2. Concatenated gene phylogeny of nine genes to infer haptophyte evolution. This phylogeny is based on the dataset used to generate the tree in Figure 5.1, with two genes removed that do not have orthologues in *C. tobin*. Node values indicating statistical support are listed as MrBAYES/Phylobayes/RAxML (posterior probability/posterior probability/bootstrap), and the tree is rooted on the *C. tobin* sequence, as it is the known outgroup to the Isochrysidales. Node values are shown on the best Bayesian topology. *G. oceanica* groups within a larger clade of *E. huxleyi* strains, although the internal topology of this clade is not well-supported.



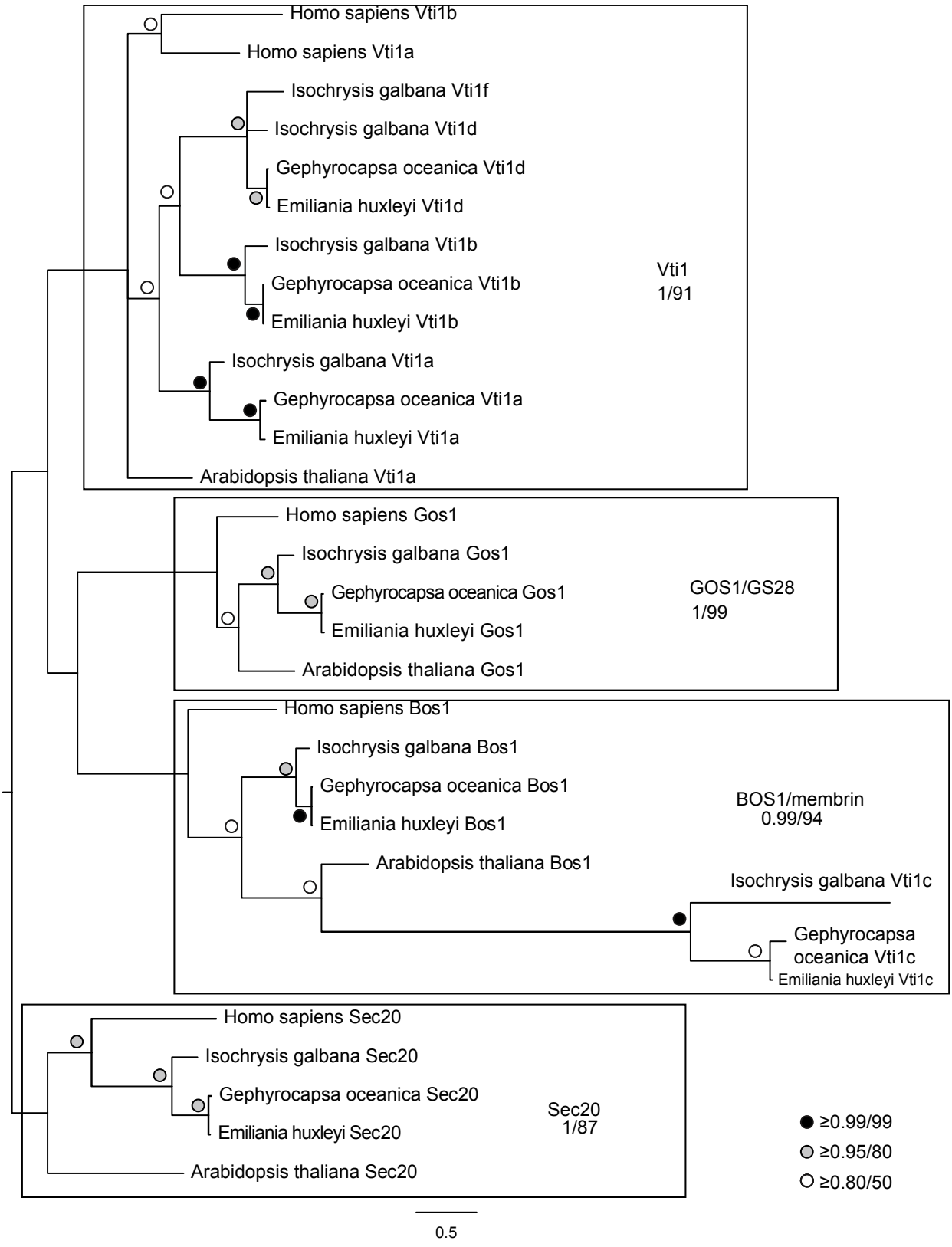
Supplementary Figure S5.2

Supplementary Figure S5.3. Phylogenetic classification of Qa SNAREs in the haptophytes. *E. huxleyi* CCMP1516, *G. oceanica*, and *I. galbana* Qa SNAREs are included with previously characterized sequences from *H. sapiens* and *A. thaliana*. *C. tobin* sequences were classified based on similarity to the sequences classified here. Node values are listed as Phylobayes/RAxML (posterior probability/bootstrap) for critical nodes, while symbols indicate nodes with a minimum level of support shown in the inset. Node values are shown on the best Bayesian topology. Important clades are boxed.



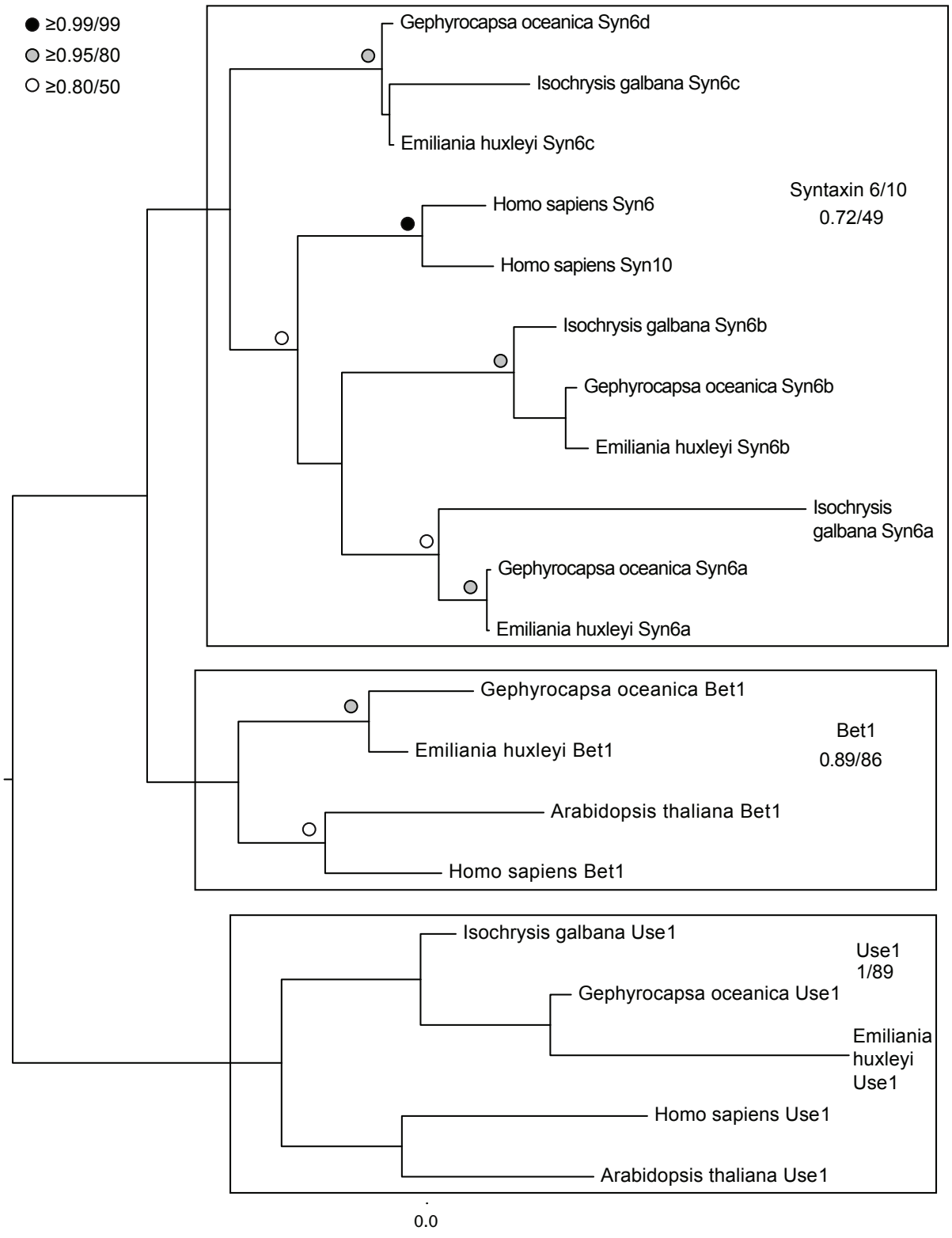
Supplementary Figure S5.3

Supplementary Figure S5.4. Phylogenetic classification of Qb SNAREs in the haptophytes. *E. huxleyi* CCMP1516, *G. oceanica*, and *I. galbana* Qb SNAREs are included with previously characterized sequences from *H. sapiens* and *A. thaliana*. *C. tobin* sequences were classified based on similarity to the sequences classified here. Node values are listed as Phylobayes/RAxML (posterior probability/bootstrap) for critical nodes, while symbols indicate nodes with a minimum level of support shown in the inset. Node values are shown on the best Bayesian topology. Important clades are boxed.



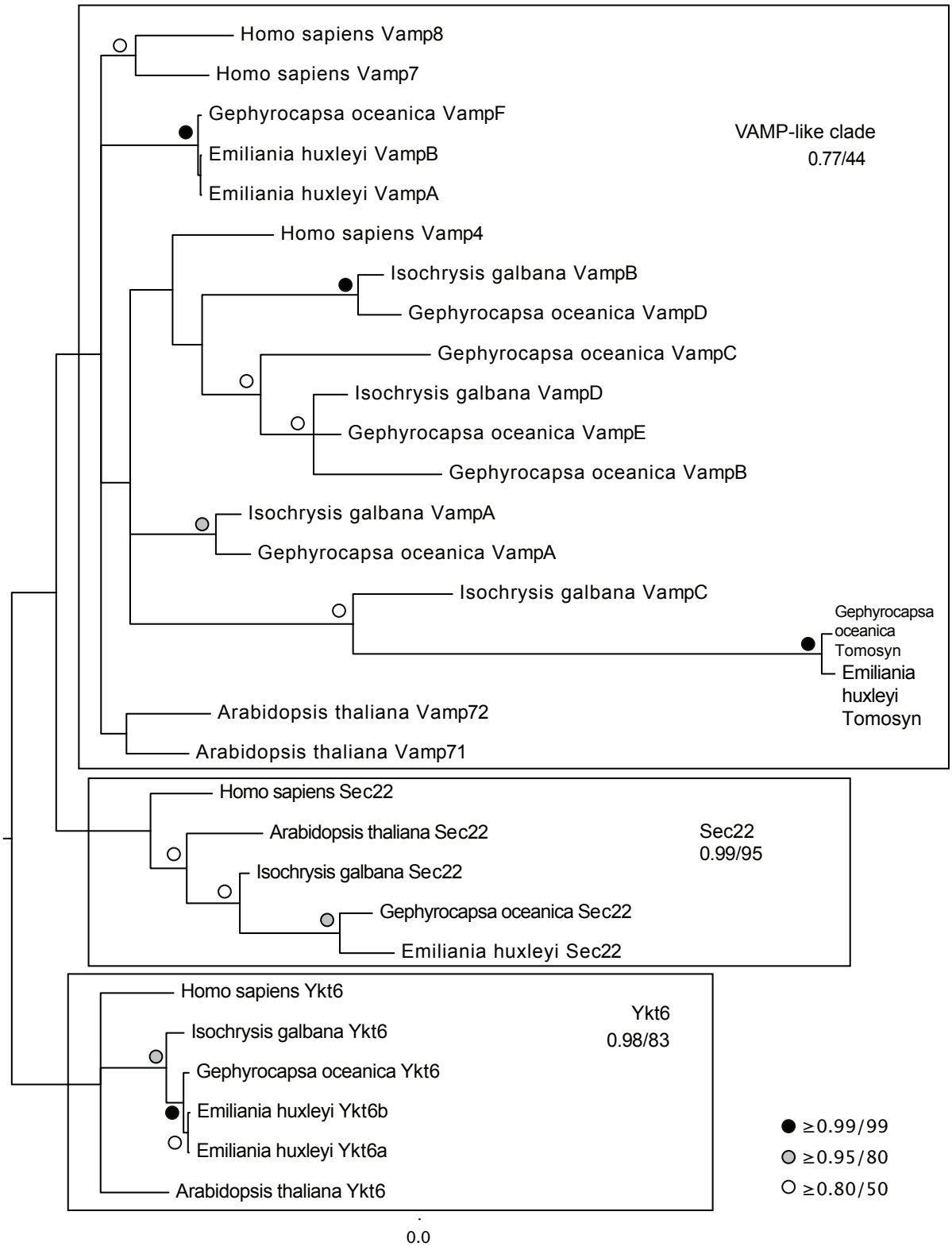
Supplementary Figure S5.4

Supplementary Figure S5.5. Phylogenetic classification of Qc SNAREs in the haptophytes. *E. huxleyi* CCMP1516, *G. oceanica*, and *I. galbana* Qc SNAREs are included with previously characterized sequences from *H. sapiens* and *A. thaliana*. *C. tobin* sequences were classified based on similarity to the sequences classified here. Node values are listed as Phylobayes/RAxML (posterior probability/bootstrap) for critical nodes, while symbols indicate nodes with a minimum level of support shown in the inset. Node values are shown on the best Bayesian topology. Important clades are boxed.



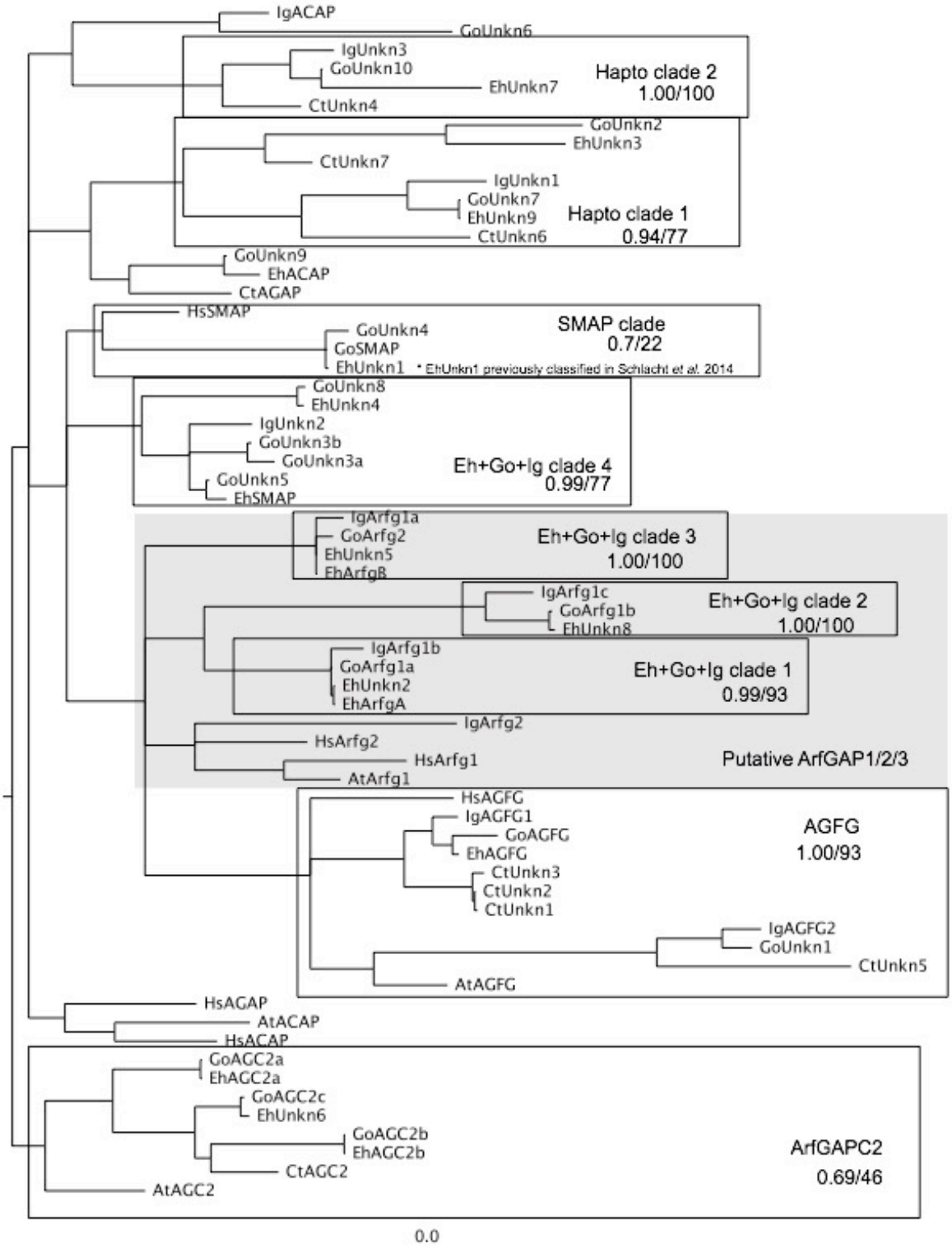
Supplementary Figure S5.5

Supplementary Figure S5.6. Phylogenetic classification of R SNAREs in the haptophytes. *E. huxleyi* CCMP1516, *G. oceanica*, and *I. galbana* R SNAREs are included with previously characterized sequences from *H. sapiens* and *A. thaliana*. *C. tobin* sequences were classified based on similarity to the sequences classified here. Node values are listed as Phylobayes/RAxML (posterior probability/bootstrap) for critical nodes, while symbols indicate nodes with a minimum level of support shown in the inset. Node values are shown on the best Bayesian topology. Important clades are boxed.



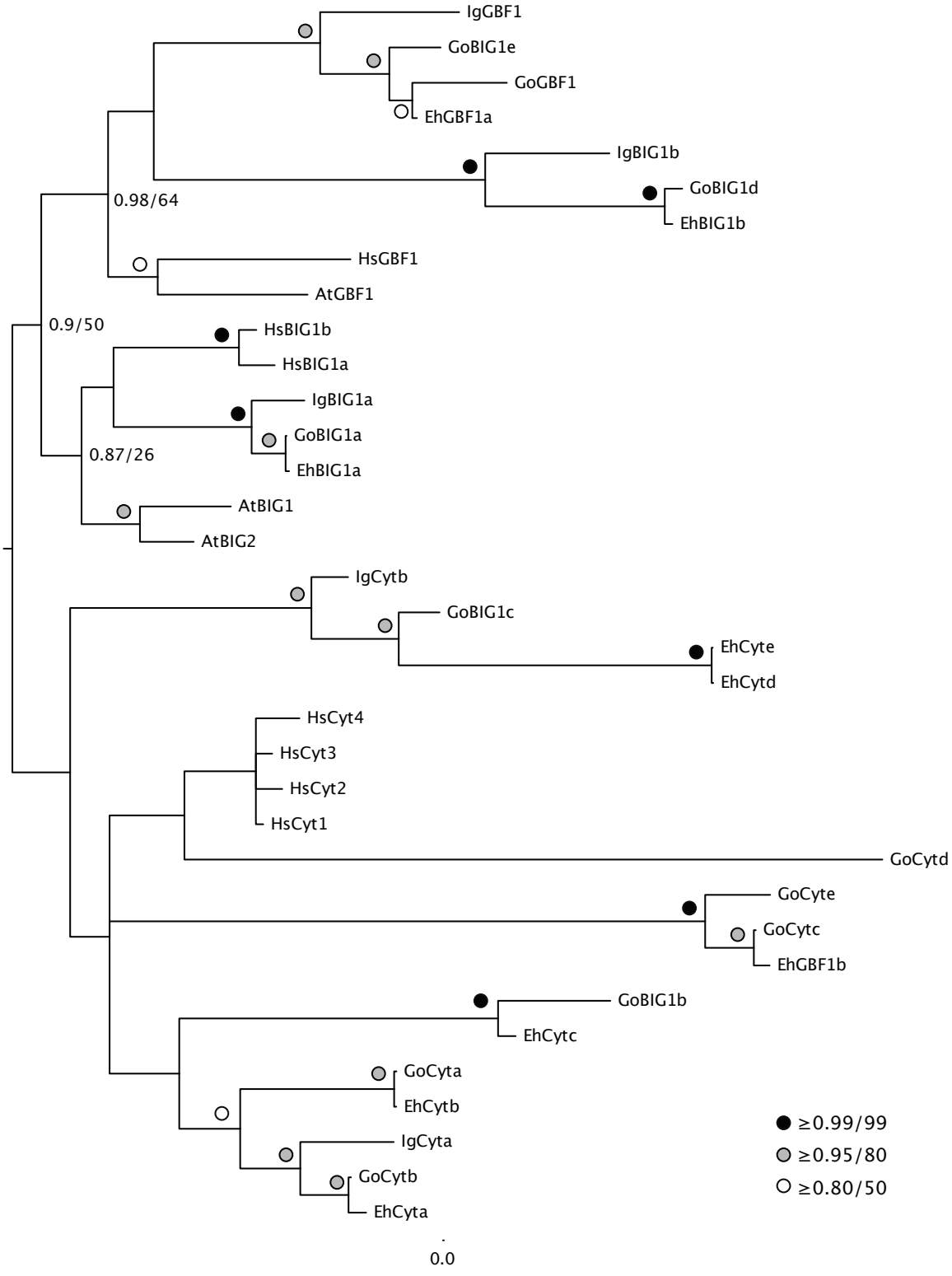
Supplementary Figure S5.6

Supplementary Figure S5.7. Phylogenetic classification of ArfGAPs in the haptophytes. *E. huxleyi* CCMP1516, *G. oceanica*, *I. galbana*, and *C. tobin* sequences are included with previously characterized sequences from *H. sapiens* and *A. thaliana*. Node values are listed as Phylobayes/RAxML (posterior probability/bootstrap) for critical nodes only. Node values are shown on the best Bayesian topology. Clades are boxed. A clade containing ArfGAP1 and ArfGAP2/3 sequences is shaded. One sequence (EhUnkn1) was shown to be an SMAP homologue in the Schlacht *et al.* (2013)[199] classification of eukaryotic ArfGAPs, and therefore is used to label this clade.



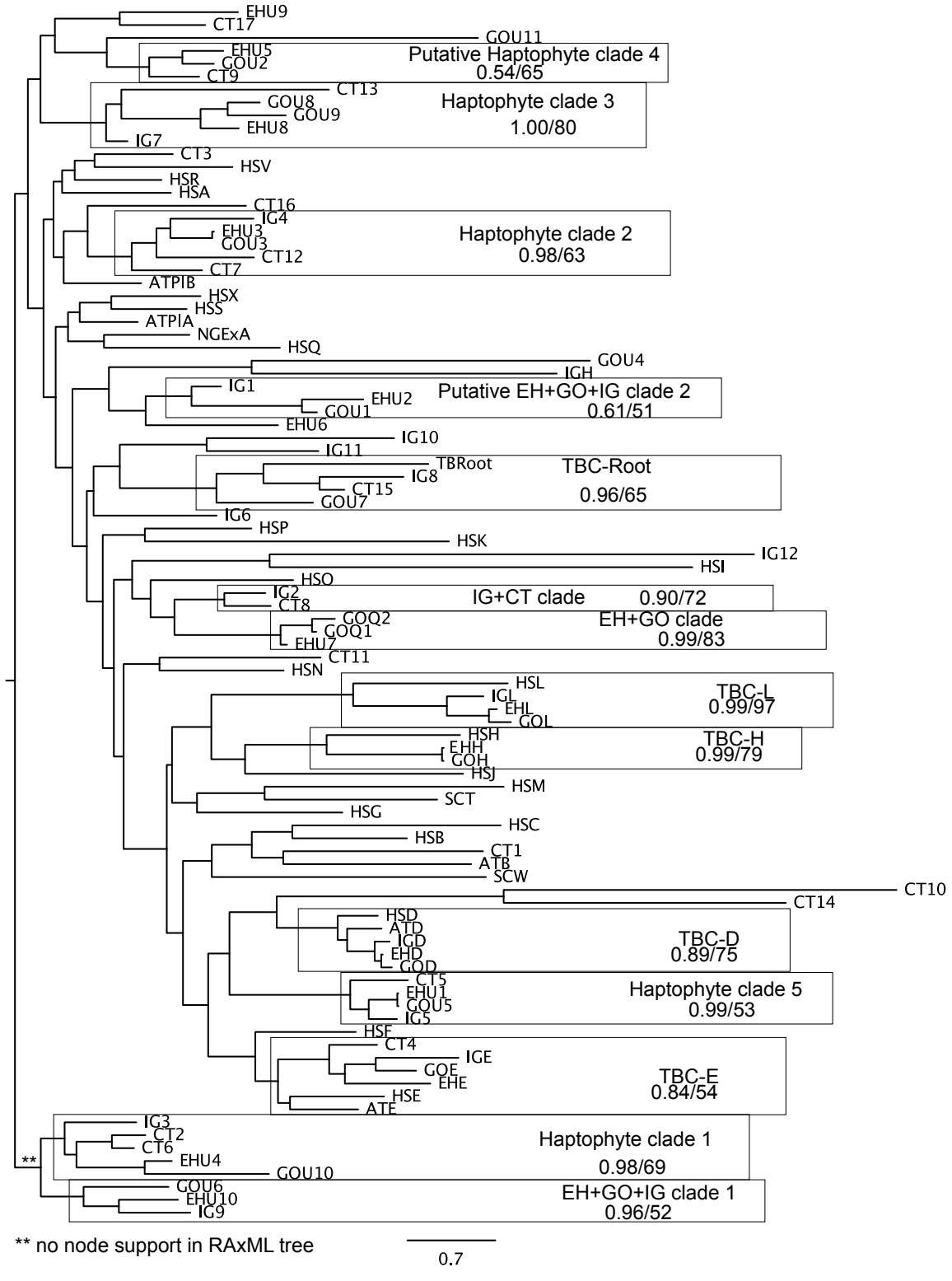
Supplementary Figure S5.7

Supplementary Figure S5.8. Phylogenetic classification of ArfGEFs in the haptophytes. *E. huxleyi* CCMP1516, *G. oceanica*, *I. galbana*, and *C. tobin* sequences are included with previously characterized sequences from *H. sapiens* and *A. thaliana*. Node values are listed as Phylobayes/RAxML (posterior probability/bootstrap) for critical nodes, while symbols indicate nodes with a minimum level of support shown in the inset. Node values are shown on the best Bayesian topology. Two large clades are apparent; one with BIG/GBF-like sequences and one with cytohesin-like sequences.



Supplementary Figure S5.8

Supplementary Figure S5.9. Phylogenetic classification of TBC RabGAPs in the haptophytes. *E. huxleyi* CCMP1516, *G. oceanica*, *I. galbana*, and *C. tobin* sequences are included with previously characterized sequences from *H. sapiens*, *A. thaliana, N. gruberi, T. brucei,* and *S. cerevisiae*. Node values are listed as MrBAYES/RAxML (posterior probability/bootstrap) only for critical nodes for readability. Node values are shown on the best Bayesian topology.

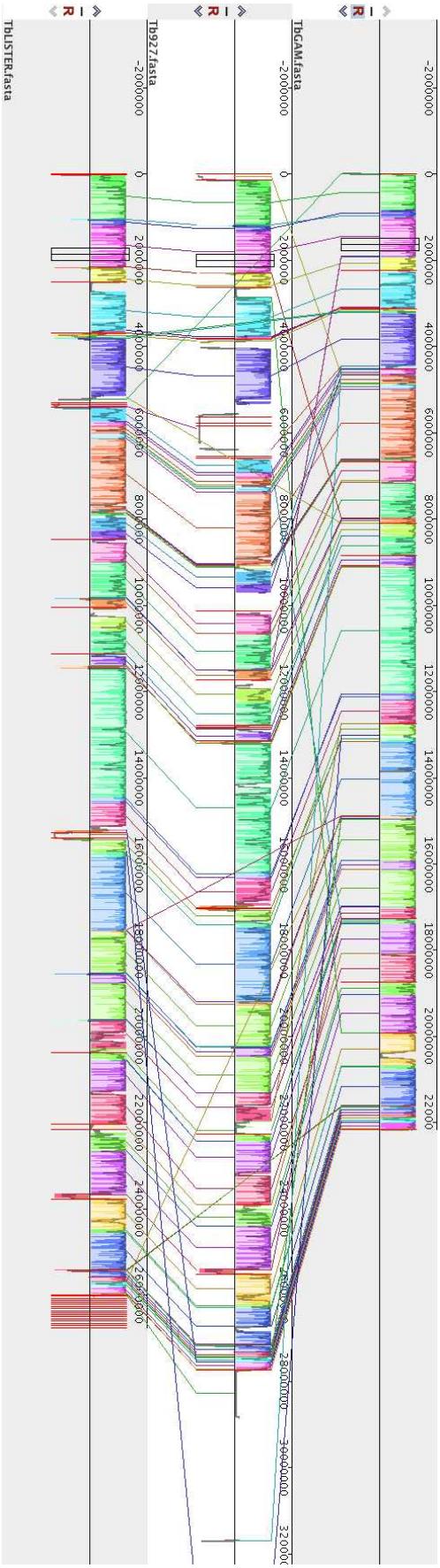


Supplementary Figure S5.9

Appendix 5: Chapter 6 Supplementary Material

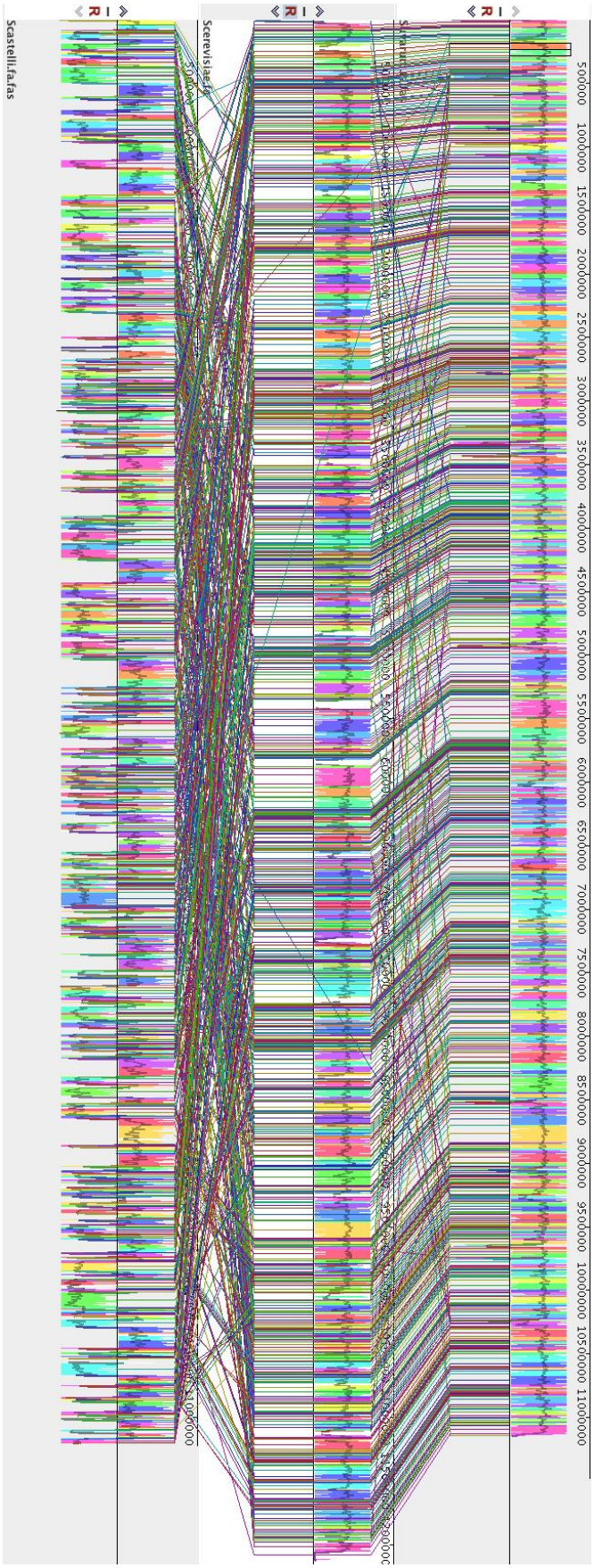
Supplementary Figure S6.1. Genomic alignment of three strains of *Trypanosoma brucei*. Mauve output showing synteny between *T. brucei gambiense* (top), *T. brucei* 927 (middle), and *T. brucei* Lister (bottom). Coloured blocks indicate regions of contiguous sequence on each strand (upper versus lower blocks). Lines connecting blocks between each genome show the extent of sequence rearrangement in the three strains. Red vertical lines delineate contigs or scaffolds.

Supplementary Figure S6.1



Supplementary Figure S6.2. Genomic alignment of three species of *Saccharomyces*. Mauve output showing synteny between *S. uvarum* (top), *S. cerevisiae* (middle), and *S. castelli* (bottom). Coloured blocks indicate regions of contiguous sequence on each strand (upper versus lower blocks). Lines connecting blocks between each genome show the extent of sequence rearrangement in the three strains. Red vertical lines delineate contigs or scaffolds.

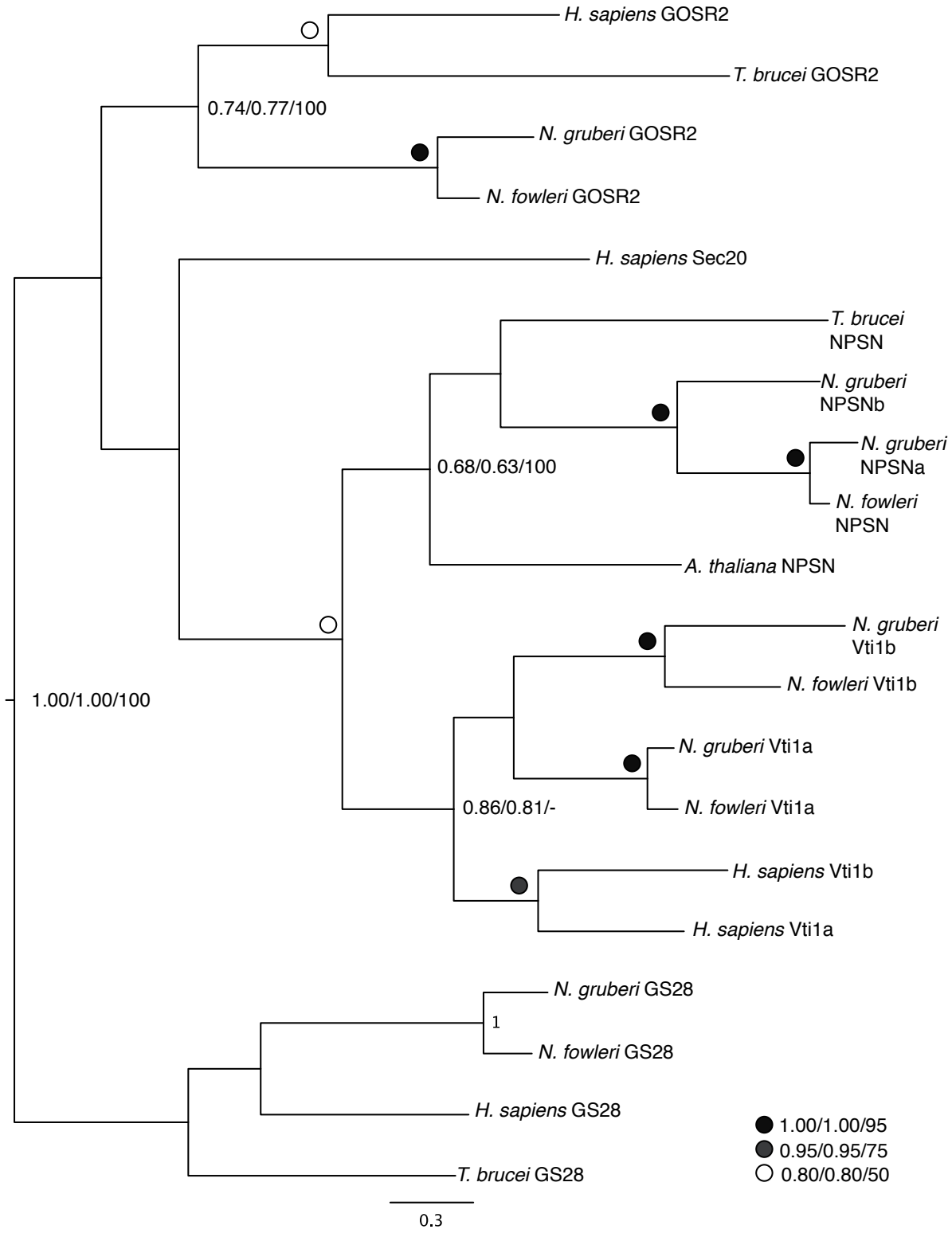
Supplementary Figure S6.2



Supplementary Figure S6.3. Consensus phylogeny of Qa SNAREs in *Naegleria* sp. *N. fowleri* V212 and *N. gruberi* sequences are classified using previously characterized sequences from *H. sapiens* and *T. brucei*. Node values are listed as MrBAYES/Phylobayes/RAxML (posterior probability/posterior probability/bootstrap) for critical nodes, and as shown in the inset.

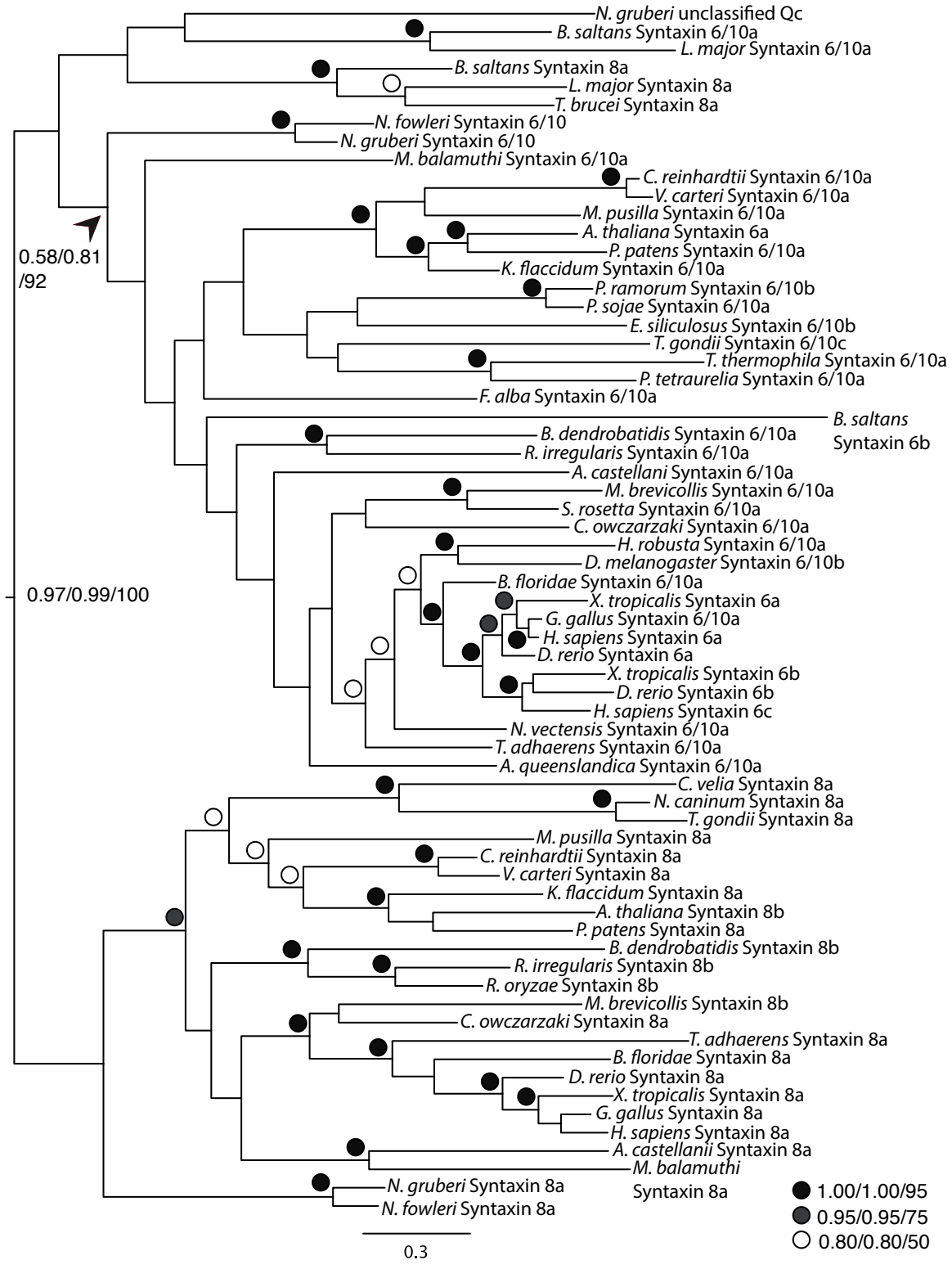
Supplementary Figure S6.3

Supplementary Figure S6.4. Consensus phylogeny of Qb SNAREs in *Naegleria* sp. *N. fowleri* V212 and *N. gruberi* sequences are classified using previously characterized sequences from *H. sapiens, A. thaliana* and *T. brucei*. Node values are listed as MrBAYES/Phylobayes/RAxML (posterior probability/posterior probability/bootstrap) for critical nodes, and as shown in the inset.



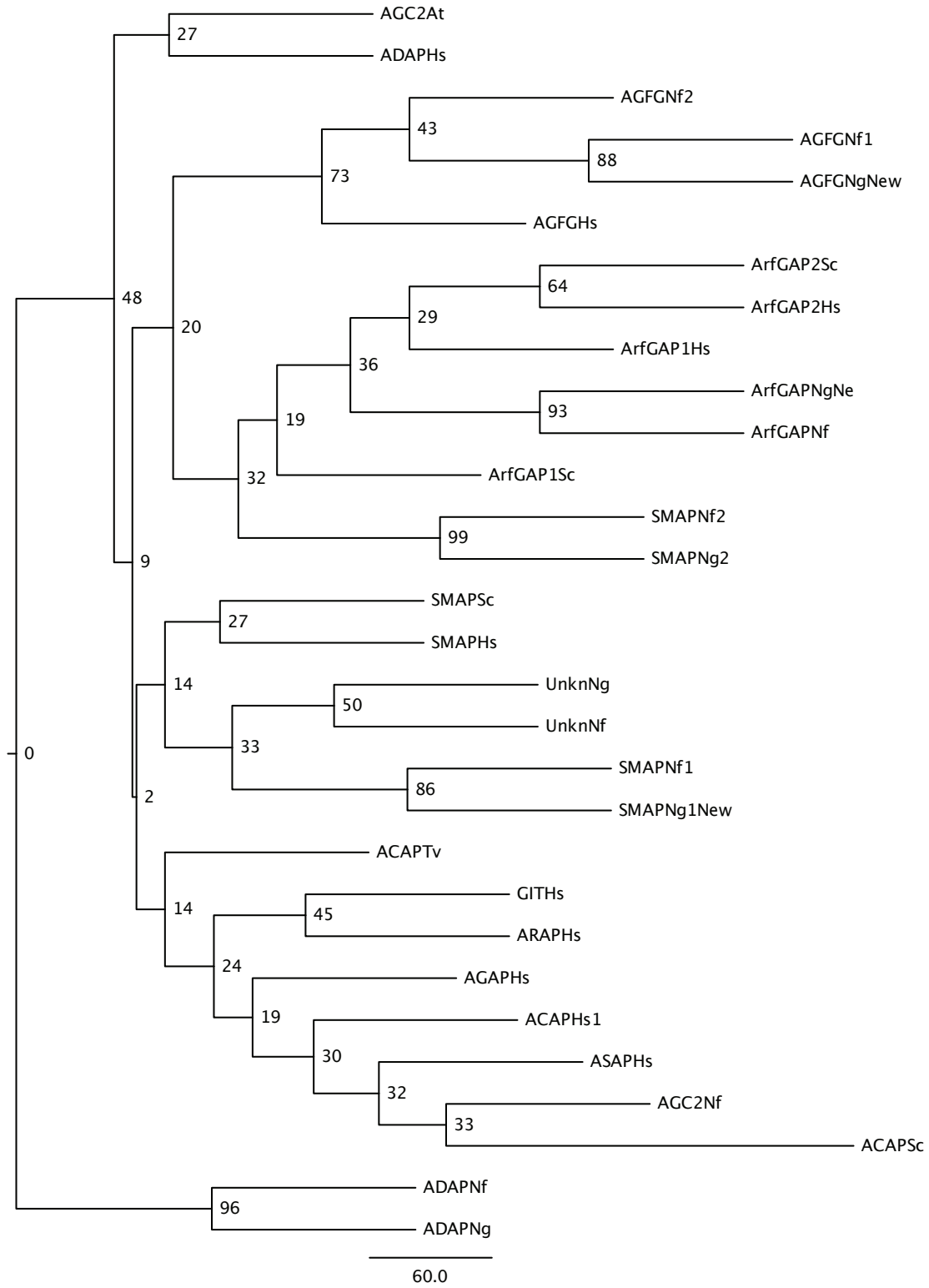
Supplementary Figure S6.4

Supplementary Figure S6.5. Consensus phylogeny of Qc SNAREs in *Naegleria* sp. *N. fowleri* V212 and *N. gruberi* sequences are classified with other eukaryotic Qc SNAREs from a variety of taxa. Node values are listed as MrBAYES/Phylobayes/RAxML (posterior probability/posterior probability/bootstrap) for critical nodes, and as shown in the inset.

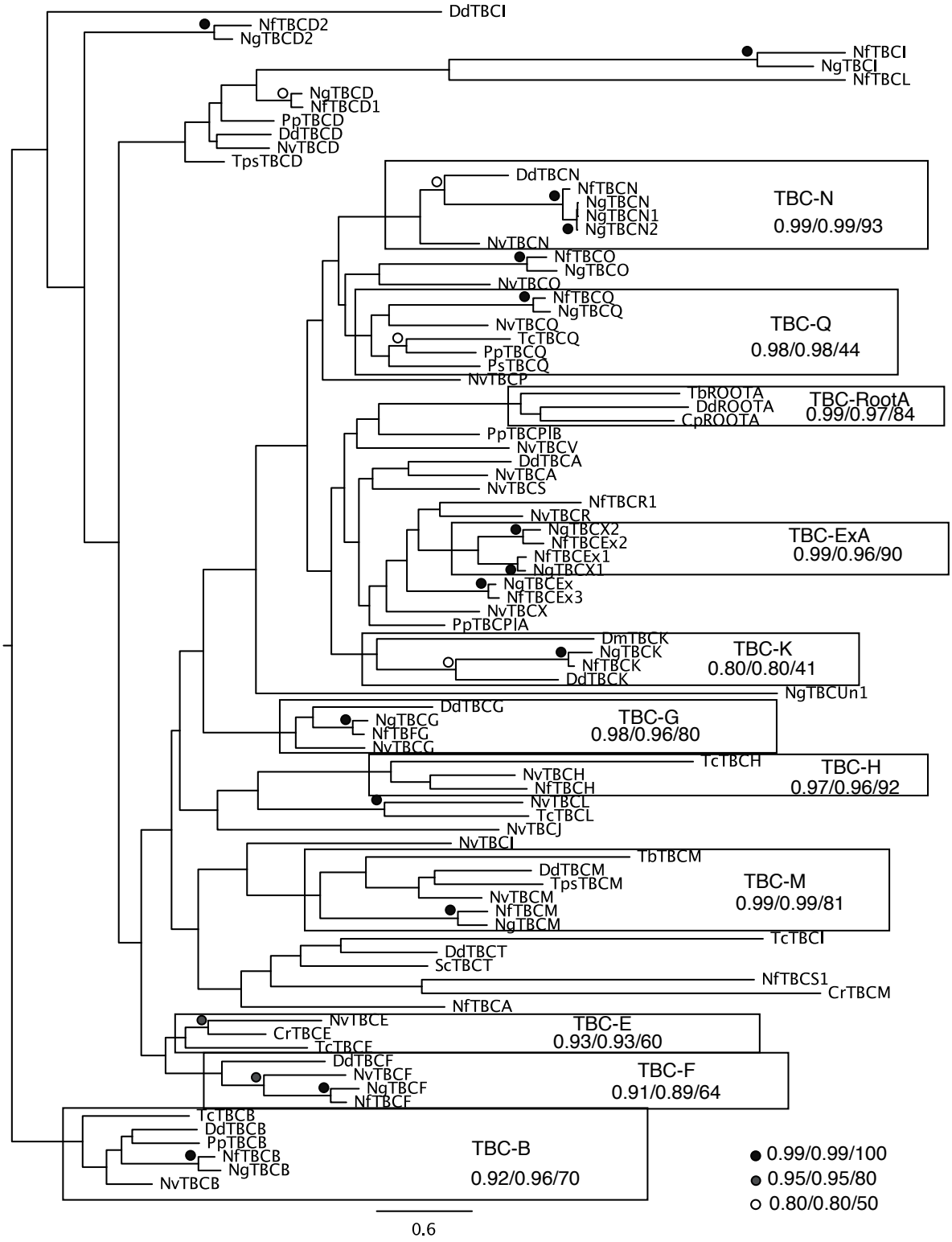


Supplementary Figure S6.5

Supplementary Figure S6.6. RAxML phylogeny of ArfGAPs in *Naegleria* sp. *N. fowleri* V212 and *N. gruberi* sequences are classified using previously characterized sequences from *H. sapiens, S. cerevisiae*, and *A. thaliana* sequences. Bootstrap values are given for all nodes.


AGC2At
27
ADAPHs
AGFGNf2
43
AGFGNf1
88
AGFGNgNew
73
AGFGHs
ArfGAP2Sc
64
ArfGAP2Hs
48
29
20
ArfGAP1Hs
36
ArfGAPNgNe
93
19
ArfGAPNf
32
ArfGAP1Sc
SMAPNf2
99
SMAPNg2
9
SMAPSc
27
SMAPHs
14
UnknNg
50
UnknNf
0
33
SMAPNf1
86
2
SMAPNg1New
ACAPTv
GlTHs
14
45
ARAPHs
24
AGAPHs
19
ACAPHs1
30
ASAPHs
32
AGC2Nf
33
ACAPSc
ADAPNf
96
ADAPNg
60.0

Supplementary Figure S6.7. Consensus phylogeny of TBC RabGAPs in *Naegleria* sp. *N. fowleri* V212 and *N. gruberi* TBC sequences are classified using previously characterized sequences from Gabernet-Castello *et al.* (2013).[330] Node values are listed as MrBAYES/Phylobayes/RAxML (posterior probability/posterior probability/bootstrap) for critical nodes, and as shown in the inset, and important clades are boxed.



Supplementary Figure S6.7

Supplementary Table ST6.1. *N. fowleri* mitochondrial genome annotation based on BLAST searching and RNA scanning software.

Annotation	Gene name	*N. gruberi* homologue[1]	Start (*fowleri*/*gruberi*)	End (*fowleri*/*gruberi*)	*N. fowleri* length (bp)	*N. gruberi* length (bp)	*Nf*/*Ng* % length	E-value[2]	Identities (%)[2]	Cove	Score (rRNA)
Rnl	Large subunit rRNA	NC_002573	1/1	2694/2573	2694	2573	121	N/A	85	N/A	741.8
Asn tRNA	Asparagine tRNA	NC_002573	2710/2698	2781/2769	72	71	101	N/A	100	74.86	N/A
Nad1	NADH dehydrogenase subunit 1	NP_066498	2810/2802	3787/3785	978	984	99	0	85	N/A	N/A
Rps7	Ribosomal protein S7	NP_066499	3788/3796	4942/5019	1155	1224	94	1.00E-84	47	N/A	N/A
Phe tRNA	Phenylalanine tRNA	NC_002573	5093/5116	5163/5186	71	71	100	N/A	99	56.34	N/A
ATP1	ATP synthase F1 subunit alpha	NP_066500	5250/5259	6902/6911	1653	1653	100	0.00E+00	94	N/A	N/A
Orf312	Orf312	NP_066501	6963/6973	7697/7911	735	939	78	2.00E-08	60	N/A	N/A
His tRNA	Histidine tRNA	NC_002573	7713/7936	7783/8006	71	71	100	N/A	90	62.61	N/A
Nad4L	NADH dehydrogenase subunit 4L	NP_066502	7811/8038	8071/8301	261	264	99	6.00E-51	92	N/A	N/A
Met tRNA	Methionine tRNA	NC_002573	8090/8319	8162/8390	73	72	101	N/A	93	67.24	N/A
Phe tRNA	Phenylalanine tRNA	NC_002573	8182/8406	8252/8476	71	71	100	N/A	100	58.04	N/A
Orf164	Orf 164	NP_066503	8414/8637	8899/9131	486	495	98	2.00E-27	37	N/A	N/A
Rps12	Ribosomal protein S12	NP_066504	8902/9135	9288/9521	387	387	100	5.00E-75	88	N/A	N/A
Gln tRNA	Glutamine tRNA	NC_002573	9289/9537	9359/9607	71	71	100	N/A	89	75.47	N/A
Cox1	Cytochrome c oxidase subunit 1	NP_066505	9612/9939	11513/11840	1902	1902	100	0.00E+00	93	N/A	N/A
Cox11	Haem biosynthesis protein	NP_066506	11612/11968	12286/12636	675	669	101	6.00E-74	52	N/A	N/A
Sdh2	Succinate:cytochrome c oxidoreductase	NP_066507	12290/12672	13120/13484	831	813	102	1.00E-146	74	N/A	N/A
Nad11	NADH dehydrogenase subunit 11	NP_066508	13223/13573	15295/15660	2073	2088	99	0	63	N/A	N/A
Cox3	Cytochrome c oxidase subunit 3	NP_066509	15390/15739	16283/16632	894	894	100	0	87	N/A	N/A
Trp tRNA	Tryptophan tRNA	NC_002573	16282/16648	16352/16718	71	71	100	N/A	100	72.62	N/A
Pro tRNA	Proline tRNA	NC_002573	16407/16754	16479/16826	73	73	100	N/A	95	63.59	N/A
Nad8	NADH dehydrogenase subunit 8	NP_066510	16513/16863	16992/17342	480	480	100	3.00E-110	91	N/A	N/A
Nad5	NADH dehydrogenase subunit 5	NP_066511	17004/17351	19007/19351	2004	2001	100	0	75	N/A	N/A
Cob	Apocytochrome b	NP_066512	19074/19432	20525/20925	1452	1494	97	0	93	N/A	N/A
ATP3	ATP synthase F1 subunit gamma	NP_066513	21018/20997	21881/21863	864	867	100	1.00E-171	79	N/A	N/A
Orf145	Orf 145	NP_066514	21947/21927	22405/22364	459	438	105	3.00E-36	69	N/A	N/A
Nad6	NADH dehydrogenase subunit 6	NP_066515	22454/22416	23059/23021	606	606	100	9.00E-98	75	N/A	N/A
Rps2	Ribosomal protein S2	NP_066516	23059/23021	23847/23815	789	795	99	3.00E-119	66	N/A	N/A
ATP8	ATP synthase F0 subunit 8	NP_066517	23923/23888	24255/24220	333	333	100	2.00E-39	78	N/A	N/A
Rps10	Ribosomal protein S10	NP_066518	<24670/24233	24912/24895	243	663	37	2.00E-26	63	N/A	N/A
Rpl11	Ribosomal protein L11	NP_066519	24909/24900	25346/25346	438	447	98	4.00E-55	56	N/A	N/A
Rpl2	Ribosomal protein L2	NP_066520	25386/25370	26186/26176	801	807	99	8.00E-133	69	N/A	N/A
Rps19	Ribosomal protein S19	NP_066521	26188/26176	26451/26442	264	267	99	6.00E-17	43	N/A	N/A
Rps3	Ribosomal protein S3	NP_066522	26457/26443	28004/28014	1548	1572	98	1.00E-150	53	N/A	N/A
Rpl16	Ribosomal protein L16	NP_066523	28006/28014	28431/28439	426	426	100	2.00E-73	75	N/A	N/A

Rpl14	Ribosomal protein L14	NP_066524	28440/28445	28811/28816	372		100	2.00E-65		82	N/A	N/A	
Rpl5	Ribosomal protein L5	NP_066525	28822/28817	29397/29395	576	579	99	3.00E-73		63	N/A	N/A	
Rps14	Ribosomal protein S14	NP_066526	29397/29395	29693/29691	297	297	100	7.00E-31		56	N/A	N/A	
Rps8	Ribosomal protein S8	NP_066527	29698/29704	30108/30129	411	426	96	1.00E-43		50	N/A	N/A	
Rpl6	Ribosomal protein L6	NP_066528	30118/30134	30603/30634	486	501	94	4.00E-54		57	N/A	N/A	
Rps11	Ribosomal protein S11	NP_066529	30604/30636	32274/32417	1671	1782	94	3.00E-150		46	N/A	N/A	
Rps13	Ribosomal protein S13	NP_066530	32264/32380	32728/32871	465	492	95	1.00E-56		65	N/A	N/A	
Nad4	NADH dehydrogenase subunit 4	NP_066531	32748/32889	34136/34313	1389	1425	97	0		77	N/A	N/A	
YejU	ABC transporter subunit	NP_066532	34137/34318	34784/34950	648	633	102	5.00E-10		45	N/A	N/A	
Nad2	NADH dehydrogenase subunit 2	NP_066533	34785/34969	36263/36447	1479	1479	100	9.00E-178		60	N/A	N/A	
Ser tRNA	Serine tRNA	NC_002573	36303/36559	36377/36633	75	75	100	N/A		93	42.59	N/A	
ATP6	ATP synthase F0 subunit 6	NP_066534	36420/36666	37163/37418	744	753	99	5.00E-131		80	N/A	N/A	
Nad3	NADH dehydrogenase subunit 3	NP_066535	37187/37443	37549/37805	363	323	100	7.00E-73		87	N/A	N/A	
Cox2	Cytochrome c oxidase subunit 2	NP_066536	37689/37967	38546/38812	858	846	101	0		89	N/A	N/A	
Orf504	Orf 504	NP_066537	38636/38895	40219/40409	1584	1515	105	9.00E-177		50	N/A	N/A	
Ser tRNA	Serine tRNA	NC_002573	40243/40428	40316/40501	74	74	100	N/A		86	64.91	N/A	
Nad7	NADH dehydrogenase subunit 7	NP_066538	40370/40546	41557/41733	1188	1188	100	0		91	N/A	N/A	
Ile tRNA	Isoleucine tRNA	NC_002573	41557/41981	41629/42053	73	73	100	N/A		89	76.42	N/A	
yejR	Haem lyase	NP_066539	41636/42054	43024/43478	1389	1425	97	3.00E-46		41	N/A	N/A	
Tyr tRNA	Tyrosine tRNA	NC_002573	43056/43506	43138/43588	83	83	100	N/A		90	73.27	N/A	
Ymf16	SecY-independent transporter protein	NP_066540	43404/43856	44192/44650	789	795	99	4.00E-24		39	N/A	N/A	
Ser tRNA	Serine tRNA	NC_002573	44210/44662	44283/44736	74	75	99	N/A		91	55.66	N/A	
Leu tRNA	Leucine tRNA	NC_002573	44289/44740	44370/44821	82	82	100	N/A		87	57.85	N/A	
Asp tRNA	Aspartic acid tRNA	NC_002573	44374/44824	44447/44896	74	73	101	N/A		96	74.55	N/A	
Leu tRNA	Leucine tRNA	NC_002573	44455/44910	44535/44992	81	83	98	N/A		95	61.14	N/A	
Rps4	Ribosomal protein S4	NP_066541	44566/45022	45897/46470	1332	1449	92	1.00E-78		45	N/A	N/A	
Met tRNA	Methionine tRNA	NC_002573	45905/46495	45976/46566	72	72	100	N/A		93	69.23	N/A	
ATP9	ATP synthase F0 subunit 9	NP_066542	46047/46636	46265/46854	219	219	100	4.00E-45		99	N/A	N/A	
Lys tRNA	Lysine tRNA	NC_002573	46288/46947	46360/47019	73	73	100	N/A		90	82.19	N/A	
Met tRNA	Methionine tRNA	NC_002573	46367/47022	46438/47093	72	72	100	N/A		89	71.01	N/A	
Arg tRNA	Arginine tRNA	NC_002573	46970/47161	47042/47233	73	73	100	N/A		96	66.09	N/A	
Nad9	NADH dehydrogenase subunit 9	NP_066543	47070/47267	47627/47827	588	561	105	5.00E-103		75	N/A	N/A	
Glu tRNA	Glutamic acid tRNA	NC_002573	47662/47858	47734/47930	73	73	100	N/A		93	54.27	N/A	
Rns	Small subunit rRNA	NC_002573	47889/48138	49451/49716	1563	1579	99	N/A		87	N/A		589.4

[1]The accession given for homologues of RNAs is that of the entire *N. gruberi* mitochondrial genome (NC_002573). Percent identity was calculated with respect to the sequence feature with a corresponding annotation. For protein-coding genes, the accession numbers of the *N. gruberi* homologues are given. [2]In relation to the *N. gruberi* orthologue.

Supplementary Table ST6.2. Read mapping rate of *N. fowleri* LEE samples to *N. fowleri* V212 predicted proteins

Sample	Read pair mapping rate (to predicted genes)
AX1	61.7% concordant pair alignment rate
AX2	58.7% concordant pair alignment rate
AX3	58.4% concordant pair alignment rate
MP2	55.3% concordant pair alignment rate
MP3	61.6% concordant pair alignment rate
MP4	66.3% concordant pair alignment rate

Supplementary Table ST6.3. Base frequencies for RDAKTTTYHGKWGTT motif found upstream of lysosomal genes up-regulated in highly pathogenic *N. fowleri*

	A	**C**	**G**	**T**
R	0.45	0.05	0.5	0
D	0.4	0	0.4	0.2
A	0.65	0	0.2	0.15
K	0	0	0.55	0.45
T	0	0	0	1
T	0.05	0	0.15	0.8
T	0	0	0.25	0.75
Y	0.05	0.35	0	0.6
H	0.2	0.4	0	0.4
G	0.2	0.1	0.6	0.1
K	0.05	0.05	0.5	0.4
W	0.35	0	0	0.65
G	0	0	0.85	0.15
T	0.1	0.1	0	0.8
T	0	0.1	0	0.9

letter-probability matrix: alength= 4 w= 15 nsites= 20 E= 1.0e-005

Online Appendices

Due to space and formatting considerations, some supplementary information is found in Online Appendices, which can be accessed at the following link at this time.

Online Appendices Google Drive