Evolution of the vesicle formation machinery

by

Alexander Douglas William Schlacht

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

Doctor of Philosophy

Department of Cell Biology
University of Alberta

© Alexander Douglas William Schlacht, 2016

Abstract

One of the features that distinguishes eukaryotes from prokaryotes is the membrane trafficking system. This system underpins much of the functionality of the eukaryotic cell, and is necessary for feeding, motility and communication. Analyses aimed at addressing the evolution of this system have revealed tremendous complexity in the Last Eukaryotic Common Ancestor, pointing to an even earlier origin. However, the events giving rise to this system and its subsequent diversification are poorly understood.

Comparative genomic and phylogenetic analyses aimed at addressing the evolution of the membrane trafficking system have focused largely on the machinery of vesicle fusion. Here, I examine the evolution of machinery involved in vesicle formation. Comparative genomic and phylogenetic analyses were used to assess the conservation and evolution of protein families involved in the regulation of vesicle coat formation. ArfGAPs and ArfGEFs are regulatory proteins for the small GTPase Arf, a key regulator of vesicle coat formation. I found that five of ten previously identified ArfGAP subfamilies were present in the LECA, and I identify a previously unreported ArfGAP subfamily that is absent from humans and yeast. I also report that the LECA possessed three of the six described ArfGEF subfamilies.

COPII is a heteromeric coat complex necessary for the transport of cargo from the endoplasmic reticulum to the Golgi apparatus. I found that all seven components were present in the LECA, five of which are ubiquitously conserved across eukaryotes. The other two are frequently missing.

TSET is a newly identified adaptin-like coat complex with a likely role in endocytosis. I found that this complex is broadly distributed, indicating its presence

in the LECA. However, it has been frequently lost from multiple eukaryotic lineages, including that giving rise to humans and fungi.

Analysis of these gene families revealed multiple patterns of protein conservation, including ubiquitous and lineage-specific patterns, but also components with a "patchy" distribution that had previously been underappreciated. That some of these are missing from traditional cell biological systems such as humans and yeast, suggests the need to consider other eukaryotes as model organisms in order to fully comprehend the diversity of eukaryotic cell biology. Analysis of these components also revealed some of the earliest events that may have occurred during the evolution of the trafficking system, in addition to convergent roles of multiple coat complexes in membrane trafficking.

Preface
(Mandatory due to collaborative work)

Much of the work in this thesis is the product of multiple international research collaborations. I collected all the data, and performed all the analyses, and generated text unless otherwise indicated.

A portion of Chapter 3 has been published as: Schlacht, A., Mowbrey, K., Elias, M., Kahn, R.A., Dacks, J.B. 2013. Ancient complexity, opisthokont plasticity, and discovery of the 11th subfamily of Arf GAP proteins. Traffic 14: 636-649. R.A. Kahn and J.B. Dacks were responsible for conception of the project. M. Elias conceived the methodological validation outlined in Chapter 3. K. Mowbrey performed the initial data collection. I completed the data collection, performed all analyses, and composed the first draft of the manuscript. All were involved in editing the manuscript. The manuscript was re-written for inclusion in this thesis.

Chapter 4 has been published as: Schlacht, A., Dacks, J.B. 2015. Unexpected ancient paralogues and an evolutionary model for the COPII coat complex. Genome Biology and Evolution 7: 1098-1109. J.B. Dacks originally conceived the project. I planned and carried out all analyses. I composed and edited the first draft of the manuscript. J.B. Dacks edited subsequent versions. An early version of the manuscript was edited for inclusion in this thesis.

Chapter 5 has been published as: Hirst, J., Schlacht, A., Norcott, J.P., Traynor, D., Bloomfield, G., Antrobus, R., Kay, R.R., Dacks, J.B., Robinson, M.S. 2014. Characterization of TSET, an ancient and widespread membrane trafficking complex. eLife 3: e02866. J. Hirst and M.S. Robinson were responsible for conception of the project. J.P. Norcott was responsible for database construction (not presented

in this thesis). J. Hirst, D. Traynor, G. Bloomfield, R. Antrobus, R.R. Kay carried out *in vivo* and biochemical analyses. J.B. Dacks and I conceived the bioinformatic analyses (*i.e.,* comparative genomic analyses, phylogenetic analyses, and structural predictions) which I then carried out. J. Hirst, M.S. Robinson, and J.B. Dacks composed the manuscript; all were involved in editing. With the exception of figure legends, which I composed for the manuscript, all of the text here was generated *de novo*.

A portion of Chapter 6 has been published as: Schlacht, A., Herman, E.K., Klute, M.J., Field, M.C., Dacks, J.B. 2014. Missing pieces of and ancient puzzle: evolution of the eukaryotic membrane-trafficking system. In *The Origin and Evolution of Eukaryotes* (Eds. Keeling, P.J. and Koonin, E.V.) Cold Spring Harbor press. I wrote the initial draft of a portion of the manuscript. E.K. Herman and M.J. Klute carried out data collection. M.C. Field and J.B. Dacks composed the remainder of the manuscript. All edited the manuscript.

Dedication

To Andrea: for your unconditional love and support.

To Isabelle: remain forever curious.

Acknowledgments

First and foremost I would like to thank my supervisor and mentor, Joel Dacks. Thank you for introducing me to the world of protistology and molecular evolution. You provided ideas in the form of projects, but always encouraged me to make them my own. You fostered an environment where I could pursue questions that I thought were important and provided the support and guidance to do so. The definition of success that you have set for yourself is one that I will forever strive to achieve. Most importantly, thank you for relentlessly pushing me beyond what I thought I was capable of myself.

Thank you to all of the members of the Dacks lab, past (Jeremy Wideman, Megan Chiu, Laura Lee, Mary Klute) and present (Emily Herman, Maria Aguillar Gonzalez, Beth Richardson, Chris Klinger, Lael Barlow, Kelly Zerr), official or squatters (Fred Mast, Giselle Walker, Anna Karnkowska). Thank you for all of the discussion, scientific or otherwise. The people present make work environments great and this lab has been truly exceptional; there's nowhere else that I would have rather spent the last four and a half years.

Thank you to all of the collaborators who have allowed me to be a part of their exciting projects. Giselle Walker and Richard Dorrell, thank you for putting up with a freshly minted undergrad who barely knew what a protist was while tasked with writing about the diversity of these amazing organisms. Richard, we'll always have Seattle. Kevin Mowbrey, Marek Elias, and Richard Kahn, thank you for all of your help and input, I will always hold onto the discovery of ArfGAPC2, which was made possible only by the contributions made by each of you. Jennifer Hirst, John

Acknowledgments

Norcott, David Traynor, Gareth Bloomfield, Robin Antrobus, Robert Kay, and Scottie Robinson, working with you on the TSET project was one of the most exciting times of my degree. Thank you for making me a part of it. Nicola De Francheschi, Klemens Wild, Irmgard Sinning, and Francesco Filippini, working with you gave me new appreciation of the information that can be gleaned from predicting protein structures. I have taken that forward with me on every project since.

I would like to especially thank Lorne Leclair and Igor Sinelnikov, who maintained and built (at least once) our local computing clusters: Raptor, Hydra, and Bioinfor. Had it not been for the two of you, this work very likely would not have been completed.

To my parents Bill and Lynn, and my brothers: Matt, Nick, and Tom, thank you for your encouragement, support, and excitement throughout this process.

To my wife Andrea, thank you for all of your unconditional love and support. You are my inspiration for doing this everyday. Thank you for embracing my inner geek.

Table of Contents

Table of Contents

Table of Contents

List of Tables

List of Figures

List of Figures

List of Figures

List of Figures

List of Figures

List of Figures

List of abbreviations

ACAP = ArfGAP and coiled-coil domain-containing protein

ADAP = ArfGAP with Dual PH domains

AGAP = ArfGAP and GTPase domain-containing protein

AGFG = ArfGAP domain and FG repeat-containing protein

AP = Adaptor Protein

ARAP = ArfGAP AND RhoGAP domain containing protein

Arf = ADP-ribosylation factor

ArfGAP1 = ADP-ribosylation factor GTPase Activating Protein 1

ArfGAP2 = ADP-ribosylation factor GTPase Activating Protein 2

ArfGAP3 = ADP-ribosylation factor GTPase Activating Protein 3

ASAP = ArfGAP and SH3 domain-containing protein

BIG = Brefeldin A-Inhibited Guanine Nucleotide Exchange Factor

BRAG = Brefeldin A-Resistant Arf GEF

CCTH = Cryptophytes, Centrohelids, Telonemids, and Haptophytes

CYTH = Cytohesin

DSCR3 = Down Syndrome Critical Region 3

EFA6 = Exchange Factor for Arf6

ER = Endoplasmic Reticulum

ESCRT = Endosomal Sorting Complex Required for Transport

FBX8 = F-box Only Protein 8

GAP = GTPase Activating Protein

GBF = Golgi-Specific Brefeldin A-Resistant Guanine Nucleotide Exchange Factor

GEF = Guanine Nucleotide Exchange Factor

GIT = G-protein Interacting ArfGAP

HMM = Hidden Markov Model

LECA = Last Eukaryotic Common Ancestor

MTC = Multisubunit Tethering Complex

List of abbreviations

NPC = Nuclear Pore Complex

Nup = Nucleoporin

OPH = Organelle Paralogy Hypothesis

PH = Pleckstrin Homology

SM = Sec1/Munc18

SMAP = Small ArfGAP

TBC = Tre-2/Bub2/Cdc16 domain-containing protein

TGN = *trans*-Golgi Network

Chapter 1: Introduction

1.1 Introduction

The evolution of the eukaryotic cell represented a monumental transition in the history of life on Earth. This landmark was a key point in the evolution of the largest, most terrifying, and most beautifully complex organisms to move about the planet. The transition from prokaryote to eukaryote has been described as "the most well-known fundamental dichotomy in biology" (Sapp, 2005). However, the nature of this transition remains largely enigmatic.

Eukaryotes are structurally distinct from their prokaryotic counterparts through the possession of discrete, membrane-bound compartments (Chatton, 1938; Stanier and van Niel, 1962). The most prominent of these is the eukaryotic namesake, the nucleus. Myriad other organelles with specific biochemical capabilities are also present. A subset of these constitutes the endomembrane system, a network of interconnected organelles responsible for the distribution of proteins and lipids throughout the cell and for communication with the extracellular environment. This membrane trafficking system is involved in cell motility and would have allowed early eukaryotes to export proteins and remodel their cell surface (Cavalier-Smith, 2002). Exocytosis would have allowed the modification of their extracellular environment, while endocytosis would have allowed the uptake of material from the environment, ultimately allowing the occupation of novel ecological niches (Cavalier-Smith, 1975; de Duve, 2007; De Duve and Wattiaux, 1966; Stanier, 1970).

Cell biological analyses in humans and yeast have been tremendously important for establishing the machinery responsible for membrane trafficking,

identifying numerous coats (Barlowe et al., 1994; Kamiguchi et al., 1998; Zhu et al., 1999), GTPases (Cukierman et al., 1995; Kahn and Gilman, 1986), SNAREs (soluble N-ethylmaleimide sensitive factor attachment protein receptor; Weber et al., 1998), and other machinery necessary for this system (Andag and Schmitt, 2003; Conibear and Stevens, 2000; TerBush et al., 1996; Whyte and Munro, 2001). However, the recognition of at least 5 other major lineages of eukaryotes prompted the search for a novel understanding of which aspects of membrane trafficking are broadly conserved across eukaryotes and which are unique characteristics of human and yeast systems.

Evolutionary biology has long been searching for the origin of eukaryotic organelles. For some, their connection to the prokaryotic world has already been made clear: mitochondria and chloroplasts are vestiges of ancient endosymbioses (Bonen et al., 1977; Doolittle and Bonen, 1981; Gray and Doolitle, 1982 *inter alia*; Margulis, 1970), as genes encoded by organellar genomes show phylogenetic affinity to α-proteobacteria and cyanobacteria, respectively. For organelles of the membrane trafficking system the situation is much less clear. Tracking the evolution of organelles involved in membrane trafficking is much more difficult. Genes that encode important trafficking factors are located in the nucleus, not in organellar genomes. However, the identification of genes whose products act at or localize to specific organelles has been used to identify equivalent structures in different organisms. Comparative genomics (*i.e.,* the identification of equivalent genes in different genomes) has been used to determine the presence or absence of membrane trafficking pathways and organelles *in silico* by searching for markers of

different trafficking steps in diverse eukaryotic genomes. Early comparative genomic analyses revealed that the major protein families involved in membrane trafficking are present across the diversity of eukaryotes (Dacks and Doolittle, 2002; Dacks and Field, 2004; Leung et al., 2008; Schledzewski et al., 1999). Subsequently, more detailed analysis of specific subfamilies revealed that many of these are also broadly conserved (Dacks and Doolittle, 2002; Dacks and Doolittle, 2004; De Franceschi et al., 2014; Elias et al., 2012; Gabernet-Castello et al., 2013; Koumandou et al., 2007; Koumandou et al., 2011; Murungi et al., 2014; Sanderfoot, 2007). This observation, paired with functional studies revealing largely conserved function and localization of many of these proteins (Chong et al., 2010; De Craene et al., 2014; Koumandou et al., 2011; Manna et al., 2015; Sauer et al., 2013; Skruzny et al., 2015; Turkewitz and Bright, 2011), suggested that by analyzing the distribution and evolution of organelle-specific members of different protein families, we can track the evolution of the organelles of the membrane trafficking system. Furthermore, homologues (*i.e.,* genes that share a common ancestor) of multiple protein families important for membrane trafficking such as small GTPases of the Ran, Rab, Rho, and Arf families have been identified in prokaryotes (Dong et al., 2007; Jékely, 2003; Wuichet and Sogaard-Andersen, 2014; Yutin et al., 2009), indicating that links between prokaryotes and the eukaryotic membrane trafficking system may, in fact, exist.

In this introductory chapter, I outline background information necessary for exploring the evolution of the membrane trafficking system. I provide an overview of eukaryote diversity, relevant for understanding the comparative genomic and

evolutionary aspects of this study. I then provide an overview of membrane trafficking, necessary to understand the system under investigation, after which follows discussion of major hypotheses concerning the origin of the eukaryotic endomembrane system. Lastly, I specify the questions addressed in this thesis.

1.2 Eukaryotic diversity

The availability of sequenced eukaryotic genomes and the tools to accurately identify common genes between organisms has allowed inquiry into the evolutionary history of the membrane trafficking system, with the ultimate goal of reconstructing the trafficking system of the Last Eukaryotic Common Ancestor (LECA); the hypothetical ancestor that gave rise to all lineages of eukaryotes. The conceptual basis for reconstructing the LECA is essentially a parsimony argument; if the majority of descendants of a single ancestor possess a specific trait, such as a gene, a pathway, or an organelle, then the most likely explanation is that the trait in question was also present in the ancestor of those organisms. The alternative explanation would either be Horizontal Gene Transfer (HGT) or convergent evolution. HGT can be ruled out using phylogenetic analysis to deduce the evolutionary history of the protein in question (Ku et al., 2015), while convergent evolution can be ruled on the basis that comparative genomic techniques should only identify related proteins, and that proteins that have evolved the same function independently should not identify one another using these techniques. The LECA is a tremendously valuable point of reconstruction as it acts as a point of reference to understand when various biological processes evolved, and provides historical

context for subsequent evolution in diverse eukaryotic lineages. The presence of broadly conserved cellular machinery can set in place expectations of function; if the homologue of a characterized protein is identified in a distantly related organism, we can use what is known about the function of that protein in one system as a working hypothesis for what it might be doing in other systems. This is especially helpful in species where *in vivo* analyses cannot be carried out, as knowing which members of a protein complex or pathway are present can provide insight into the basic biology of that species. For example, the broad conservation of nuclear pore subunits (Nups) across eukaryotes sets in place an expectation of nuclear pore function in these organisms (Field et al., 2014). Similarly, the identification of midbody proteins, a transient structure that bridges the two daughter cells at then end of cytokinesis, in diverse eukaryotes (Eggert et al., 2006; Eme et al., 2009) sets in place functional hypotheses of how cytokinesis occurs in these organisms. By identifying broadly conserved biological pathways, we can begin to reconstruct the pathways that were present in the LECA and begin to understand some of its biology.

Large-scale comparative genomic analyses have identified the major protein families involved in membrane trafficking (*e.g.* syntaxins, Rabs, coats, *etc.*) in all major eukaryotic lineages, indicating that they would also have been present in the LECA (Dacks and Doolittle, 2001; Dacks and Field, 2004). More detailed analyses of organelle-specific trafficking pathways and machinery also identified near complete systems that would have been present in the LECA (Field et al., 2007b; Koumandou et al., 2013; Leung et al., 2008). For example, major coat complexes (COPI, COPII,

adaptins, clathrin, retromer), small GTPases (Arfs, Sar, Rabs), syntaxins, EpsinR, and ESCRTs (Endosomal Sorting Complex Required for Transport) are all found across the diversity of eukaryotes, indicating that they were also present in the LECA. Moreover, analyses of paralogous protein families, such as, adaptins (Hirst et al., 2011), syntaxins (Dacks and Doolittle, 2002; Dacks and Doolittle, 2004), Rabs (Elias et al., 2012), and TBC (Tre-2/Bub2/Cdc16; Gabernet-Castello et al., 2013) proteins, showed the conservation of organelle-specific paralogues across eukaryotes and by extension, their presence in the LECA.

In order to reconstruct the LECA, we first need an accurate picture of eukaryote diversity. Understanding how different eukaryotic lineages are related allows us to map the distribution of membrane trafficking genes across these lineages and infer points of origin, loss, and duplication. For example, identifying genes that are found across all eukaryotic lineages would suggest that these genes were present in the LECA. By contrast, genes that are present only in a few closely related lineages would likely have arisen in the ancestor of those few lineages, and therefore evolved much more recently.

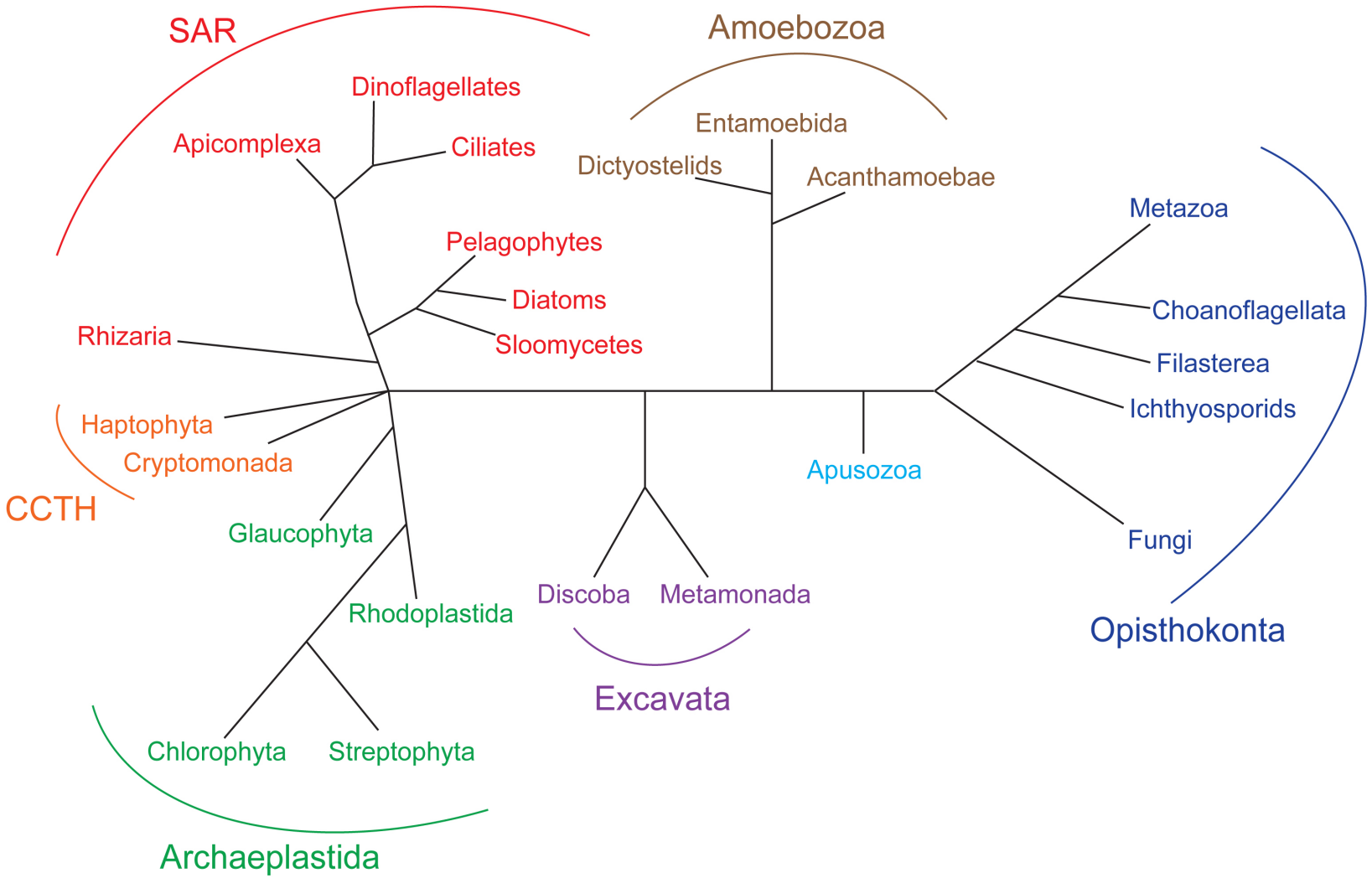
Large phylogenomic analyses concatenating (*i.e.,* stringing together) hundreds of genes supported the division of eukaryotes into six major eukaryotic supergroups, as well as an array of additional lineages that do not belong to any one supergroup, or whose phylogenetic affinities are not yet clear (Adl et al., 2005 *inter alia*; Adl et al., 2012 *inter alia*; Bapteste et al., 2002; Burki et al., 2007; Burki et al., 2008; Burki et al., 2009; Hampl et al., 2009; Parfrey et al., 2010; Rodríguez-Ezpeleta et al., 2005). The classification and resolution of diverse eukaryotic lineages has

been enormously informative for dissecting the relative timing and the order of major evolutionary transitions, such as the acquisition of primary plastids, the multiple origins of multicellularity, and for placing parasitic or pathogenic organisms in an evolutionary context with free-living relatives (Parfrey and Lahr, 2013; Walker et al., 2011). Although a perfectly resolved tree of eukaryotes has yet to be produced, efforts have been greatly helped by the steady increase in the number of sequenced eukaryotic genomes from taxa spanning the diversity of eukaryotes. To give greater context to the sequenced genomes sampled the analyses presented in this thesis, each of the six eukaryotic supergroups is described here. Formal taxonomic names will be provided when taxa are introduced and are capitalized (*e.g.,* Metazoa, Opisthokonta) and will subsequently be used interchangeably with informal taxonomic names (*e.g.,* metazoans, opisthokonts).

1.2.1 Opisthokonta

Opisthokonta comprise Metazoa (animals), Choanozoa (choanoflagellates, *Capsapora owczarzaki, Sphaeroforma arctica*), and Fungi (Figure 1-1, dark blue). Choanoflagellates include *Monosiga brevicollis* and *Salpingoeca rosetta*, and along with *C. owczarzaki* and *S. arctica,* represent the most closely related single-celled organisms to Metazoa. Comparative genomic and transcriptomic analyses of these organisms have revealed that they all possess proteins and domains identified as important for multicellularity in animals, such as integrins (Sebé-pedrós et al., 2010), fibronectin (Sebé-Pedrós et al., 2013), and cadherin (Suga et al., 2013). Opisthokonta was originally supported by multiple morphological and genomic

Figure 1-1. Overview of eukaryote diversity. Supergroups are colour coded, with internal branches representing major lineages sampled in subsequent chapters. The Apusozoa is a distinct lineage from all other supergroups, but is often associated with Opisthokonta. Lineage names and relationships are based on names and definitions from Adl et al., (2005, 2012), Walker et al., (2011) and the Origins of multicellularity sequencing project (Broad institute of Harvard and MIT, www.broadinstitute.org).


SAR
Amoebozoa
Dinoflagellates
Apicomplexa
Ciliates
Entamoebida
Dictyostelids
Acanthamoebae
Metazoa
Pelagophytes
Diatoms
Choanoflagellata
Rhizaria
Sloomycetes
Filasterea
Ichthyosporids
Haptophyta
Cryptomonada
Apusozoa
CCTH
Glaucophyta
Fungi
Discoba
Metamonada
Rhodoplastida
Opisthokonta
Excavata
Chlorophyta
Streptophyta
Archaeplastida

features including: a single posteriorly directed flagellum (Cavalier-Smith, 1988), a pair of centrioles (Cavalier-Smith, 1987a), flat mitochondrial cristae (Cavalier-Smith, 1987a), and a unique 12 amino acid insertion in the elongation factor 1-alpha gene (Baldauf and Palmer, 1993). Opisthokonta is also supported by phylogenomic evidence (Baldauf, 2000; Hampl et al., 2009; Torruella et al., 2012). Opisthokonta is the source of the majority of our understanding of cellular function as it contains the vast majority of model organisms, including *Homo sapiens* and *S. cerevisiae*, in addition to other animal and fungal systems such as *Mus musculus, Rattus norvegicus, Drosophila melanogaster, Caenorhabditis elegans, Neurospora crassa,* and *Pichia pastoris*. Sequenced genomes include: *H. sapiens*, *S. cerevisiae, M. musculus*, *D. melanogaster, C. elegans*, and *C. owczarzaki*, among many others. Phylogenetic analyses have placed *Thecamonas trahens* as the closest outgroup to opisthokonts and is often grouped with them even though it does not satisfy the morphological characteristics that define the Opisthokonta. Specifically, *T. trahens* possesses two flagella, one anterior and one posterior (Vickerman et al., 1974) and tubular mitochondrial cristae (Karpoff and Zhukov, 1986; Molina and Nerad, 1991; Figure 1-1, light blue).

1.2.2 Amoebozoa

The Amoebozoa is a group of primarily single-celled organisms with variable cellular structure (Figure 1-1, brown; Adl et al., 2005; Page, 1987; Shadwick et al., 2009). Members of the Amoebozoa include pathogenic organisms such as

Entamoeba histolytica and *Acanthamoeba castellanii* the causative agents of amoebic dysentery and amoebic encephalitis, respectively (Dart et al., 2009; Stanley, 2003). The internal branching of this clade has not entirely been resolved. Major groups include Entamoebida, Acanthamoebae, and Dictyostelids (Adl et al., 2005; Bapteste et al., 2002; Walker et al., 2011). Amoebozoans typically possess branched, tubular mitochondrial cristae, and when flagellated possess a flagellum supported by a single basal body (Cavalier-Smith, 1998). The monophyly of Amoebozoa is also supported by phylogenomic evidence (Baldauf, 2000; Hampl et al., 2009). Organisms from this supegroup in which tagging and knockouts can be carried out include: *Dictyostelium discoideum* and *E. histolytica.* Genome sequences for the above organisms, as well as *Polysphondylium pallidum* have been completed (Clarke et al., 2013; Eichinger et al., 2005; Heidel et al., 2011; Loftus et al., 2005; Sucgang et al., 2011).

1.2.3 Excavata

The Excavata is home to a diverse group of eukaryotes, many of which are resident to oxygen-deficient environments and may possess hydrogenosomes or mitosomes, anaerobic mitochondria that have undergone reductive evolution (Figure 1-1, purple; Walker et al., 2011). Hydrogenosomes produce iron-sulphur clusters for incorporation into nascent proteins and they produce molecular hydrogen as a by-product of ATP production. By contrast mitosomes are only involved in the production of iron-sulphur clusters (For reviews of hydrogenosomes and mitosomes, see Hjort et al., 2010; Makiuchi and Nozaki, 2014). Most excavates

were originally united by the presence of a ventral feeding groove, into which food particles are directed by a posteriorly directed flagellum (Simpson et al., 2002). The monophyly of the Excavata is somewhat contentious, as phylogenomic support for this group is relatively weak compared to other supergroups (Hampl et al., 2009). Nonetheless, two major lineages have been recognized in the Excavata: Metamonada, which includes the human parasites *Giardia lamblia* and *Trichomonas vaginalis* (Cavalier-Smith, 2003), and Discoba, which includes free-living organisms such as *Naegleria gruberi*, and the human parasites *Trypanosoma brucei* and *Leishmania major* (Hampl et al., 2009). Organisms from this supergroup in which tagging, knockouts, RNAi can be carried out include: *L. major, T. brucei,* and *G. lamblia*. Sequenced genomes available for this supergroup include the organisms listed here, in addition to *Naegleria gruberi, Bodo saltans*, and many others, mostly from the Discoba (Kinetoplastida).

1.2.4 Archaeplastida

Archaeplastida is a group of photosynthetic organisms united by the presence of a primary plastid, resulting from an endosymbiotic event with a cyanobacterium (Figure 1-1, green; Delwiche et al., 1995; Mereschkowsky, 1905). The monophyly of the Archaeplastida is supported by morphological evidence, including the possession of flat mitochondrial cristae and a primary photosynthetic plastid, in addition to multiple lines of phylogenomic evidence (Adl et al., 2005, *inter alia*; Price et al., 2012; Rodríguez-Ezpeleta et al., 2005). Archaeplastida includes green algae (*e.g., Chlamydomonas reinhardtii*), land plants (*e.g., Physcomitrella*

patens, *Arabidopsis thaliana*), red algae (*e.g., Cyanidioschyzon merolae*), and glaucophytes (*e.g., Cyanophora paradoxa*). Organisms from this supergroup in which tagging and knockouts can carried out include: *A. thaliana, P. patens,* and *V. carteri*. At least 50 genomes are available for this group in addition to the species mentioned here.

1.2.5 The SAR clade

The SAR clade is composed of three major lineages: Stramenopiles, Alveolates, and Rhizarians and are supported by phylogenomic evidence (Figure 1-1, red; Burki et al., 2007; Burki et al., 2008; Cavalier-Smith, 2010). Stramenopiles include multicellular and unicellular brown algae, (*e.g., Ectocarpus siliculosus* and *Nannochloropsis gaditana*, respectively), diatoms (*e.g., Thalassiosira pseudonana*), and sloomycetes (*e.g., Phytophthora sojae*). Alveolates include ciliates, (*e.g., Paramecium tetraurelia*), and apicomplexans, (*e.g., Plasmodium falciparum,* the causative agent of malaria; Walker et al., 2011). Rhizaria includes an array of photosynthetic, (*e.g., Bigelowiella natans*), and non-photosynthetic (*Reticulomyxa filosa*) species. Organisms from this supergroup in which tagging and knockouts can be carried out include: *P. falciparum, Toxoplasma gondii, Tetrahymena thermophila,* and *P. tetraurelia*. Genome sequences are available for all of the organisms mentioned above, as well as at least 40 additional alveolate and stramenopile species.

1.2.6 The CCTH group

The CCTH clade is composed of cryptophytes, centrohelids, telonemids, and haptophytes (Figure 1-1, orange; Burki et al., 2008; Burki et al., 2009). However, recent evidence suggests that these lineages may not form a single taxonomic unit, but rather are independent lineages (Burki et al., 2012). Despite this uncertainty, it is clear that they are closely related to the SAR group and to the Archaeplastida. No protocols for tagging or knockouts have been developed for any organisms in this supergroup. Sequenced genomes are available for the haptophyte *Emiliania huxleyi* and the cryptomonad *Guillardia theta* (Curtis et al., 2012; Read et al., 2013).

1.2.7 Incertae sedis taxa

Some eukaryotes have been identified whose position on the eukaryotic tree is currently uncertain. Examples of these groups include Breviatea (*e.g., Breviata anathema*; Cavalier-Smith et al., 2004), Collodictyonidae (*e.g., Collodyction tricilliatum*; Brugerolle et al., 2002), and Malawimonads (*e.g., Malawimonas jakobiformis;* O'Kelly and Nerad, 1999). The difficulty in placing these species on the tree of eukaryotes often stems from inconsistency between morphological and phylogenetic analyses. For example, *M. jakobiforms* possesses features, that would suggest it is related to excavates, notably a ventral suspension-feeding groove (O'Kelly and Nerad, 1999); however, recent phylogenetic analyses have suggested that it lies elsewhere on the eukaryote tree (Derelle et al., 2015; Katz and Grant, 2015). Although these lineages may be small compared to the other supergroups,

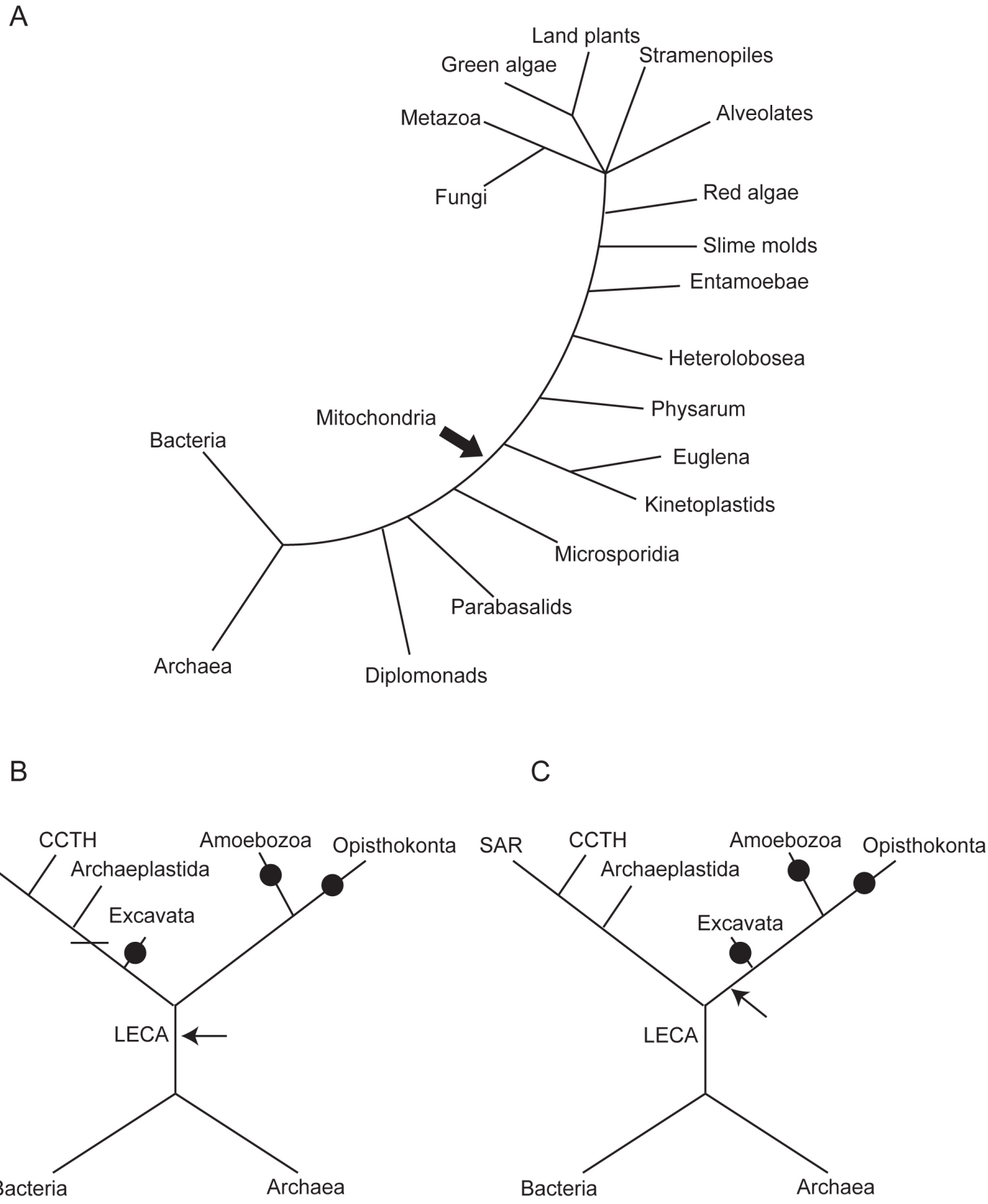
identifying their place on the eukaryotic tree can have enormous ramifications. For example, finding that the putatively bikont *B. anathema* along with *T. trahens* branches within the unikonts, help to over turn the bikont-unikont hypothesis (Minge et al., 2009). Therefore, placing these taxa on the eukaryotic tree will greatly aid our understanding of eukaryote evolution. At present, no genomes have been sequenced for these organisms, nor have any protocols for tagging or knockdowns been developed.

1.2.8 Rooting the tree of eukaryotes

One of the major outstanding questions in evolutionary biology is the position of the root of the eukaryotic tree. The root represents the first bifurcation after the LECA and would have produced two lineages from which all supergroups would have evolved. Pinpointing the location of the root of the eukaryotic tree would not only provide a clearer understanding of eukaryote evolution, but would also allow us to polarize transitions in all eukaryotic lineages.

One of the earliest rooting hypotheses for the tree of eukaryotes was the 'Archezoa hypothesis' (Cavalier-Smith, 1987b). The Archezoa hypothesis suggested that some lineages of eukaryotes, such as microsporidians (*e.g., Encephalitozoon*), *Giardia*, and *Trichomonas* are direct descendants of pre-mitochondriate eukaryotes because they lacked detectable mitochondria, as well as Golgi bodies and peroxisomes (Figure 1-2A). These lineages also appeared to possess a bacterial-like 70S ribosome with 16S and 23S subunits rather than the eukaryotic-like 80S

Figure 1-2. Three rooting hypotheses for the tree of eukaryotes. A) Archezoa hypothesis (Cavalier-Smith, 1987b). Early branching eukaryotes such as Diplomonads (*Giardia*), Parabasalids (*Trichomonas*), and Microsporidia are descendants of pre-mitochondriate eukaryotes. Phylogenetic trees supporting this hypothesis are characterized by the laddered branching of protist groups along the stem of the tree followed by a large radiation of eukaryotic groups giving rise to opisthokonts, archaeplastids, stramenopiles, and alveolates. Tree is a summary of inferences made by Sogin (1989), Vossbrinck et al. (1987). Arrow labelled "Mitochondria" denotes the acquisition of the mitochondrion B) Bikont-Unikont rooting of the eukaryotic tree proposed by Stechmann and Cavalier-Smith (2002) that divides eukaryotes into SAR + CCTH + Archaeplastida + Excavata on one side, and Amoebozoa + Opisthokonta on the other. Black dots indicates the presence of a gene in the Excavata, Amoebozoa, and Opisthokonta. This gene likely appeared in the ancestor of these lineages (the LECA, arrow) as they span the backbone of the tree. Line indicates loss of the gene from the ancestor of Archaeplastida, SAR, and CCTH. C) Alternative rooting of the eukaryotic tree. The same gene with the same distribution as in (B) is now interpreted as having evolved in the ancestor of Excavata, Amoebozoa, and Opisthokonta (arrow), and not present in the LECA, as its distribution does not span the entire backbone of the tree.


A
Land plants
Green algae
Stramenopiles
Metazoa
Alveolates
Fungi
Red algae
Slime molds
Entamoebae
Heterolobosea
Mitochondria
Physarum
Bacteria
Euglena
Kinetoplastids
Microsporidia
Parabasalids
Archaea
Diplomonads
B
SAR
CCTH
Amoebozoa
Opisthokonta
Archaeplastida
Excavata
LECA
Bacteria
Archaea
C
SAR
CCTH
Amoebozoa
Opisthokonta
Archaeplastida
Excavata
LECA
Bacteria
Archaea

ribosome that contains 18S and 28S subunits (Vossbrinck and Woese, 1986). Additional support for this hypothesis came from phylogenetic analysis of the small subunit ribosomal DNA (SSU rDNA) gene that placed the Archezoa basal to the rest of eukaryotes (Sogin, 1989; Vossbrinck et al., 1987). The Archezoa hypothesis was eventually refuted on two lines of evidence. First, the discovery that all Archezoans possessed either hydrogenosomes or mitosomes indicated that they were not pre-mitochondriate eukaryotes, but are highly derived eukaryotes with mitochondria Second, the SSU rDNA sequences were shown to be highly divergent sequences (Arisue et al., 2004; Shirakura et al., 2001); the clustering of these sequences at the base of eukaryotes was the shown to be the result of Long Branch Attraction (LBA), the artificial grouping of rapidly evolving sequences (Dacks et al., 2002; Philippe and Germot, 2000). These findings indicated that all eukaryotes are descendants of a mitochondrion-bearing ancestor.

Many attempts have since been made to root the tree of eukaryotes (Derelle and Lang, 2012; Katz et al., 2012; Pusnik et al., 2011; Richards and Cavalier-Smith, 2005; Rogozin et al., 2009; Serfontein et al., 2010; Stechmann and Cavalier-Smith, 2002). One of the most prominent of these rooting hypotheses is the Bikont-Unikont root (Figure 1-2B; Stechmann and Cavalier-Smith, 2002). Unikonts (*i.e.,* Opisthokonta and Amoebozoa) possess the ancestral condition of a single basal body supporting the flagellar root, whereas bikonts (*i.e.,* Excavata, Archaeplastida, SAR, and CCTH) possess the derived state of two basal bodies (Stechmann and

Cavalier-Smith, 2002). Additional support for this bifurcation came in the form of rare genomic changes. Specifically, unikonts possess separate DHFR (dihydrofolate reductase) and TS (thymidylate synthase) genes whereas these genes are fused in bikonts (DHFR-TS; Stechmann and Cavalier-Smith, 2002). Unikonts were also thought to possess a unique glycine insertion in the Myosin II gene (Richards and Cavalier-Smith, 2005). The Bikont-Unikont root was subsequently rejected by phylogenomic evidence that placed the Apusozoa and the protist *Breviata anathema* in the unikont clade; however, both of these organisms possess bikont flagellar structures (Kim et al., 2006; Minge et al., 2009). Apusozoans were also found to possess the fused DHFR-TS gene thought to only be found in bikonts (Kim et al., 2006). Finally, the genome sequence of the bikont *N. gruberi* revealed the presence of the unikont-type Myosin II gene indicating that these supposedly rare genomic changes were insufficient to pinpoint the eukaryotic root (Fritz-Laylin et al., 2010).

A similar rooting to the Bikont-Unikont hypothesis has recently been proposed (Derelle et al., 2015), that places the root between Opisthokonta, Amoebozoa, Malawimonads, and Collodictyonids on one side, and Discoba, Archaeplastida, SAR, and CCTH on the other. This root is base on concatenated phylogeny of two previously used datasets (Derelle and Lang, 2012; He et al., 2014), both of which use bacteria as outgroups. Although this rooting may appear to be topologically similar to the bikont-unikont rooting, it is distinct, especially with respect to the monophyly of the Excavata. In this analysis, the Discoba is supported as a monophyletic group, but *Malawimonas*, typically thought of as an excavate, branches with the opisthokont – Amoebozoa clade, suggesting that it is either not a

true excavate, or that the Excavata is a polyphyletic group. The uncertainty in this interpretation is further compounded by the exclusion of any metamonads from the analysis. The exclusion of this group was the result of the gene selection process which resulted in a dataset with a large proportion of metabolic genes associated with the mitochondrion (Derelle and Lang, 2012; He et al., 2014), often lost from the highly reduced mitochondrion-related organelles of metamonads.

Although no consensus has been reached on the location of the root of the eukaryotic tree, it is generally agreed that the root most likely lies between Opisthokonta + Amoebozoa and Archaeplastida + SAR + CCTH, with the position of the Excavata lying on one side other the other, or with the root lying within the Excavata (Derelle and Lang, 2012; Derelle et al., 2015; Pusnik et al., 2011; Wideman et al., 2013). The position of the Excavata, and thus the root, has the potential to alter our interpretation of when eukaryotic genes evolved. For example, if a gene is found in most or all supergroups, then it was likely present in the LECA. If a gene is found in the Opisthokonta, Amoebozoa, and Excavata it could have been present in the LECA if a Bikont-Unikont root is considered (Figure 1-2B). In this case, the gene would have been lost from the ancestor of Archaeplastida + SAR + CCTH. Alternatively, if the root lies between Opisthokonta + Amoebozoa + Excavata and Archaeplastida + SAR + CCTH (Figure 1-2C) then parsimony would dictate that the gene was not present in the LECA, but arose in the ancestor of Opisthokonta + Amoebozoa + Excavata. Thus, in the absence of a definitive root of eukaryotes broad sampling of genomes across the eukaryote tree is necessary to understand how and when genes and pathways evolved.

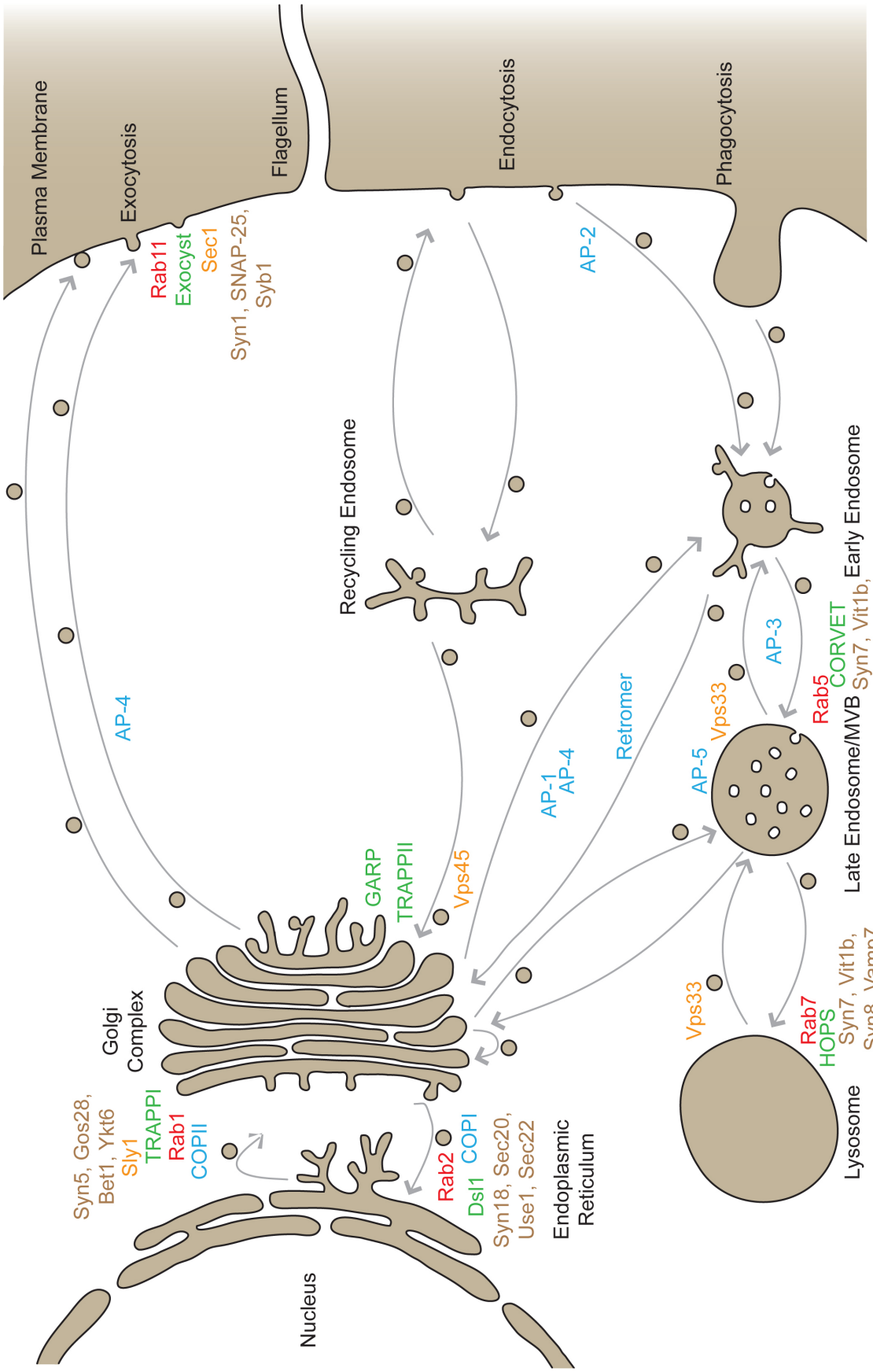
1.3 Overview of the membrane trafficking system

Now that we have an understanding of eukaryotic diversity, we can being to discuss our cellular system of interest: the membrane trafficking system, a network of internal compartments found in all eukaryotes. Although this system has been subject to lineage-specific evolution (Adung'a et al., 2013; Klinger et al., 2013; Sanderfoot, 2007), *in vivo* analyses and comparative genomic analyses suggest that the organelles, transport steps, and protein families that govern transport through this system are largely conserved across the diversity of eukaryotes (Brady et al., 2008; El-Kasmi et al., 2011; Field et al., 2007a; Hall et al., 2004; Langhans et al., 2008; Turkewitz and Bright, 2011; Veltman et al., 2011). Therefore, in order to understand how this system evolved, an understanding of how this system works is required.

1.3.1 Organelles of the membrane trafficking system

The first organelle involved in membrane trafficking, the ER, is a reticulating network of membranes continuous with the nuclear envelope (Figure 1-3; for a recent review, see Lynes and Simmen, 2011). The ER can be divided into multiple subdomains. The rough ER is covered in ribosomes actively translating proteins as they are inserted into the ER lumen or the ER membrane (Blobel and Dobberstein, 1975), the smooth ER is devoid of ribosomes and is responsible for the synthesis of lipids and is important for Ca^{2+} signalling (Bell et al., 1981; Rizzuto et al., 2009), and the transitional ER which marks the site of vesicle formation (ER exit sites) for

Figure 1-3. Overview of the eukaryotic membrane trafficking system. Generic eukaryotic cell depicting a generalization of organelles typically present, along with trafficking pathways (denoted by arrows), and the location of action for major proteins and complexes in the membrane trafficking system. Colour code: blue = coats, red = Rabs, green = tethering complexes, orange = SM proteins, brown = SNAREs. Abbreviations: Syn = syntaxin, Syb = synaptobrevin. Modified from Schlacht et al., 2014.


Plasma Membrane
Exocytosis
Flagellum
Endocytosis
Phagocytosis
Rab11
Exocyst
Sec1
Syn1, SNAP-25,
Syb1
AP-2
Recycling Endosome
AP-4
AP-1
AP-4
Retromer
AP-3
AP-5
Vps33
Rab5
CORVET
Syn7, Vit1b,
Syn6, Vamp7
Early Endosome
Late Endosome/MVB
GARP
TRAPPII
Vps45
Golgi
Complex
Syn5, Gos28,
Bet1, Ykt6
Sly1
TRAPPI
Rab1
COPII
Rab2
Dsl1
COPI
Syn18, Sec20,
Use1, Sec22
Endoplasmic
Reticulum
Nucleus
Vps33
Rab7
HOPS
Syn7, Vit1b,
Syn8, Vamp7
Lysosome

transport of cargo to the Golgi complex (Bannykh et al., 1996; Orci et al., 1991; Palade, 1975).

In mammalian cells, vesicles budding from the ER accumulate, and fuse at the *cis*-face of the Golgi complex. The classical view of the Golgi structure is as a stack of membrane compartments (Warren and Mellman, 2007). Although this Golgi organization has been observed in many eukaryotes (*e.g., Dictyostelium discoideum, Tetrahymena thermophila*; Kurz and Tiedtke, 1993; Schneider et al., 2000), other morphologies, including unstacked (punctate), as in some fungi (Franzusoff et al., 1991), or ribbons (laterally connected stacks), as in metazoans have also been reported (Ladinsky et al., 1999). Transport through the Golgi is typically described by one of two classic models, although neither is sufficient to account for all experimental observations (Jackson, 2009, *inter alia*). The forward vesicular-trafficking model proposes that cargo proteins are transported between permanent cisternae by transport vesicles in a *cis* to *trans* direction (Palade, 1975). The cisternal maturation model proposes that new cisternae form at the *cis*-Golgi that progress through the Golgi while Golgi processing enzymes undergo retrograde vesicular transport to earlier cisterna (Bonfanti et al., 1998; Morre and Mollenhauer, 2007).

Once modified in the Golgi, proteins accumulate at the *trans*-Golgi network (TGN) where they are sorted and transported to their final destinations. Plasma membrane and secretory proteins (*e.g.* glycosylphosphatidylinositol (GPI)-anchored proteins; Paladino et al., 2004) are transported to the cell surface where they are incorporated into the plasma membrane or are released into the extracellular

environment, respectively (Figure 1-3). Alternatively, TGN-derived vesicles fuse with endocytic vesicles from the plasma membrane, generating early endosomes, or are transported to late endosomes or lysosomes (Burgos et al., 2010; Hirst et al., 1999). Late endosomes subsequently fuse with lysosomes, resulting in degradation of endosomal cargo.

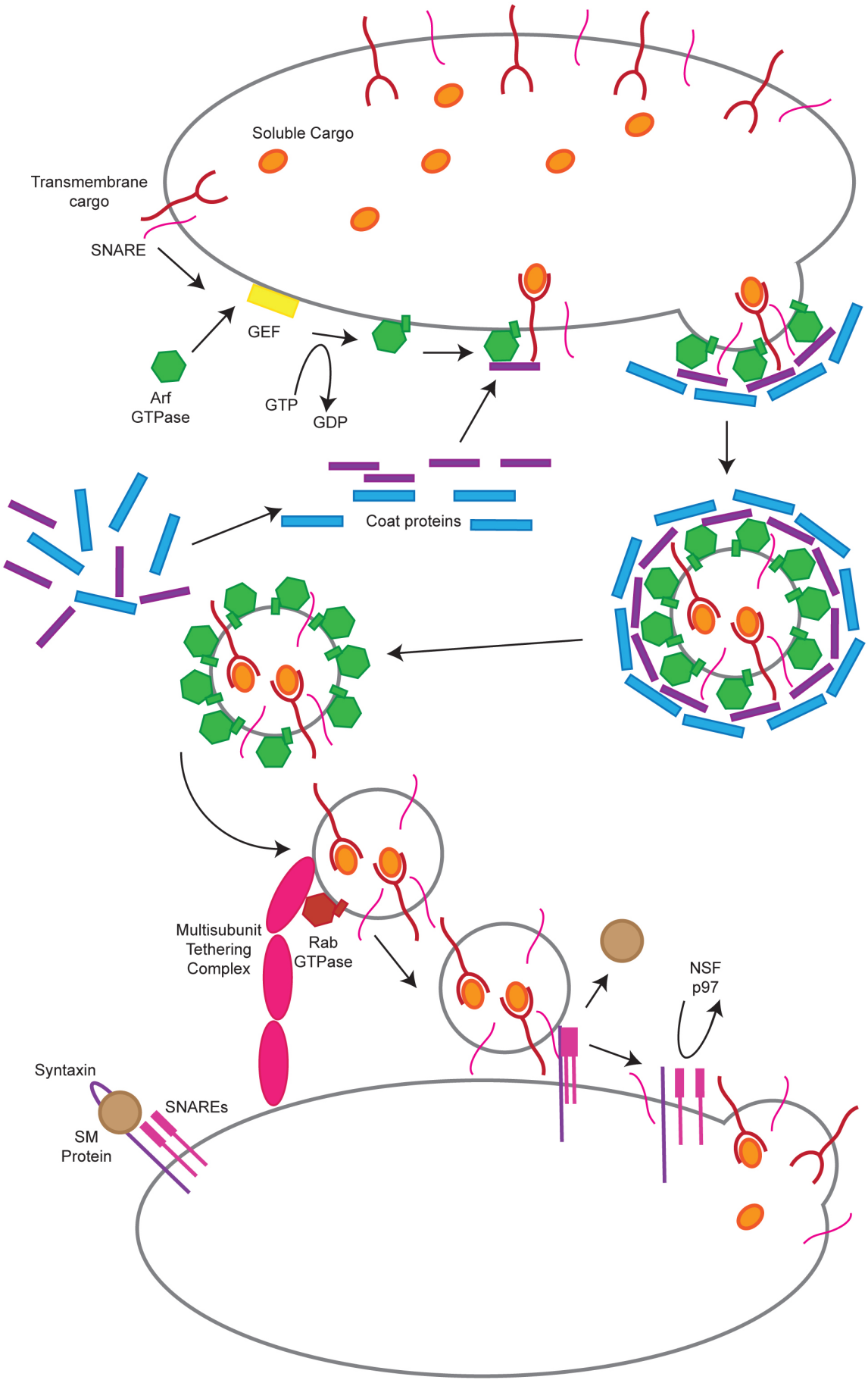
The endosomal system generally displays more plasticity than earlier steps in the membrane trafficking pathway. The overview here is broadly what is thought to occur in mammalian and yeast cells; however, the endocytic system has been modified extensively in some lineages. For example, secretory granules in mammalian cells are thought to be specialized lysosome-related organelles, similar to mucocysts or trichocysts in ciliates, also thought to be derived from lysosomes (Elde et al., 2007).

1.3.2 Mechanisms of trafficking

1.3.2.1 Vesicle formation

The principles and process of vesicle formation and fusion are largely the same at each step in transport pathway; membrane trafficking is mediated by a limited set of protein families, with organelle-specific members carrying out essentially the same function at each trafficking step (Figure 1-4; Bonifacino and Glick, 2004). Nucleation of vesicle formation occurs when an Arf GTPase is recruited to the donor membrane (Spang et al., 1998). Arfs cycle between activated GTP-bound, and inactivated GDP-bound states. Arf activation is mediated by guanine-nucleotide exchange factors (GEFs) that catalyze the exchange of GDP for GTP

Figure 1-4. Overview of steps involved in vesicle formation and fusion. Vesicle formation: First, cargo is concentrated at the donor membrane through recognition by a variety of adaptor proteins. An activated GTPase of the Arf family (Arf, Sar) then recruits coat proteins from the cytosol (COPI, COPII, APs, clathrin) to the donor membrane in order to bind and concentrate cargo (Bi et al., 2002; Owen and Evans, 1998). Polymerization of the coat complex induces membrane deformation, resulting in the budding of a nascent vesicle. Coat polymerization continues until the vesicle has completely budded and has pinched off of the donor membrane, completing vesicle formation. Arf is released after hydrolyzing GTP. Finally, the coat disassembles and is reused in another round of vesicle formation. Vesicle fusion: Fusion begins when the approaching vesicle is tethered to the target membrane by a multisubunit-tethering complex (Hughson and Reinisch, 2010; Jackson et al., 2012). Interaction with tethers and activated Rab GTPases apposes the vesicle and target membranes, allowing the interaction of SNARE proteins on either membrane to interact. Dissociation of SM proteins from syntaxin frees the requisite SNAREs on the target membrane, allowing interaction with SNAREs on the vesicle to form a *trans*-SNARE complex. Formation of this complex overcomes the energetic barrier required for membrane mixing (Nickel et al., 1999; Weber et al., 1998) allowing delivery of the vesicle contents to the target organelle. The action of the AAA ATPase NSF untwists the SNARE complex, priming the membrane for another round of fusion.


Soluble Cargo
Transmembrane cargo
SNARE
GEF
Arf GTPase
GTP
GDP
Coat proteins
Multisubunit Tethering Complex
Rab GTPase
NSF p97
Syntaxin
SNAREs
SM Protein

(D'Souza-Schorey and Chavrier, 2006). Arf-GTP binds membranes through insertion of an N-terminal amphipathic α-helix and a myristate group into the donor membrane (D'Souza-Schorey and Stahl, 1995; Franco et al., 1996). At the ER, the Sar1 protein, a member of the Arf family, also binds the membrane through the insertion of an N-terminal amphipathic α-helix (Lee et al., 2005).

At the membrane, activated Arf/Sar1 recruit subunits of membrane deforming coat complexes and contribute to the stabilization of the forming vesicle (Figure 1-4). Coats are generally recruited in two stages: first, cargo-binding subunits bind signals in the cytosolic portion of membrane proteins, or cargo receptors, and is followed by the recruitment of the outer coat resulting in membrane deformation and budding of the nascent vesicle (Bonifacino and Glick, 2004). At the ER, the Sec23/Sec24 complex binds Sar1 and cargo and is followed by binding of Sec13/Sec31 (Barlowe et al., 1994; Fromme et al., 2007; Shaywitz et al., 1997). All of these components are thought to contribute to membrane deformation in nascent COPII coats (Bi et al., 2002; Lee et al., 2005; Stagg et al., 2006; Stagg et al., 2008). Retrograde transport from the Golgi, as well as intra-Golgi transport is mediated by the COPI complex, whose seven subunits are recruited en-bloc to the Golgi membrane. Similarly, at both the plasma membrane and endolysosomal organelles, adaptor protein (AP) complexes are recruited by Arf GTPases, recognize cargo proteins, and in some cases, recruit clathrin to trigger membrane deformation (Cocucci et al., 2012).

During the vesicle formation process, Sar1/Arf dissociate from vesicle as the result of GTP hydrolysis. Arf family proteins possess low intrinsic GTPase activity

and are activated by GTPase Activating Proteins (GAPs) to stimulate hydrolysis (Cukierman et al., 1995; Kahn and Gilman, 1986). GTP hydrolysis is stimulated by the insertion of an arginine-finger into the active site of the GTPase, as is the case for other Ras GTPases (Scheffzek et al., 1998). ArfGAP proteins provide the necessary arginine residue for Arfs (Cukierman et al., 1995; Ismail et al., 2010), whereas Sec23 carries out this function for Sar1 (Bi et al., 2002; Yoshihisa et al., 1993). However, the Sec13/31 complex also accelerates GTP hydrolysis by Sar1 upon recruitment (Antonny et al., 2001). Finally, the membrane deforming coat is shed, freeing the vesicle to fuse with the target membrane (Figure 1-4).

1.3.2.2 Vesicle fusion

Fusion of vesicles at different target membranes is also dependent on a set of highly paralogous protein families and occurs in multiple steps. First, vesicles are recognized and tethered to the target membrane through the interaction with a multisubunit-tethering complex (MTC; Figure 1-3; Cai et al., 2007, *inter alia*). Different MTCs localize to different organelles to mediate specific transport steps. The Dsl1 (Dependence on SLY1-20) complex recognizes COPI coats at the ER (Andag and Schmitt, 2003), the COG (Conserved Oligomeric Golgi) complex functions within the Golgi (Whyte and Munro, 2001), the GARP (Golgi Associated Retrograde Protein) complex tethers endosome-derived vesicles to the TGN (Conibear and Stevens, 2000), whereas the exocyst complex tethers secretory vesicles to the plasma membrane (TerBush et al., 1996). The HOPS (Homotypic Vacuole Fusion and Protein Sorting) and CORVET (Class C Core Vacuole/Endosome Tethering)

complexes share a core set of subunits, but also possess subunits unique to each complex (Nakamura et al., 1997; Peplowska et al., 2007; Radisky et al., 1997). HOPS and CORVET are responsible for tethering vesicles at the lysosome/vacuole and at endosomes, respectively (Chen and Stevens, 1996; Nakamura et al., 1997; Ostrowicz et al., 2010). The TRAPP (Transport Protein Particle) complex is primarily involved in fusion of vesicles at the Golgi (Sacher et al., 1998; Sacher et al., 2001). Interaction with MTCs and the target membrane is partially mediated by the small GTPase Rab, which, like other protein families involved in membrane trafficking, possess organelle-specific family members (Stenmark, 2009).

Finally, as the MTCs bind the approaching vesicle, SNARE proteins on the vesicle interact with SNAREs on the target membrane by interacting with Rab GTPases (Figure 1-4; for review, see Hong and Lev, 2014). The interaction of multiple SNARE proteins forms a *trans*-SNARE complex which is necessary to overcome the energetic barrier required for membrane fusion (Nickel et al., 1999; Weber et al., 1998). Similar to the MTCs and the Rabs, the interaction between SNAREs is specific; certain SNAREs will only interact with particular sets of other SNAREs to mediate fusion (Sutton et al., 1998). Following cargo delivery, the SNARE complex is disassembled, by the action of the ATPase NSF and its partner α-SNAP (Mayer et al., 1996; Rice and Brunger, 1999), priming the SNAREs for another round of vesicle fusion (Figure 1-4).

1.4 Evolution of the membrane trafficking system

1.4.1 Endosymbiosis as an origin of organelles

We have seen that the LECA was a highly complex organism with multiple cellular pathways and organelles, but the question remains, how did these organelles evolve? Mitochondria and chloroplasts are the result of endosymbiotic events between the proto-eukaryote and an α-proteobacterium and cyanobacterium, respectively (Gray and Doolittle, 1982, *inter alia*). However, the origin of the compartments involved in membrane trafficking is less clear. Endosymbiosis has been proposed as a possible mechanism giving rise to the membrane trafficking system (Martin and Müller, 1998); the lack of a double membrane and organellar genomes, traits characteristic of endosymbiotically derived organelles, argues against such an origin (Gray and Doolittle, 1982), leaving open the question of how and when the eukaryote membrane trafficking system evolved.

1.4.2 Autogenous origins of organelles

Many theories that attempt to explain the evolution of the membrane trafficking system have been proposed, often associating the earliest endomembrane compartments with origins of the eukaryotic cell itself. One of the earliest hypotheses concerning the evolution of eukaryotes was the Archezoa hypothesis. As mentioned earlier, the Archezoa hypothesis proposed that 'primitive eukaryotes' (*Giardia, Trichomonas,* and Microsporidia), called Archezoa, represent the transition state between 'higher eukaryotes' (animals, plants, fungi) and an early

nucleus-containing eukaryotic ancestor (Cavalier-Smith, 1987b). Archezoans were characterized primarily by the absence of detectable mitochondria, Golgi apparatus, and peroxisomes. Although not explicitly stated, it was presumed that the membrane trafficking system evolved autogenously prior to the endocytic event that gave rise to the mitochondrion. However, with the fall of the Archezoa hypothesis, a new class of hypotheses that linked eukaryogenesis with appearance of the membrane trafficking system were proposed.

1.4.3 Fusion hypotheses and the origin of the endomembrane system

The fall of the Archezoa hypothesis suggested that features thought to be present only in higher eukaryotes are in fact present in diverse eukaryotic lineages, and therefore arose prior to the LECA. Detailed microscopic analyses identified stacked Golgi complexes in the majority of eukaryotes, with some sporadic lineages, including some Archezoans, possessing unstacked Golgi complexes (Mowbrey and Dacks, 2009). The pervasive nature of stacked Golgi suggests that the ancestor of eukaryotes also possessed a stacked Golgi, implying that the divergent Golgi structures observed in these putatively ancient eukaryotes are secondarily derived.

Hypotheses for the origin of the membrane trafficking system that followed continued to equate the formation of the trafficking system with eukaryogenesis. In these hypotheses eukaryotes are the product of the fusion of an archaeon with a bacterium to produce a chimeric cell, giving rise to eukaryotes. In these scenarios, the bacterium would have acquired the ability to undergo phagocytosis, taking up an archaeon, which would then give rise to the eukaryotic nucleus (Forterre, 2011;

Gupta and Golding, 1996). These hypotheses were largely based on the observation that eukaryotic informational genes (*i.e.,* transcription, translation, replication) are generally of archaeal origin, whereas membrane phospholipids and metabolic genes are generally derived from bacteria (Gribaldo et al., 2010; Poole and Penny, 2007). These hypotheses account for an origin of the endoplasmic reticulum, as a by-product of engulfing the archaeon, forming the nucleus through endosymbiosis. They then require that the membrane trafficking system evolve beyond the ER through some unstated mechanism.

One recent, and more plausible fusion hypothesis postulates the origin of eukaryotes as the result of the fusion of a bacterium from the PVC (Planctomycete – Verrucomicrobia – Chlamydiae) superphylum with an archaeum from the phylum Thaumarchaeota (Forterre, 2011). The combination of these two lineages would account for the major genetic contributions that make up the eukaryotic genome. Additionally, the PVC bacteria possess proteins structurally similar to the β-propeller-α-solenoid proteins pervasive throughout the eukaryotic membrane deformation machinery (Santarella-Mellwig et al., 2010), providing a hypothesis for the origin of the eukaryotic coat proteins and nuclear pore complex. Additionally, these bacteria also possess an intracytoplasmic membrane (ICM) that surrounds the bacterial nucleoid (Fuerst and Webb, 1991). The thaumarchaeon would have provided eukaryotic features such as eukaryote-like histones and members of the ESCRT complex (Cubonova et al., 2005; Makarova et al., 2010).

In this scenario, viruses provided the pressure driving eukaryotes towards complexity and to produce proteins and processes not found in the other two

domains of life. New proteins and protein folds may have been introduced into the genomes of early eukaryotes through the integration of existing viral genomes. Forterre argues that the co-option of viral proteins by early eukaryotes may explain the origin of major eukaryotic structures such as the nucleus. Some viruses, such as Mimiviruses, recruit perinuclear organelles to build large viral factories (Novoa et al., 2005; Suzan-Monti et al., 2007, *inter alia*). Forterre imagines a scenario where the PVC bacterium uses the viral factory machinery provided by integrated viruses to build a nucleus from the ICM to protect its genome from further viral invasion (Forterre, 2011).

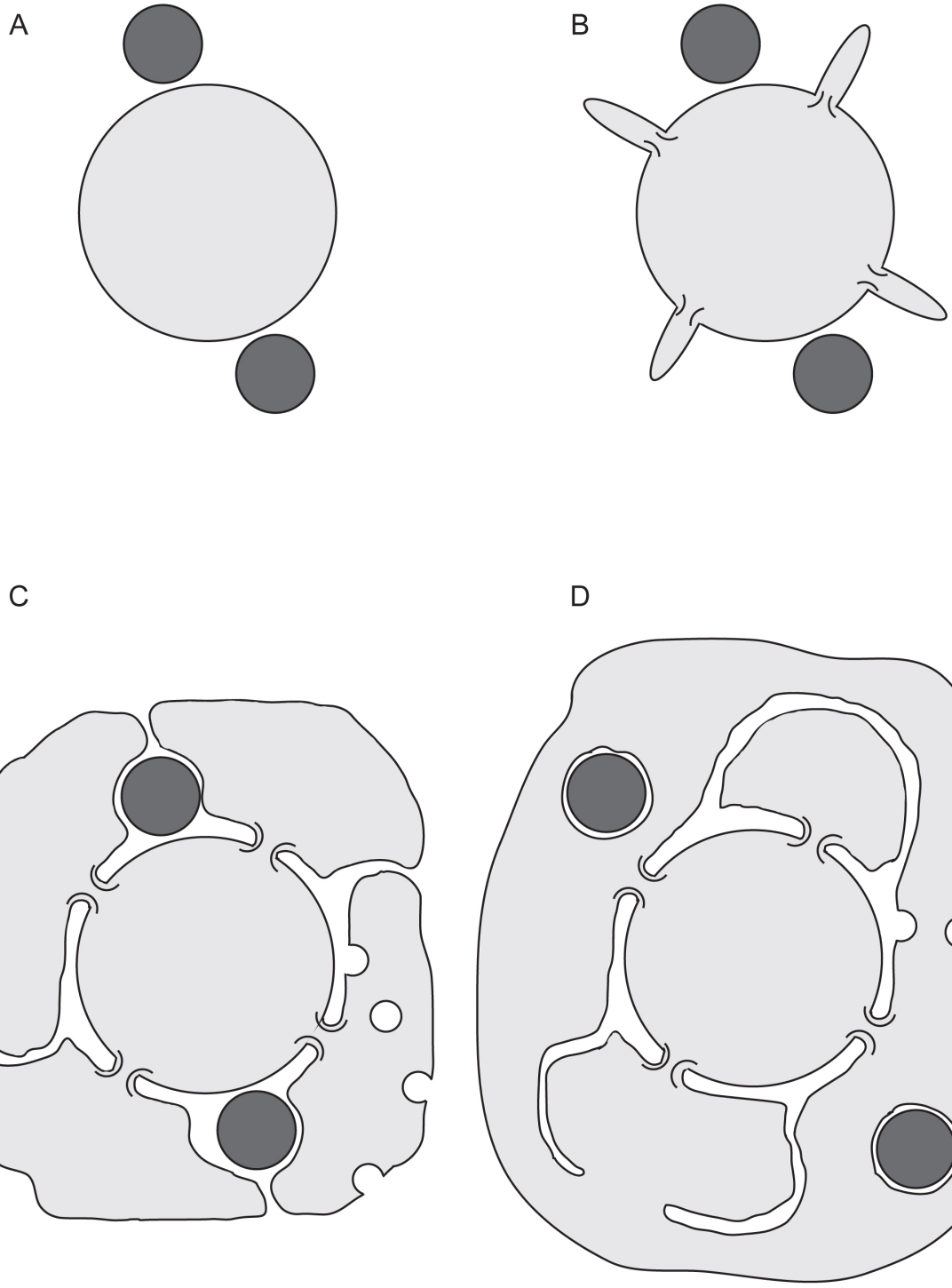
Although this hypothesis is intriguing to consider, major flaws have been pointed out that apply to this hypothesis and to fusion hypotheses generally. The first concerns the apparent loss of the archaeal membrane that is characterized by the presence of isoprenoid lipids (Langworthy and Pond, 1986). Secondly, of the known endosymbiotically derived organelles, loss of the endosymbiont membrane has not occurred, raising the question of what happened to the archaeal-derived lipids (Forterre, 2011). Third, is that ribosomal phylogenies reconstruct three domains: bacteria, archaea, and eukaryotes (Woese and Fox, 1977). Forterre himself stated that it would be difficult to explain why the ribosomal subunit from the archaeal symbiont was retained over that of the bacterial host, and why the new eukaryotic ribosome underwent accelerated sequence evolution from the archaeal version as to be recognized as a different domain of life (Forterre, 2011).

1.4.4 Autogenous origins of the membrane trafficking system

A more recent take on the evolution of the eukaryotic cell and by extension, the membrane trafficking system removes the phagocytic event from eukaryogenesis, and instead postulates an autogenous origin (*i.e.,* not endosymbiotic in origin) of the nucleus. Called the 'inside-out' hypothesis (Baum and Baum, 2014), the authors propose that membrane deforming complexes related to those found in the nuclear pore complex and coat proteins, induced membrane deformation at the plasma membrane, resulting in outward membrane blebbing (Figure 1-5A, B). In this hypothesis, the outward growth gave rise to different regions of the eukaryotic cytoplasm, with the original sites of deformation eventually giving rise to nuclear pores (Figure 1-5C). The outward membrane blebs would have greatly increased the surface area of the cell, allowing increased contact and association with ectosymbiotic bacteria (*i.e.,* symbiotic bacteria that adhere to the surface of their symbiotic partners) that would eventually give rise to the mitochondria by being surrounded by the growing membrane blebs, incorporating it into the host cell. The space between the fusing blebs would give rise to the ER and other compartments of the membrane trafficking system, with the separation of the inter-bleb space from the outside world mediated by dynamin acting at plasma membrane, resulting in the formation of the ER (Figure 1-5D).

Although an interesting thought experiment, it is somewhat unwieldy, as it begins with an archaeon possessing isoprenoid lipids, and invokes a shift to bacterial-like lipid membrane contributed by the ectosymbiont. This model would

Figure 1-5. Inside-out hypothesis for the origin of eukaryotes and of the membrane trafficking system. A) An archaeon (light circle) interacts with ectosymbiotic bacteria (dark circles). B) Membrane blebs form as the result of membrane deformation by an early protocoatomer complex (curved lines inside blebs). These protrusions would have facilitated the exchange of metabolites between the archaeon and the bacterium. C) Expansion of the membrane blebs would have enclosed the ectosymbiont, giving rise to the mitochondrion. The ancestral protocoatomer complex gives rise to a full nuclear pore complex (semi-circles). D) Fusion of the blebs would have created a continuous plasma membrane, isolated the primordial mitochondria from the outside world, and would have generated precursors to the organelles of the membrane trafficking system.


A
B
C
D

also require that the membrane deformation machinery is either pre-adapted, or undergoes a shift from binding archaeal membranes to bacterial lipids.

1.4.5 Organelle Paralogy Hypothesis

One recent proposal for the evolution and diversification of autogenously derived organelles of the membrane trafficking system is the Organelle Paralogy Hypothesis (OPH; Dacks and Field, 2007). The OPH is based on two primary observations: first, as above, that the majority of protein families involved in membrane trafficking [GTPases, SNAREs, syntaxins, SM (Sec1/Munc18-like) proteins, coats] were present in the LECA, indicating that they evolved in an even earlier eukaryotic ancestor (Dacks and Doolittle, 2002; Dacks and Doolittle, 2004; Devos et al., 2004; Field et al., 2007b; Jékely, 2003; Koumandou et al., 2007).

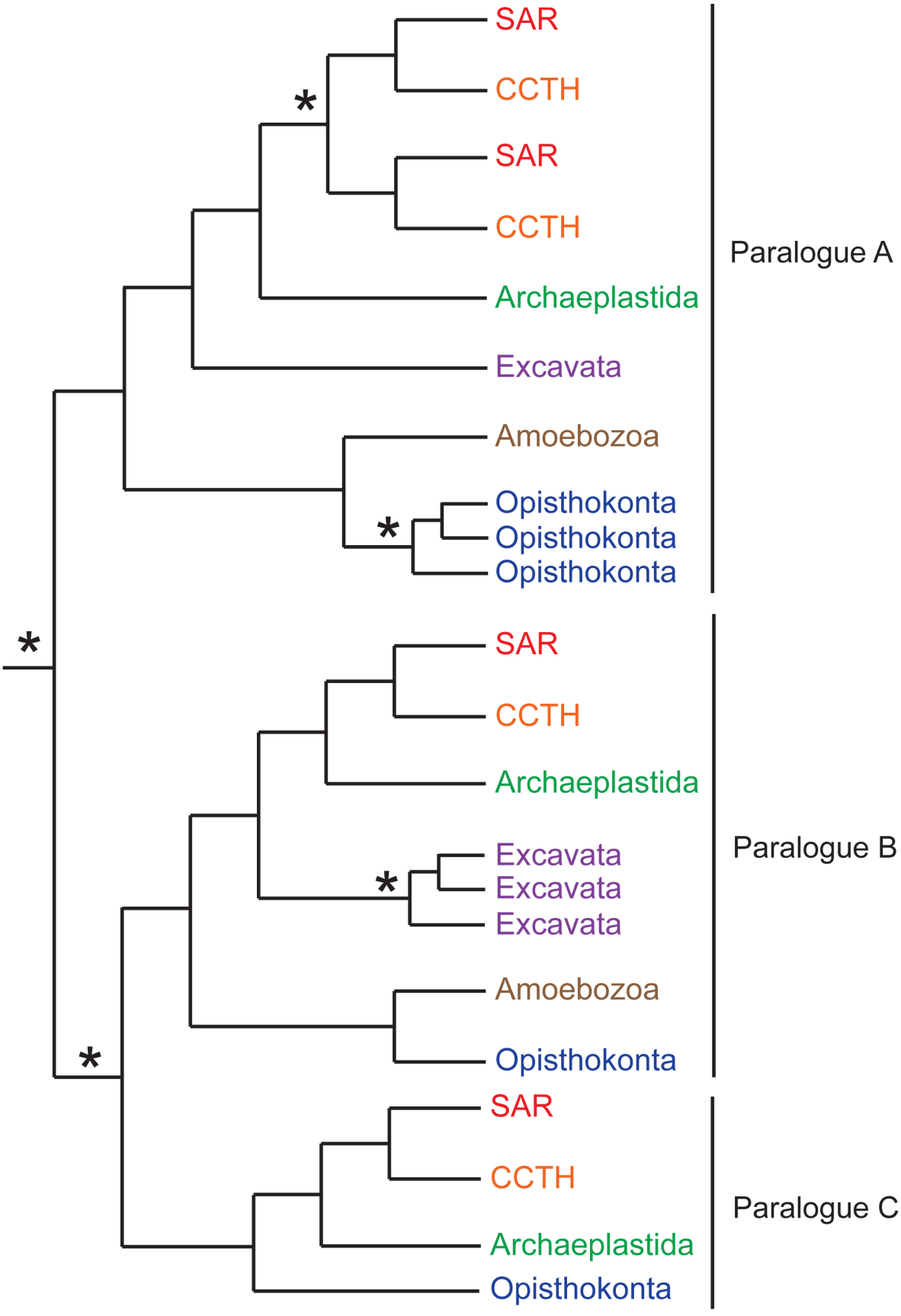
Second, these protein families contain multiple members that carry out similar functions at different subcellular locations (*i.e.,* paralogues; Bonifacino and Glick, 2004). Based on sequence similarity, and in some cases structural similarity, it is determined that different proteins share a common ancestor, *i.e.,* are homologous (Fitch, 1970). Homologues can be of two types: paralogues, which arise through gene-duplication events and orthologues, which arise through speciation events (Fitch, 1970). Evidence points to the presence of different paralogues from each family in the LECA (Dacks and Doolittle, 2002; Dacks and Doolittle, 2004; Elias et al., 2012; Gabernet-Castello et al., 2013; Hirst et al., 2011; Vedovato et al., 2009). If we assume that the function of each paralogue has been conserved in extant

eukaryotes, we can then infer that the LECA possessed a complex membrane trafficking system, similar to that observed in living eukaryotes. In phylogenetic analyses, these ancient paralogues should assemble into groups containing the full range of eukaryotic taxa, indicating that they descended from a single ancestor (*e.g.,* paralogues A, B, and C from Figure 1-6), and by extension, the organelles at which these paralogues act should also have been present in the LECA. By contrast, paralogues resulting from lineage-specific expansions should assemble into a clade only containing sequences from that lineage (Figure 1-6). Loss of an ancient paralogue can only be inferred if a clade contains a broad, but incomplete array of supergroups representing eukaryote diversity (Figure 1-6).

Ancient duplication and lineage-specific expansion were shown to be major drivers of the evolution of membrane trafficking, where phylogenetic analysis of SynE (endosomal syntaxins), Rab5, and the shared β-subunit of AP-1 and AP-2 identified sequences that grouped by taxonomic lineage, revealing multiple independent expansions within each of these protein subfamilies (Dacks et al., 2008). Phylogenies of SM proteins (Koumandou et al., 2007), and more recently analyses of the TBC (RabGAPs; Gabernet-Castello et al., 2013) and the Rab families (Elias et al., 2012), revealed sequences grouping by subfamily rather than by taxonomic lineage, indicating that they are the product of multiple ancient gene duplications that occurred prior to the LECA.

The OPH predicts that, since these gene families arose by gene duplication, sequence divergence, and co-evolution, an ancestral organelle should have existed that possessed the ancestor of each protein family (*e.g.,* the ancestral coat complex,

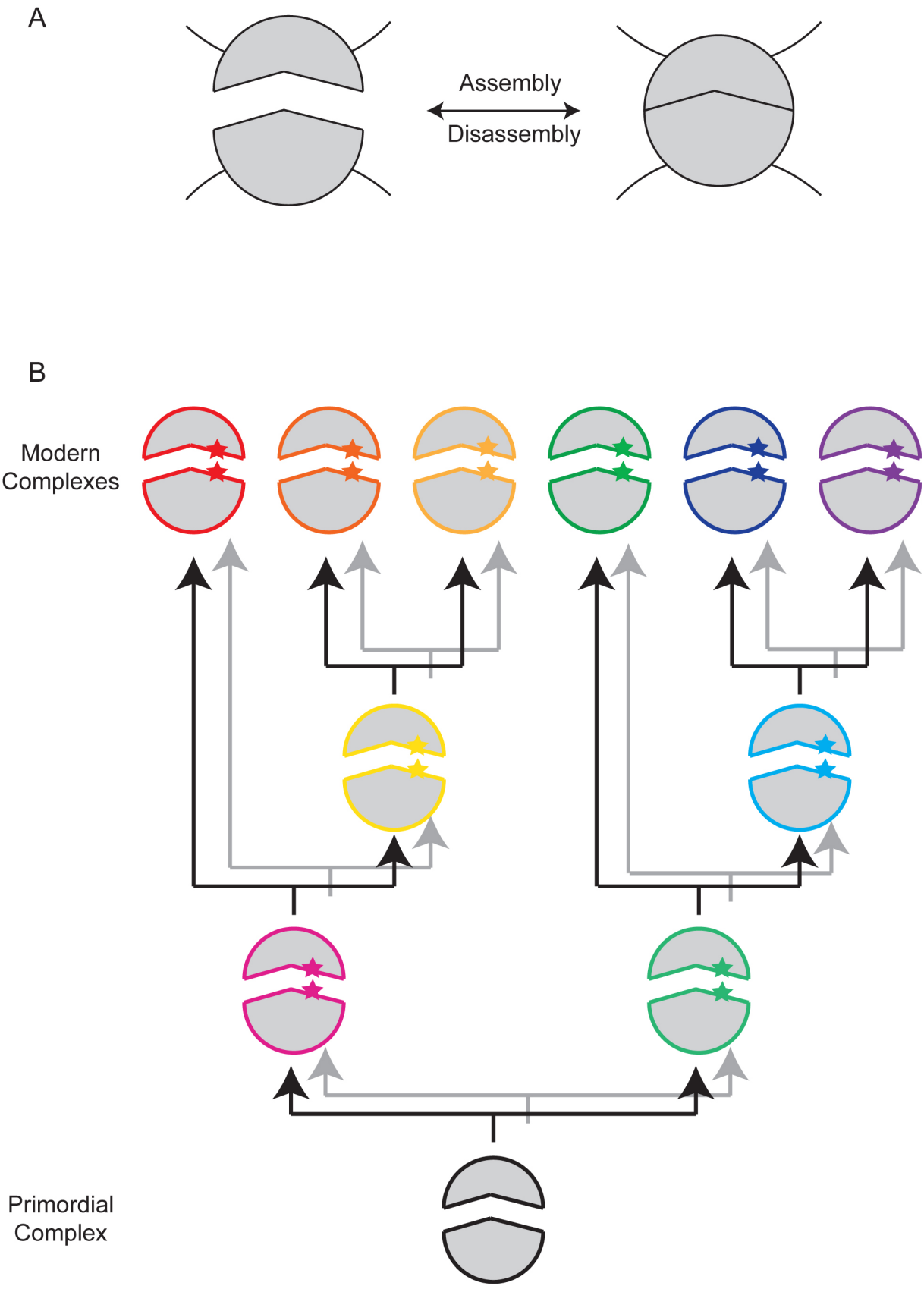
Figure 1-6. Example of a gene family with ancient and lineage-specific gene duplications. Hypothetical gene family with ancient and lineage-specific gene duplications. Three ancient eukaryotic paralogues (A, B, C) of this family are depicted here. They are considered ancient because all three are found in multiple, diverse eukaryotic lineages. They are considered paralogues because representatives of nearly every supergroup are found within each clade, indicating these groups are the result of an ancient gene duplication event (left-most star). Lineage-specific gene duplications have occurred for the SAR+CCTH clade and opisthokonts of paralogue A. These are viewed as independent duplications and not loss from Amoebozoa, Archaeplastida, and Excavata because they group separately from each other. It is also more parsimonious than the alternative, three independent losses in Amoebozoa, Excavata, and Archaeplastida. If the opisthokont sequences grouped with each of the SAR+CCTH clades, then loss in the other supergroups would be a more likely scenario because the clade contains multiple supergroups from across the eukaryotic tree. Lineage-specific expansion of paralogue B has also occurred in the Excavata . Paralogue C would be interpreted as having undergone loss in the Excavata and Amoebozoa because the rest of the diversity of eukaryotes are represented. Asterisks (*) indicate gene duplication events.


SAR
CCTH
SAR
CCTH
Archaeplastida
Excavata
Amoebozoa
Opisthokonta
Opisthokonta
Opisthokonta
Paralogue A
SAR
CCTH
Archaeplastida
Excavata
Excavata
Excavata
Amoebozoa
Opisthokonta
Paralogue B
SAR
CCTH
Archaeplastida
Opisthokonta
Paralogue C
*
*
*
*

the ancestor of all Rabs, the ancestor of all SNAREs, *etc*.) in the proto-eukaryote (Figure 1-7). Novel organelles would then arise via gene duplication of the ancestral homologues. Sequence divergence of the newly evolved paralogues would result in multiple subpopulations that are only able to interact with paralogues of other protein families that have co-evolved with those paralogues, resulting in exclusionary interactions between sets of proteins. These exclusionary interactions would define different membrane-bound compartments, producing novel organelles. Iterations of gene duplication and sequence divergence would eventually produce a large array of membrane bound organelles differentiated from one another by specific combinations of paralogues of trafficking factors (Figure 1-4). This hypothesis has been supported by computer simulations; a simple system consisting of a coat, a SNARE, and a single membrane-bound compartment could give rise to a complex endomembrane system through gene duplication and divergence (Ramadas and Thattai, 2013). While gene duplication-divergence was sufficient to generate novel organelles in this simulation, increasing specificity between interactions of coats and SNAREs across the system was required to maintain a large number of distinct organelles.

Conceivably, elucidating the order in which organelle-specific paralogues of different membrane trafficking families emerged would allow the deduction of the order of organelle evolution (Dacks and Field, 2007). One major limitation of the OPH is that most of the members of protein families encoding specificity in the membrane trafficking system are made up of relatively short sequences. Combined

Figure 1-7. The Organelle Paralogy Hypothesis for the evolution of autogenously derived organelles. A) Assembly and disassembly of a hypothetical protein complex, or set of interacting proteins, composed of two distinct members (upper and lower). B) Evolution of the single primordial complex/interacting proteins into multiple complexes with different cellular locations. The primordial complex (black) undergoes a gene duplication event, represented by the bifurcating arrows. The Black arrow represents the evolution of the upper subunit. The grey arrow represents the evolution of the lower subunit. The primary sequence of the subunits acquired mutations resulting in two types of complexes, pink and light green. Some mutations, represented by stars, fix specific interactions such as the pink upper subunit with the pink lower subunit or the light green upper subunit with the light green lower subunit, resulting in two distinct subpopulations of complexes. These subpopulations are only able to interact with members of the same subpopulation, but that are unable to interact with members of the other subpopulation (*i.e.,* pink only interacts with pink, but not with light green) If these complexes associate with membrane bound organelles, then, by extension, the evolution of these two distinct complexes would also produce two novel organelles. Iterations of this process would result in a large number of distinct, but evolutionarily related membrane bound compartments whose evolutionary histories can be traced by elucidating the evolutionary relationships between the subunits of the various complexes.


A
Assembly
Disassembly
B
Modern
Complexes
Primordial
Complex

with large numbers of paralogues, this renders obtaining phylogenetic resolution problematic. Nonetheless, recent developments in phylogenetic approaches have begun to resolve some of the earliest events in the evolution of the membrane trafficking system. Analyses using a novel phylogenetic approach, called Scrollsaw, have been able to determine the order of gene duplications that occurred to produce the set of Rab and TBC paralogues that were present in the LECA (Elias et al., 2012; Gabernet-Castello et al., 2013). Thus, these gene duplications would presumably have occurred in an earlier eukaryotic ancestor than the LECA. In particular, analysis of the Rab GTPases identified 19-23 Rab subfamilies in the LECA (Diekmann et al., 2011; Elias et al., 2012), which correspond broadly to two large groups containing endocytic and exocytic functions, respectively (Elias et al., 2012). Additionally, phylogenetic analysis stringing together multiple genes through concatenation of the subunits of COPI and the adaptin complexes resolved the internal branching order of these coats (Hirst et al., 2011). An earlier analysis had suggested that the first bifurcation, which produced COPI and the adaptin clade, would have correlated with the evolution of a Golgi and a TGN-like organelle, effectively bridging the endocytic and secretory systems (Dacks et al., 2008). Hirst et al*.,* (2011) took this further and suggested that the TGN-like organelle was a TGN/endosome hybrid, based on the observation that AP-3 and AP-5, the two earliest branching AP complexes, are both involved endocytic trafficking (Hirst et al., 2011; Peden et al., 2002). AP-4 branches after AP-3 and AP-5 and is involved in trafficking from the TGN (Dell'Angelica et al., 1999a; Hirst et al., 2011). Hirst et al.,

(2011) also suggested that the evolution of the AP-4 complex coincided with the evolution of distinct TGN and endosomes.

Thus far, analyses testing the OPH have produced largely consistent results. With the exception of the analysis of COPI and the AP complexes, analyses testing the OPH have focused primarily on protein families involved in membrane fusion (Dacks and Doolittle, 2002; Dacks and Doolittle, 2004; Dacks et al., 2008; Elias et al., 2012; Field et al., 2007b; Gabernet-Castello et al., 2013; Koumandou et al., 2007; Sanderfoot, 2007; Vedovato et al., 2009); it remains uncertain whether machinery involved in vesicle formation display the same pattern. Analysis of proteins involved in this step of membrane trafficking should provide further insight into the origin and evolution of the trafficking system.

1.4.6 Protocoatomer hypothesis

The OPH itself does not propose an origin for the first endomembrane compartment, but rather a mechanism for organelle diversification. However, a specific example of the OPH, the protocoatomer hypothesis, posits a common origin for the nuclear pore complex (NPC) and the various vesicle coats and is based on the observation of shared structural elements, namely β-propeller/α-solenoid domain-containing proteins (Devos et al., 2004). Structural analysis of the *S. cerevisiae* Nup84 subcomplex revealed that each of its components, Seh1, Sec13, Nup84, Nup85, Nup120, Nup133, and Nup145C consist of an α-solenoid, a β-propeller, or both (Berke et al., 2004; Boehmer et al., 2008; Brohawn et al., 2008; Devos et al., 2004; Fath et al., 2007; Hsia et al., 2007; Leksa et al., 2009; Nagy et al., 2009).

Similarly, elements of each major vesicle coat complex possess the same architecture: the trunk domains of the large subunits of COPI and the adaptin complexes are α-solenoids (Hoffman et al., 2003; Owen et al., 1999; Traub et al., 1999; Watson et al., 2004), whereas clathrin and α/β'-COPI, the membrane deformation complexes, are composed of one or two β-propellers, followed by an α-solenoid domain, respectively (Devos et al., 2004; Lee and Goldberg, 2010; ter Haar et al., 1998). The COPII complex also shares this organization, with Sec31 comprising the β-propeller/α-solenoid configuration (Fath et al., 2007). Sec13, a subunit of the Nup84 complex, is also a member of the COPII coat complex. Sec13 binds Sec31 and Nup145C through a conserved mechanism, involving the insertion of a seventh blade, completing the propeller structure of Sec13 (Brohawn and Schwartz, 2009), and is consistent with reports of common ancestry between Sec31 and Nup145 as well as Nup84 and Nup85 (Brohawn et al., 2008). Other cellular complexes have also recently been shown to possess this 'protocoatomer' type structure, such as the intraflagellar transport system (IFT; van Dam et al., 2013), and the recently described SEA (Seh1-associated) complex, a vacuole associated complex that modulates TOR (Target of Rapamycin) signalling (Algret et al., 2014; Dokudovskaya et al., 2011), and that shares Seh1 with the nuclear pore complex, and Sec13 with COPII and the NPC. Additionally, the SEA complex possesses subunits thought to be structurally similar to Sec31, providing another link between this novel complex, vesicle coats, and the nuclear pore (Algret et al., 2014). Subunits of the HOPS/CORVET MTCs such as Vps3, Vps8, Vps11, Vps16, Vps18, Vps33, Vps39, and Vps41 are also predicted to contain the β-propeller/α-solenoid domain

composition (Plemel et al., 2011). However, crystal structures have yet to be determined for these proteins. Some may argue that these structural similarities may be the result of convergent evolution. If this were the case, we would not expect that all of the β-propeller/α-solenoid domain-containing proteins would possess these structural elements in the same order; the α-solenoid should precede the β-propeller in some cases. We would also not expect subunits to be shared among the different complexes, but rather, multiple proteins with analogous functions should have evolved in each complex. Similarly, binding mechanisms such Sec13 with Sec31 or Nup145 are not expected to have occurred if these proteins are of separate evolutionary origins.

1.5 Focus of this thesis

Past analyses have shown that the LECA was a highly complex ancestor, that possessed much of the known membrane trafficking machinery (Koumandou et al., 2013, *inter alia*). However, these analyses focused extensively on the machinery of vesicle fusion such a SNAREs (Dacks and Doolittle, 2002; Dacks and Doolittle, 2004; Dacks et al., 2008), tethering complexes, SM proteins (Koumandou et al., 2011), or Rabs (Diekmann et al., 2011; Elias et al., 2012; Gabernet-Castello et al., 2013; Pereira-Leal and Seabra, 2001). These analyses indicated that much of the membrane fusion machinery observed in extant eukaryotes was also present the LECA. However, the extent to which the machinery involved vesicle formation is conserved is much less clear. These observations raise the question of whether or not machinery involved in vesicle formation are equally well conserved. To address

this question, three separate analyses have been carried out: a comparative analysis of the ArfGAP and ArfGEF proteins, a detailed comparative genomic and phylogenetic analysis of the COPII coat complex, and a comparative genomic and phylogenetic analysis leading to the discovery of a novel ancient coat complex: TSET.

1.5.1 Comparative genomic analysis Arf GTPase regulators

Arfs are small GTPases of the Ras superfamily, and are involved in regulating membrane trafficking, phospholipid biosynthesis, and cytoskeletal remodelling (Brown et al., 1993; Cockcroft et al., 1994; Honda et al., 1999; Ooi et al., 1998; Paleotti et al., 2005). Up to six Arf proteins are found in mammalian cells, but only one is found in most other eukaryotes. Additionally, current evidence points to the presence of a single Arf protein in the LECA (Berriman et al., 2005; Li et al., 2004), suggesting that regulatory proteins may have provided functional diversity and encoded specificity in Arf signalling. As mentioned in section 1.3.2.1, the transition between the active and inactive state of Arf is mediated by GEFs and GAPs, respectively. ArfGEFs are defined by the presence of Sec7 domain and in humans are subdivided into six subfamilies (Cox et al., 2004). ArfGAPs are defined by the presence of the catalytic ArfGAP domain and are divided into ten subfamilies in humans (Kahn et al., 2008). I carried out comparative genomic and phylogenetic analyses in order to determine the extent to which each subfamily is conserved and to identify gene duplications in different eukaryotic lineages. By analyzing the patterns of gene duplications, I have generated hypotheses regarding the

complement of ArfGAPs and ArfGEFs that were present in the LECA. Analysis of accessory domains also provided insight into the functional complexity of ArfGAP and GEF proteins in the LECA and its descendant lineages. Overall, this analysis will fill a hole in our current understanding of the evolution of the Arf regulatory system as it relates to the OPH.

1.5.2 Comparative genomic analysis of the COPII coat complex

The COPII coat complex is responsible for trafficking cargo and membranes from the ER to the Golgi (Barlowe et al., 1994; Fromme et al., 2007; Gershlick et al., 2014; Jones et al., 2003; Miller et al., 2003; Roberg et al., 1999; Venditti et al., 2012; Wendeler et al., 2007). A previous comparative genomic analysis assessing the presence or absence of the nuclear pore and related complexes identified components of the COPII coat in all major eukaryotic lineages, suggesting their presence in the LECA (Neumann et al., 2010). Although core members of the coat (Sar1, Sec23, Sec24, Sec13, and Sec31) are ubiquitously conserved, both Sec16 and the *S. cerevisiae* Sec24-like protein, Sfb3, were found to be frequently missing. In order to determine the extent to which these two coat components are absent, I carried out an extended comparative genomic analysis. I increased the sampling of non-opisthokont representatives and included recently sequenced, key taxonomic sampling points to determine if these patchy distributions are the result of multiple lineage-specific losses or are the result of taxon selection. The Sar1 GEF Sec12 was excluded from the analysis by Neumann et al., (2010) and therefore I included it here. Extensive phylogenetic analysis of each component identified lineage-specific

expansions and determined the number of paralogues of each subunit present in the LECA. This was be especially exciting for the phylogeny of Sec24, where multiple ancient paralogues have been proposed, but never conclusively demonstrated using phylogenetic methods (Pagano et al., 1999; Tang et al., 1999).

1.5.3 Comparative genomic analysis of the TSET complex

Chapter 5 analyzes the evolution of the recently discovered TSET coat complex (Gadeyne et al., 2014). TSET was identified as an archaeplastid-specific coat complex (Gadeyne et al., 2014). However, our collaborators identified putative TSET subunits in representatives of two other supergroups, suggesting that this novel coat complex may be more broadly distributed than previously thought. Therefore, I carried out a comparative genomic analysis to determine the distribution of the TSET complex among a set of representative eukaryotic genomes. The TSET complex is thought to possess a high degree of sequence similarity to the heterotetrameric coat complexes (APs and F-COPI; Gadeyne et al., 2014). To determine the relationship of TSET, the APs, and COPI, phylogenetic analysis of these coat complexes was carried out. Understanding the relationship between TSET, the APs, and COPI will provide a more detailed understanding of the order in which their associated organelles evolved.

Chapter 2: Materials and Methods

2.1 Overview

A combination of computational methods was used to identify and analyze the evolution of genes involved in membrane trafficking. The specific tools, criteria, and their implementation apply equally to all analyses discussed hereafter, unless otherwise indicated. Modifications of these criteria (*e.g.,* changes in phylogenetic methodologies and approaches to comparative genomic analyses) and the rationale for those modifications will be addressed in their respective chapters.

Comparative genomics is the process of identifying homologues in the genomes of different organisms. Comparative genomics is a powerful approach for identifying and analyzing biological pathways in diverse organisms by making use of publicly available genome sequencing projects and advanced computational methodologies. The main advantage of these approaches is that they do not depend on the availability of genetically tractable systems in diverse eukaryotic lineages to be available for the study of a cell biological system. Nonetheless, there are two major limitations for this approach. First, is the sensitivity and selectivity of the methodologies used to identify homologous sequences. In the past, analyses have relied solely on sequence-sequence comparisons (*e.g.,* FASTA, BLAST; Altschul et al., 1990; Pearson and Lipman, 1988), which are informative so long as they share sufficient similarity as to be recognized by the algorithm. The advent of profile-based searches (*e.g.,* PSI-BLAST, HMMER; Altschul et al., 1997; Eddy, 1998; http://hmmer.org) that incorporate information from multiple sequences into a statistical model, have dramatically increased the power of comparative genomic approaches in their ability to detect distant homologues. The second limitation is

the availability of sequenced genomes and associated genomic databases. Compared to the known diversity of eukaryotes, a relatively small number of fully sequenced genomes are available for analysis. However, the available databases are broadly representative of eukaryote diversity, allowing a proper sampling of major eukaryotic lineages. Additionally, the number of fully sequenced genomes representing diverse lineages is continuing to increase, permitting a deeper sampling of these lineages.

Phylogenetic analysis reconstructs the evolutionary relationships of homologous genes and is able to discriminate between orthologues and paralogues. Importantly, these analyses are able to discriminate between different types of paralogues, such as ancient paralogues that predate the LECA, paralogues found within a specific taxonomic range (*i.e.,* supergroup or genus), or genome-specific paralogues. A clear picture of when these gene duplications occurred allows us to hypothesize when particular functions or pathways evolved, clarifies which aspects of cellular function are homologous between organisms, and which pathways may have arisen through convergent evolution. The results of these methods can be used as hypotheses to be tested in genetically tractable systems, allowing *in silico* methods to inform functional analyses just as functional analyses are able to provide queries for comparative genomics.

2.2 Comparative Genomics

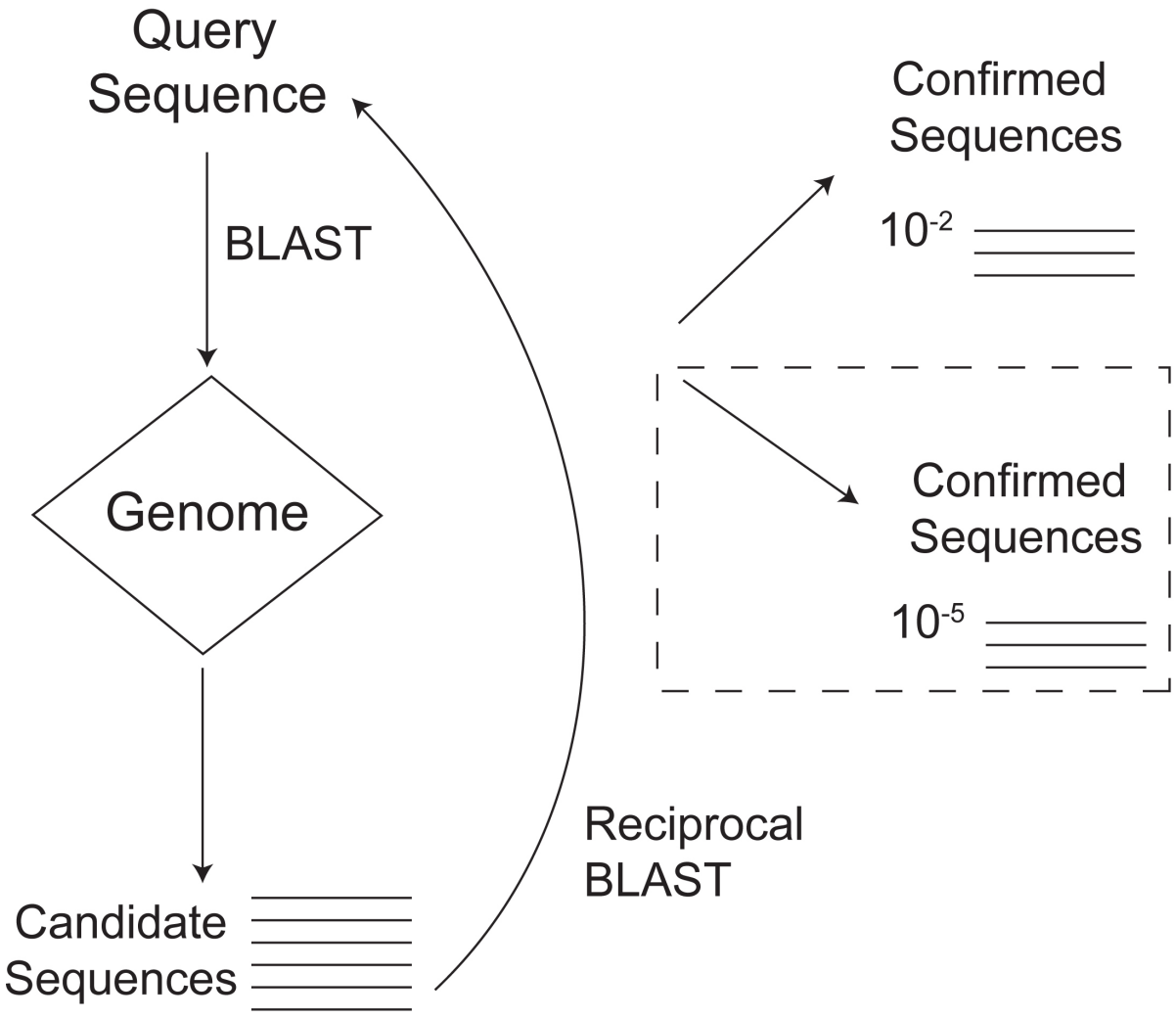
Comparative genomic analyses were carried out to identify homologues of selected genes in representative sets of eukaryotic genomes. The most

representative set of eukaryotic genomes available at the time each study was conducted were selected for analysis. The genomic databases sampled are displayed in Figures 3-3, 3-4, 4-2, and 5-1.

2.2.1 BLAST

Basic Local Alignment Search Tool (BLAST; Altschul et al., 1997) was used to identify protein homologues in selected genomes. Amino acid sequences were used for analyses as 20 states are possible for each position compared to four states for nucleic acids. This reduces the likelihood that codon bias or differences in GC (guanine-cytosine) content will impact BLAST results (Steel et al., 1993). As our understanding of membrane trafficking stems largely from studies of humans and yeast, each analysis addressed whether components of the trafficking systems of those organisms are broadly conserved or not. Therefore, protein sequences from *H. sapiens* or *S. cerevisiae* were used as queries in BLASTp searches (*i.e.,* protein query against a protein database; Figure 2-1). Forward BLAST searches (*i.e.,* into genomes of interest) generated a list of candidate sequences that required confirmation of homology. Sequences with E-values less than or equal to 0.05 were retained for verification. This cut-off was used to remove false positive sequences that were retrieved due to the presence of shared accessory domains or because of random sequence similarity. Homology was assessed using the reciprocal best-hit method (Bork et al., 1998; Tatusov, 1997), whereby the candidate sequences must retrieve the original query with an E-value at least two orders of magnitude smaller than the next best hit (Figure 2-1). In the event that no homologues were confidently

Figure 2-1. The comparative genomic workflow. Query sequences are used to search for homologues in target genomes. 'Forward BLAST' experiments produce a list of candidate sequences that contain some similarity to the original query sequence. Homology is confirmed if upon 'reciprocal BLAST' against the genome of the original query, candidate sequences retrieve the query as the best BLAST-hit with an E-value at least two orders of magnitude better than the next best sequence. Similar criteria were implemented for searches using HMMs. An additional level of stringency required candidate sequences to retrieve the query with an E-value five orders of magnitude better than the next best hit was also considered for the analysis of ArfGAP and ArfGEF proteins (see chapter 3).

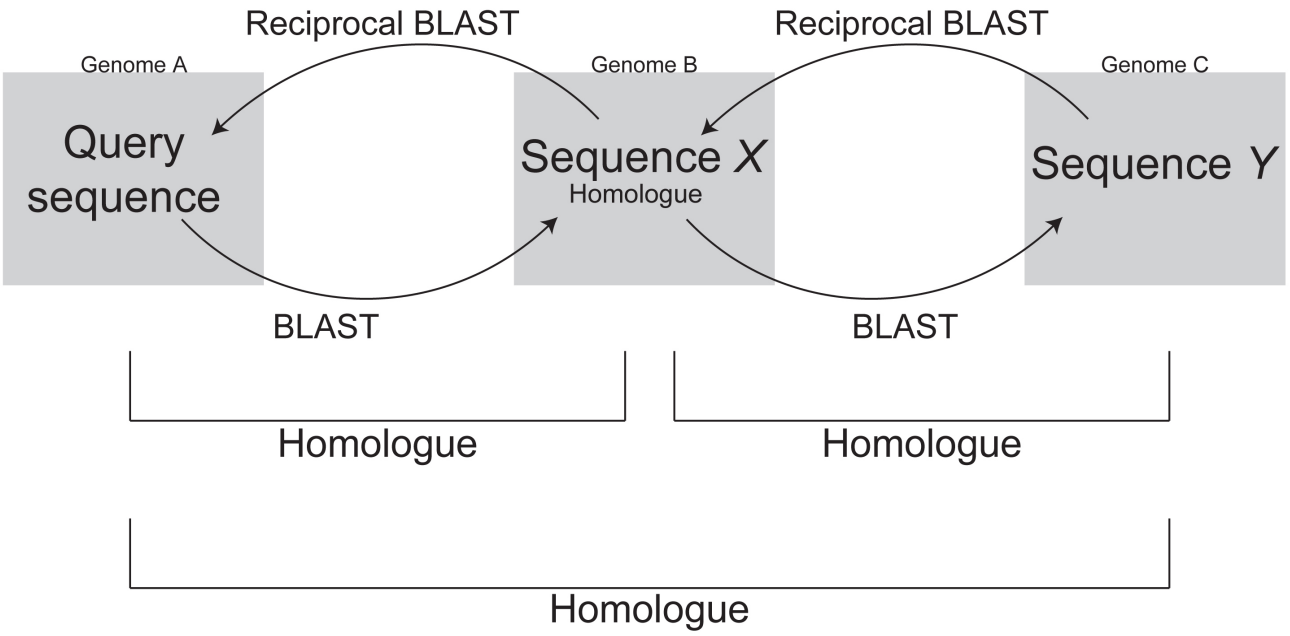

Query
Sequence
BLAST
Genome
Candidate
Sequences
Reciprocal
BLAST
Confirmed
Sequences
10-2
Confirmed
Sequences
10-5

identified using the human or yeast queries, then 'taxon-jumping' was carried out. In this approach, if no homologues were identified in a particular genome, then the homologues of a closely related organism were used as queries (Figure 2-2). This approach is based on the assumption that if a sequence, *X,* is considered homologous to the *H. sapiens* or *S. cerevisiae* query sequence, then if another sequence, *Y*, is found to be homologous to *X*, then by extension it is also homologous to the original query (Figure 2-2).

2.2.2 HMMER

Although most sequences were identified using BLAST, sequence-sequence comparison is not always able to detect extremely divergent homologues. Protein databases were searched using Hidden Markov Models (HMM) implemented through the HMMER suite (http://hmmer.org) to identify homologues missed due to sequence divergence. HMMER take a multiple sequence alignment and generate a statistical model (HMM) of the alignment that describes position-specific information about the conservation of each column in the alignment and the probability of finding specific amino acids at specific positions (Eddy, 1996; Eddy, 1998; Krogh et al., 1994). The model is then used to search sequence databases for sequences that fit the model (Eddy, 2009; Eddy, 2011). Multiple sequence alignments were generated using MUSCLE (Multiple Sequence Comparison by Log-Expectation; Edgar, 2004) by aligning homologues identified by BLAST. HMMs were generated using the HMMbuild program, and used to search protein databases with the HMMsearch program. Candidate sequences identified by HMMsearch with E-values below HMMER's inclusion threshold (0.01), which indicates that the

Figure 2-2. Illustration of taxon jumping. The query sequence from Genome A identified and confirmed Sequence *X* from Genome B as a homologue of the query through forward and reciprocal BLAST experiments. The confirmed homologue can then be used to confirm homology of divergent sequences, or identify additional sequences missed by the original query (Sequence *Y* from Genome C). Successful reciprocal BLAST of Sequence *Y* against Genome B can then confirm the homology of Sequence *Y* to Sequence *X*, and by extension, the original query.


Reciprocal BLAST
Reciprocal BLAST
Genome A
Genome B
Genome C
Query sequence
Sequence X
Homologue
Sequence Y
BLAST
BLAST
Homologue
Homologue
Homologue

sequence's fit to the model is not likely to be the result of random similarity, were retained for reciprocal confirmation, maintaining the 2-orders criterion using either BLASTp or phmmer. phmmer is a BLAST-like program implemented in the HMMER package and was used for reciprocal analyses for the analysis of ArfGEFs (chapter 3) and COPII (chapter 4). The HMMER package uses a different approach to calculating E-values than BLAST (Eddy, 2008; Karlin and Altschul, 1990; Karlin and Altschul, 1993), therefore, using phmmer for the reciprocal analyses maintained consistency and comparability between forward and reverse searches. phmmer was not implemented in the other analyses as it was not available at the time they were performed.

2.2.3 Nucleotide searches

If neither BLAST nor HMMer succeeded in identifying homologues, genomic nucleotide databases (*e.g.,* contigs, scaffolds, *etc.*) were searched for the missing sequences in the event that poor gene models or missing open reading frames resulted in false negatives. An algorithm entitled tBLASTn (*i.e.,* protein query against a translated nucleotide database) was used to identify sequences that may have been excluded from the protein databases. As above, sequences with an E-value less than or equal to 0.05 were used as queries in reciprocal BLASTx searches (*i.e.,* translated nucleotide sequence against protein database) to confirm homology. If at this step homologous sequences could not be identified, these sequences were considered 'not identified'. It should be noted that definitive absences (or losses) cannot be unambiguously proven by these methods. The possibility remains that

some sequence may have evolved beyond detection by our methods but may exist and function in the organisms in question. Additionally, one can never state with 100 percent confidence that an entire genome has been sequenced; it is always possible that a small fragment has been missed and that this may account for the 'missing' gene. Nonetheless, exhausting organismal databases (*i.e.,* protein, nucleic acids), in combination with searches carried out using multiple methods, and including multiple genomes from closely related lineages into the analyses provides some confidence to suggest that these sequences may have indeed been lost.

2.2.4 Deducing presence in the LECA

Individual genes or subfamilies of proteins were considered present in the LECA based on their presence in at least four supergroups. This is a parsimony argument based on observations from the literature. First, multiple phylogenomic analyses using multiple methods and datasets consistently recover similar relationships between eukaryotic supergroups. Two eukaryotic "Megagroups" have recently been recognized (Adl et al., 2012, *inter alia*): Amorphea, which comprises Opisthokonta, Amoebozoa, and Apusomonada, and Diaphoretickes, which is an amalgamation of Archaeplastida, SAR, and the CCTH group. The Excavata forms a clade unto itself between the Amorphea and Diaphoretickes. Therefore, a gene or subfamily present in at least four supergroups would overcome the need to rely on a rooted tree of eukaryotes since the gene in question would be present in all three major lineages. Parsimony would argue that the gene most likely arose in the ancestor of those lineages, the LECA.

The second observation is that recent rooting hypotheses suggest that the eukaryotic root lies either between Amorphea and Excavata, between Diaphoretickes and the Excavata, or in the Excavata (Cavalier-Smith, 2010; Derelle et al., 2015; Wideman et al., 2013). Therefore, proteins with distributions that span the backbone of the eukaryotic tree can be considered to have been present in the LECA, whereas proteins with distributions restricted to specific taxonomic groups can be considered to have arisen more recently. For example, proteins identified in three supergroups could also be considered to have been present in the LECA, if they are broadly distributed. If a gene family is found in Opisthokonta, Amoebozoa, and Archaeplastida, it would likely have been present in the LECA because the distribution spans the backbone of the eukaryotic tree. By contrast, proteins found in Opisthokonta, Amoebozoa, and Excavata (or Archaeplastida, SAR, and CCTH), could either have been present in the LECA, depending on the placement of the root of eukaryotes, or maybe the result of lineage-specific expansion (see section 1.2.7). Proteins present in two supergroups could be ancient proteins that have undergone substantial loss or could be the result of convergent evolution and are therefore analyzed more deeply using phylogenetic methods. Proteins present in one supergroup is considered to be a lineage-specific expansion.

2.2.5 Methodology used only in Chapter 3

2.2.5.1 Comparative genomics and identification of ArfGEF homologues

With the advent of high quality genomes and more sensitive homology searching tools, the comparative genomic analysis of ArfGEFs was approached

differently than for the other analyses presented here. Identification of candidate Sec7 domain-containing proteins was performed using HMMER (http://hmmer.org). An HMM was constructed using the Sec7 domains, which are approximately 200 amino acids in length, from all ArfGEFs identified in *H. sapiens* and *S. cerevisiae* and was used to search a close relative of *H. sapiens* (*Rattus norvegicus*). Candidate sequences were confirmed as GEFs by the identification of a Sec7 domain by InterProScan at the European Molecular Biology Laboratory (EMBL; Hunter et al., 2009; Mitchell et al., 2014). In addition, reciprocal BLAST against the *H. sapiens* genome was performed in order to assess membership to a particular subfamily (see section 2.2.1). The Sec7 domains of all identified sequences were then incorporated into the existing HMM, by aligning the new sequences to the existing multiple sequence alignment and using the new alignment to build a new HMM resulting in a more accurate model. This iterative searching and incorporation of new sequences continued from genomes most closely related to *H. sapiens* to those most distantly related, at which point, a final search of all genomes was carried out using an HMM built from all of the identified sequences in order to identify any highly divergent sequences that may have been missed.

2.3 Phylogenetic Analysis

Phylogenetic analysis was carried out to confirm statements of orthology made by BLAST. Generally, the results from the phylogenetic analyses did not contradict the BLAST assignments; however, in many cases, phylogenies were able to classify sequences that were not classified by BLAST. Phylogenetic analysis can

also clarify patterns of gene expansion. For example, it can identify paralogous genes resulting from recent gene duplications (limited to one or a few genomes) versus paralogous genes that arose from more ancient gene duplications and are present across multiple supergroups.

Phylogenetic analysis of protein sequences was used rather than nucleotide sequences because each position has 20 states instead of four, and is potentially more informative, especially over the time scales being examined, whereas nucleotides often do not retain enough phylogenetic signal to be informative. Moreover, protein sequences are less prone to convergent or parallel substitutions, and are not affected by differences in GC content, reducing the probability of artefact being introduced into the analysis by these evolutionary processes (Steel et al., 1993). Multiple sequence alignments were generated using MUSCLE (Edgar, 2004). Alignments were viewed using either MacClade (Maddison and Maddison, 2005) or Mesquite (Maddison and Maddison, 2015), and adjusted manually as necessary to correct any obvious misalignments made by the program. Alignments were masked and trimmed manually. This allowed a greater degree of control over which segments of the alignment to include in the analysis than is normally available with automated masking and trimming programs. One risk of masking alignments manually is the introduction of subjectivity into the analysis; it raises the question of 'what is a conserved site?' To maintain consistency between analyses and reduce the amount of subjectivity introduced into the analysis, the following guideline was followed to determine whether or not to retain or exclude sites from the analysis. Constant or invariant sites in the alignment were identified and were used as

reference points, signalling that the region of the alignment in question was properly aligned. These reference points, along with any intervening sequence, were retained for the analysis, while excluding any insertion-deletions. Regions of the alignment not anchored by an invariant or constant site and that appeared to be randomly aligned sequence were not included in the mask. All alignments can be found at the following link: http://www.ualberta.ca/~aschlach.

To determine which model of sequence evolution best fit the data and the parameters associated with it, model testing of trimmed alignments was carried out using ProtTest (Abascal et al., 2005). ProtTest analyzes the input alignment and identifies an empirically determined substitution matrix that best represents the data. Available substitution matrices are: WAG (Whelan and Goldman, 2001), Dayhoff (Dayhoff et al., 1978), JTT (Jones et al., 1992), mtREV (Adachi and Hasegawa, 1996), mtMam (Cao et al., 1998), mtArt (Abascal et al., 2007), VT (Müller and Vingron, 2000), RtREV (Dimmic et al., 2002), CpREV (Adachi et al., 2000), Blosum62 (Henikoff and Henikoff, 1992), LG (Le and Gascuel, 2008), DCmut (Kosiol and Goldman, 2005), HIVw/HIVb (Nickle et al., 2007), and FLU (Dang et al., 2010). These substitution matrices provide the probability that one amino acid (*e.g.,* lysine) is replaced by another (*e.g.* glutamate), in every pairwise combination. While this does not represent an exhaustive list, it encompasses the most commonly used substitution matrices not designed for a specific protein type (*e.g.*, coiled-coil proteins, transmembrane proteins, *etc.*). In addition, ProtTest determines whether or not all sites in the alignment evolve at the same rate. It does so to account for a proportion of invariant sites (+I), gamma rate categories (+G), the observed

frequency of amino acids (+F), or a combination of these parameters. These latter parameters inform the phylogenetic program how to account for differences in the rate of evolution of different sites in the alignment. The +I parameter indicates that some proportion of columns in the alignment do not change, and the +G parameter follows a gamma distribution and essentially bins all of the columns in the alignment into a pre-set number of rate categories to reduce the complexity of the dataset and the computational load (Hasegawa et al., 1985; Yang, 1994). Typically four rate categories provide an optimal trade off between fitting the data and approximating all of the rates in the alignment (Yang, 1994). An alternative to the gamma model is the CAT model (named as such because it classifies sites into categories), which evaluates the data and determines the best number of rate categories (Lartillot and Philippe, 2004; Le et al., 2008). However, this model requires very large alignments (*i.e.,* 100's of taxa and 1000's of positions, typically larger than most membrane trafficking proteins) for the calculation to be accurate. The +F parameter assess the amino acid content (*i.e.,* frequency of each amino acid) of the input alignment and compares it with the amino acid content of the data set used to generate the substitution matrix (Cao et al., 1994). If sufficiently different, the +F option indicates that the phylogeny program should use the observed frequencies of amino acids found in the alignment, rather than the frequencies used to generate the empirical model (Cao et al., 1994). For a model to be chosen, it needed to be deemed the best fit by at least two selection criteria, *i.e.,* negative log-likelihood, Akaike Information Criteria, or Bayesian Information Criteria. If no consensus was reached, the simplest model (least number of parameters) was

chosen. This process was implemented, as not all phylogeny programs (*i.e.,* PhyML and RAxML) are able to incorporate model selection as part of the tree building process. Additionally, this maintained parameter consistency across each method. Exceptions to this process occurred for earlier versions of MrBayes and RAxML, for which a limited repertoire of substitution matrices were available. If the best model determined by ProtTest was unavailable for MrBayes, then the 'mixed' model parameter was used which incorporates model selection into the tree building process. If the chosen model was unavailable for RAxML, then the WAG model was selected.

Phylogenetic trees were generated using multiple methods including: bayesian analysis (MrBayes and Phylobayes; Lartillot et al., 2009; Ronquist and Huelsenbeck, 2003), and maximum-likelihood analysis (PhyML and RAxML;Guindon and Gascuel, 2003; Stamatakis, 2006). In bayesian analysis, the program generates multiple randomized starting trees. A change is made to each tree (a generation) that alter the branching order or branch length, resulting in a different topology. The new tree is compared against the previous one and if it is a better tree, then it is kept, if not then it is rejected and the previous tree is kept. Multiple chains are run to increase the probability that the single best tree will be found. The best tree is found when all off the chains converge on the same phylogenetic tree. The program monitors the different chains through a statistical measure, the splits frequency, which measures the similarity of tree samples of the independent runs by evaluating all aspects of tree topology including branching order and branch length.

The program requires that the analysis reach a minimum threshold regardless of the number of generations.

MrBayes analyses were carried out using the model specified by ProtTest (see above), using 4 Gamma rate categories, and accounting for invariant sites when necessary. Analyses were run for a minimum of 1,000,000 generations and until convergence, measured by a splits frequency of 0.01 or smaller. Phylobayes analyses were carried out in a similar fashion, but rather than specifying a minimum number of generations, analyses were run until convergence (splits frequency of 0.1) and minimum sample size of 100 trees.

In maximum-likelihood analysis, the program generates a 'best-tree' using the maximum-likelihood calculation (Guindon and Gascuel, 2003; Stamatakis, 2006). Support for each branching point, or node, is determined by bootstrapping where the program takes the input alignment and randomly samples positions with replacement generating a subsample of the same length as the original alignment (Felsenstein, 1985). A maximum-likelihood tree is then built from the subsampled alignment and compared to the best likelihood tree. If the trees are in agreement then this is bootstrap support for the entire tree. If only some relationships are in agreement, then this sampling is support for those relationships only. Each iteration is called a 'pseudoreplicate', because each bootstrap is constructed from a subsample of the original dataset. The number of times a specific relationship (or node) is reconstructed is displayed on the tree as a percentage of the total number of bootstraps conducted. The underlying assumption in bootstrapping is that if the best maximum-likelihood tree is the best representative for the data, then any

subsampling of the input alignment should produce the same tree. By contrast, a tree that is biased by a few strongly conserved residues (positions in the alignment) and poor conservation in the remainder of the alignment, would receive low bootstrap support because the majority of the positions would not support the maximum-likelihood tree.

PhyML and RAxML analyses were run, accounting for the model specified by ProtTest, and bootstrapping using 100 pseudoreplicates. Large phylogenetic analyses were performed using the Cyberinfrastructure for Phylogenetic Research (CIPRES) webserver (Miller et al., 2010). The remaining analyses were performed on local computing clusters. Support values generated by each method were mapped onto either the best MrBayes or Phylobayes tree generated from the original alignment.

2.4 Tertiary Structure Prediction

Structural analyses can provide information that cannot be gleaned from linear sequence comparisons. For example, the presence of some domains, *e.g.,* α-solenoid, is difficult to predict from primary sequence alone, but is easily identifiable from the tertiary structure. Therefore, to predict the tertiary structures of potential α-solenoid and β-propeller domain-containing proteins homology modeling was carried out using the Phyre2.0 server (Kelley and Sternberg, 2009; Kelley et al., 2015).

Phyre2.0 generates predicted tertiary structures by first identifying homologues of the query sequence by searching a specially curated database using

the HHblits program. HHblits is a sequence-profile search program (*i.e.,* searches a single sequence against a database of HMMs; Remmert et al., 2011). At the same time, the secondary structure of the query is predicted using PSIPRED (PSI-BLAST – based secondary structure prediction; Jones, 1999). An HMM is built using the query, its homologues, and its secondary structure and is then used to search a library of experimentally determined structures using HHsearch (Söding, 2005). Indels are modelled using a library of fragments of known protein structures. Side chains are fit to the backbone using a side chain rotamer library and the R3 library (Xie and Sahinidis, 2006).

I predicted the tertiary structure of Sed4 (homologue of Sec12) suspecting a β-propeller fold, and the subunits of TSET (homologues of the AP complexes and COPI) suspecting α-solenoids, β-propellers, and longin domains. Homology modeling was carried out using default settings (Kelley and Sternberg, 2009; Kelley et al., 2015). The resulting structures were visualized using MacPymol (www.pymol.org).

Chapter 3: Comparative genomic and phylogenetic analysis of ArfGAP and ArfGEF proteins identifies a complex Arf regulatory system present in the LECA

A portion of this chapter has been published as:

Schlacht, A., Mowbrey, K., Elias, M., Kahn, R.A., Dacks, J.B. 2013. Ancient complexity, opisthokont plasticity, and discovery of the 11th subfamily of ArfGAP proteins. *Traffic* 14: 636-649

3.1 Overview

ADP-ribosylation factor (Arf) GTPases are important regulators of membrane trafficking, remodelling of the actin cytoskeleton, and the synthesis of phospholipids (for review, see D'Souza-Schorey and Chavrier, 2006). Comparative genomic and phylogenetic analyses have suggested that the LECA possessed as single Arf sequence, as most extant eukaryotic taxa only possess one Arf (Berriman et al., 2005; Li et al., 2004). Because Arfs do not appear to have undergone large paralogous expansions prior to the LECA, they are unable to provide evidence for, or against, autogenous organelle evolution. Two different protein families regulate Arf GTPases: GEFs and GAPs, both are highly paralogous and contain members that are broadly distributed across eukaryotes. These regulators may be able to provide insight into the evolution of the Arf system, as has a recent analysis of the RabGAPs (TBCs: Gabernet-Castello et al., 2013).

In this chapter, I use comparative genomics to determine which ArfGAPs and ArfGEFs are conserved across eukaryotic diversity, and by extension, which were present in the LECA. I also use phylogenetic analysis to elucidate the order of gene duplications giving rise to the diverse paralogues, especially those found in humans. These analyses will also reveal the number of paralogues of each GAP and GEF subfamily present in the LECA, helping us to understand how specificity may be encoded in the Arf system.

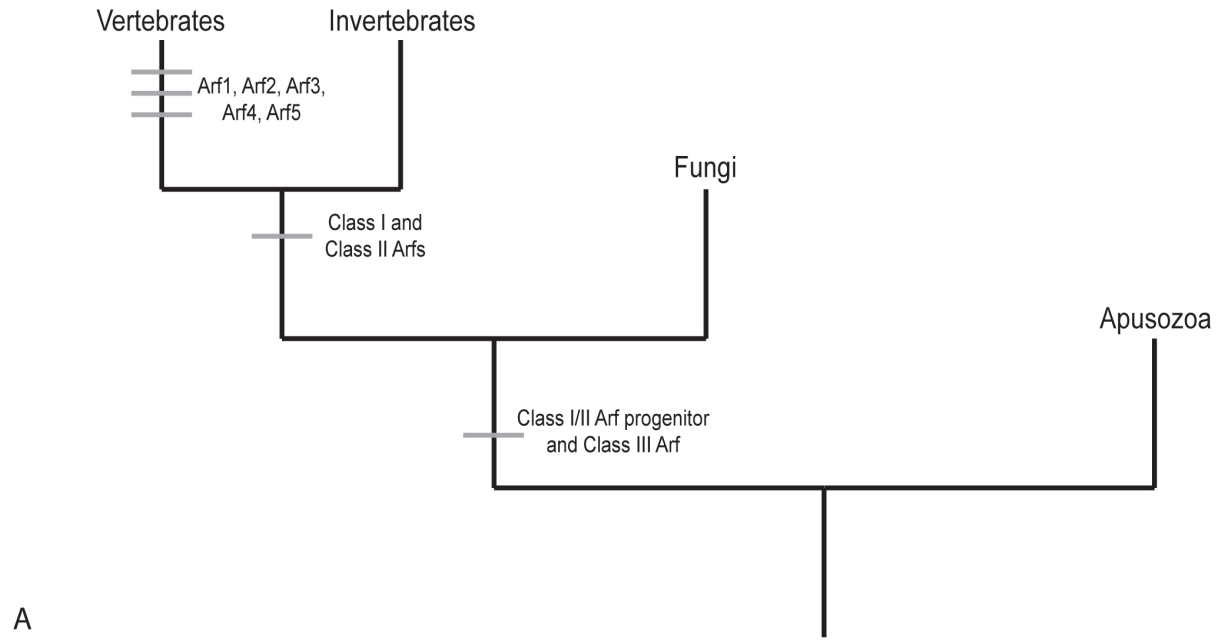
3.2 Introduction

Arfs are small, ~21kDa GTPases within the Ras superfamily (Kahn and Gilman, 1984; Kahn and Gilman, 1986). Arfs act as molecular switches to regulate a

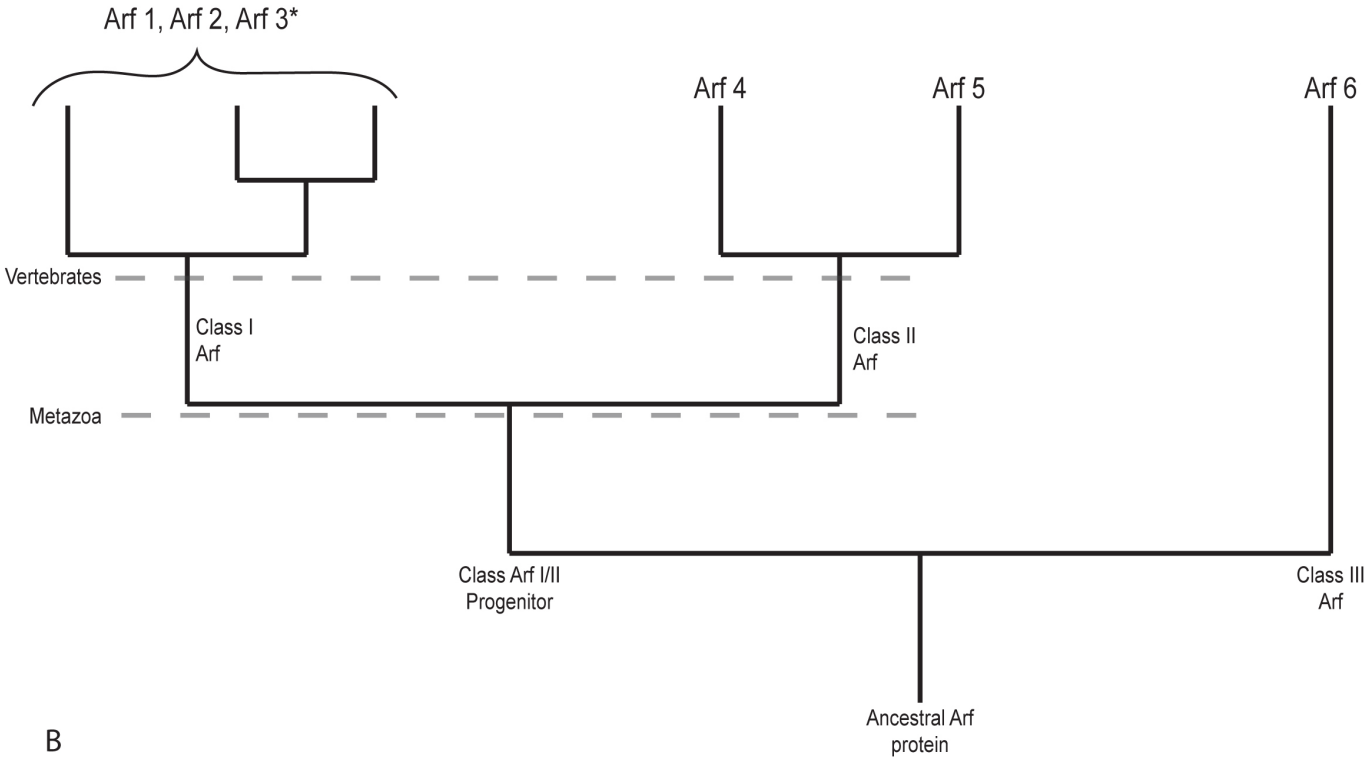
variety of cellular activities including membrane trafficking, remodelling of the actin cytoskeleton, the synthesis of phosphoinositides, and are able to activate phospholipase D (Brown et al., 1993; Cockcroft et al., 1994; D'Souza-Schorey and Chavrier, 2006, *inter alia*; Donaldson and Jackson, 2011, *inter alia*; Honda et al., 1999; Ooi et al., 1998; Paleotti et al., 2005; Yorimitsu et al., 2014, *inter alia*). Arfs cycle between an active GTP-bound state, in which Arf is able to interact with effector proteins, and an inactive GDP-bound state, where Arf signalling is terminated (East and Kahn, 2011; Wright et al., 2014).

Six genes encoding different Arf paralogues have been identified in vertebrates and are divided into three classes: class I (Arf1, Arf2, and Arf3), class II (Arf4 and Arf5), and class III (Arf6; Tsuchiya et al., 1991). Each Arf paralogue has been shown to act at distinct compartments within the endomembrane system or at the plasma membrane (Chun et al., 2008; Paleotti et al., 2005; Volpicelli-Daley et al., 2005). The division between these classes is based on sequence similarity and shared evolutionary histories. Arf1, Arf2, and Arf3 are the product of multiple gene duplications of the invertebrate class I Arf near the vertebrate transition (Figure 3-1; Manolea et al., 2010). The same is thought for the gene duplication of the invertebrate class II Arf giving rise to Arf4 and Arf5; however, resolution at the relevant nodes has not yet been obtained (Figure 3-1; Li et al., 2004). Moreover, the progenitor of class I and of class II Arfs arose from the duplication of a single gene

Figure 3-1. Overview of Arf evolution in opisthokonts. A) Multiple gene duplications (grey bars) in opisthokonts produced the six mammalian Arf paralogues. The first duplication is thought to have occurred in the ancestor of Holozoa and Fungi resulting in the class III Arf (*H. sapiens* Arf6 and *S. cerevisiae* Arf3) and the progenitor of the class I and II Arf proteins. The second gene duplication occurred near the base of Metazoa, producing the class I and class II Arfs. Subsequent gene duplications of the class I Arf near the vertebrate transition produced Arf1, Arf2, and Arf3. Duplication of the class II Arf produced Arf4 and Arf5 is also thought to have occurred at or near the base of vertebrates. B) Information from panel A is reconfigured as the hypothetical Arf gene tree. Data used to generate this figure is from Li et al., (2004) and Manolea et al., (2010). *It should be noted that the order in which class I Arfs evolved has not yet been confidently determined.


Vertebrates
Invertebrates
Arf1, Arf2, Arf3,
Arf4, Arf5
Fungi
Class I and
Class II Arfs
Apusozoa
Class I/II Arf progenitor
and Class III Arf

A


Arf 1, Arf 2, Arf 3*
Arf 4
Arf 5
Arf 6
Vertebrates
Class I
Arf
Class II
Arf
Metazoa
Class Arf I/II
Progenitor
Class III
Arf
Ancestral Arf
protein

B

prior to the divergence of animals (Figure 3-1). The class III/Arf6 is thought to have arisen from a single Arf gene in the ancestor of opisthokonts that also gave rise to the progenitor of the class I and class II Arfs (Figure 3-1; Li et al., 2004).

This scenario of iterative gene duplication giving rise to the complexity of Arfs observed in mammals, in combination with the observation that the LECA likely only possesses a single Arf (Berriman et al., 2005; Li et al., 2004), suggests that organisms with multiple Arfs arose convergently via lineage-specific expansion of the single Arf ancestor. This observation prompted the question: how was specificity encoded if the LECA only possessed a single Arf?

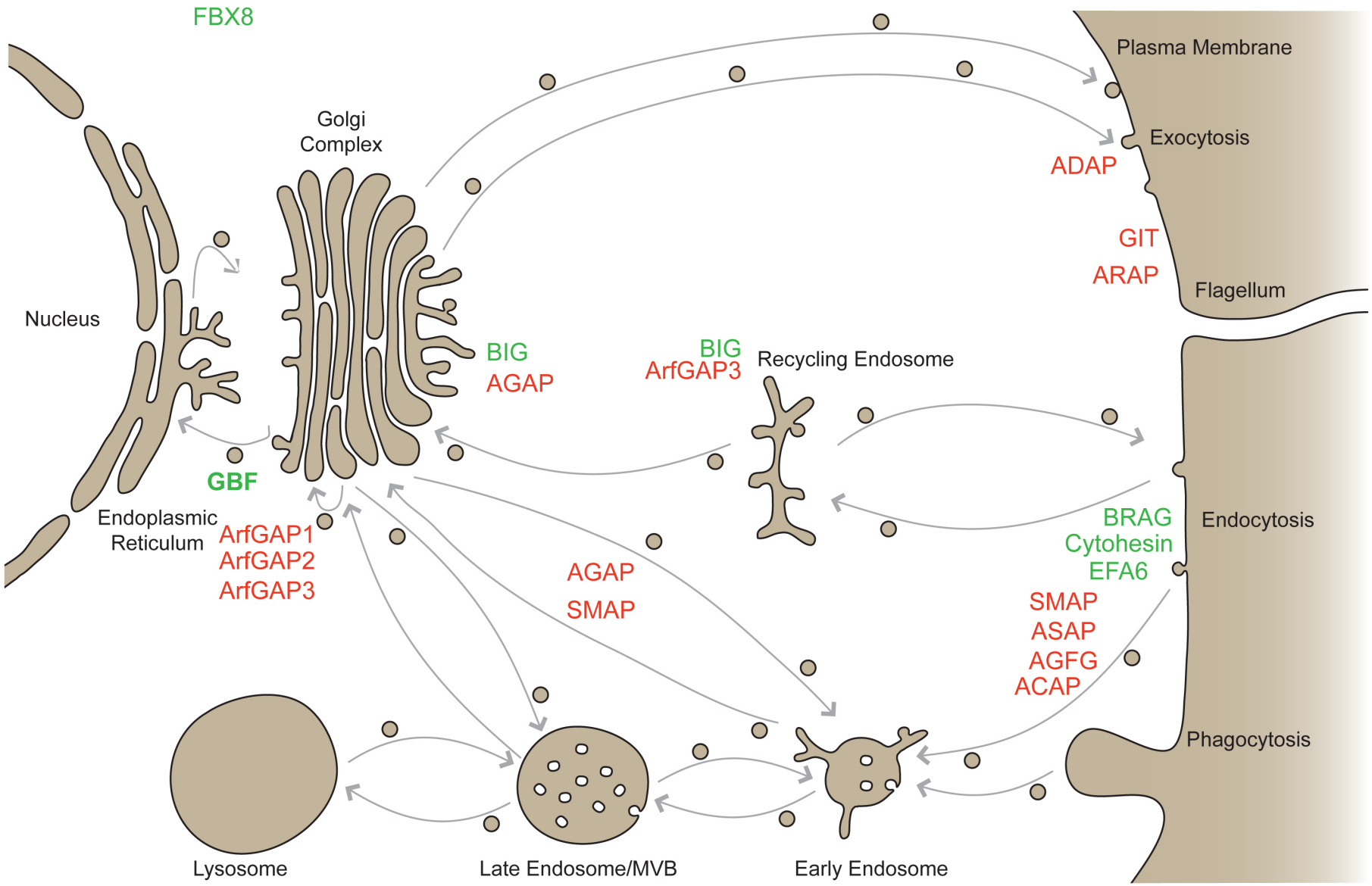
In order to generate a suitable hypothesis, one must examine how Arfs are regulated. Arfs are unable to efficiently exchange GDP for GTP or to hydrolyze GTP in the absence of other factors (D'Souza-Schorey and Chavrier, 2006). Arf activation depends on the action of GEFs, defined by the presence of the Sec7 domain, to promote the exchange of GDP for GTP (Cox et al., 2004). Similarly, hydrolysis of GTP to GDP depends on the interaction with GAPs, defined by the presence of the ArfGAP domain, to stimulate hydrolysis (Kahn et al., 2008). Both ArfGAPs and GEFs are thought to act as Arf effectors, as well as Arf regulators (East and Kahn, 2011; Padovani et al., 2014). For example, the *S. cerevisiae* Sec7 protein, the orthologue of the *H. sapiens* BIGs, is both an activator of the yeast Arf1 and an effector (Richardson et al., 2012). Arf1 binding to the HDS1 (Homology Downstream of Sec7 1) region of Sec7p relieves the autoinhibitory effect of HDS1 and targets Sec7p to the TGN (Richardson et al., 2012). Richardson et al., also found that this activation occurs though a positive feedback loop, as the addition of increasing amounts of Arf-GTP

accelerated the rate of GDP-GTP exchange (Richardson et al., 2012). Arf6-GTP is able to negatively regulate GTP exchange by EFA6 (exchange factor for Arf6) through an interaction with the PH domain and C-terminal region of EFA6 (Padovani et al., 2014). This not only allows EFA6 to be regulated by its own product, Arf6-GTP, but also by other Arf6 GEFs such as, cytohesin and BRAG (Brefeldin A-resistant Arf guanine nucleotide exchange factor) that also activates Arf6 (Frank et al., 1998; Someya et al., 2001).

The human ArfGAP2 and ArfGAP3 proteins and their yeast orthologue, Glo3p, are thought to act downstream of Arf1. ArfGAP2 and ArfGAP3 are recruited to Golgi membranes in a COPI-dependent manner through an interaction with γ- and β'-COPI (Eugster et al., 2000; Frigerio et al., 2007; Weimer et al., 2008). In yeast, binding of one of the SNAREs Bet1p, Bos1p, or Sec22p to Glo3p is necessary for the formation of a priming complex with Arf1 to recruit COPI (Rein et al., 2002). In this *in vitro* analysis, addition of Glo3 to the reaction mixture prior to COPI was required for the incorporation of SNAREs into the budding vesicle, indicating that binding of the SNAREs to Glo3p occurs before binding to COPI. These analyses suggest that ArfGAP2 and ArfGAP3 and the yeast orthologue Glo3 not only act as terminators of Arf signalling, but also as Arf effectors.

Ten subfamilies of ArfGAP domain-containing proteins and six subfamilies of Sec7 domain-containing proteins have been identified in animals, each with different subcellular functions and locations (Figure 3-2; Casanova, 2007; Kahn et al., 2008). It is worth noting that, even in organisms such as humans that possess

Figure 3-2. Subcellular localization of ArfGAP and GEF subfamilies. Overview of the membrane trafficking system with the location of action of each ArfGAP and GEF subfamily indicated in red and green, respectively. ArfGAPs: ArfGAP1, ArfGAP2, and ArfGAP3 localize to the Golgi complex and are primarily involved in COPI-dependent transport. SMAP is involved in endocytosis and transport between the TGN and endosomes. ADAP is involved in secretion of regulated secretory vesicles. AGAP regulates trafficking between the TGN and endosomes. ASAP is responsible for regulating specialized plasma membrane structures (*e.g.*, focal adhesions) and endocytosis. GIT is involved in signal integration with Rho proteins at focal adhesions. ARAP is also involved in the integration of Arf and Rho signalling pathways. ArfGEFs: GBF is primarily found at the *cis*-Golgi and is responsible for regulation of COPI vesicle formation. BIG is primarily found at the TGN and interacts with AP-1 and GGAs. EFA6 regulates endocytosis and cytoskeletal dynamics at the plasma membrane. Cytohesin is involved in both exo- and endocytosis, in addition to regulating cell motility. BRAG is plasma membrane localized and selectively regulates endocytosis. FBX8 is set to the side as its function and localization are currently unknown.


FBX8
Plasma Membrane
Golgi
Complex
Exocytosis
ADAP
GIT
ARAP
Flagellum
Nucleus
BIG
AGAP
BIG
ArfGAP3
Recycling Endosome
GBF
Endoplasmic
Reticulum
ArfGAP1
ArfGAP2
ArfGAP3
AGAP
SMAP
BRAG
Cytohesin
EFA6
Endocytosis
SMAP
ASAP
AGFG
ACAP
Phagocytosis
Lysosome
Late Endosome/MVB
Early Endosome

multiple Arf paralogues, the number of GAPs and GEFs both outnumber the Arfs (Casanova, 2007; Kahn et al., 2008), prompting the question of how many ArfGAPs and GEFs were present in the LECA? To addresses this question, I analyzed the conservation of each ArfGAP and ArfGEF subfamily across a representative set of eukaryotic taxa. I found that six ArfGAP and three ArfGEF subfamilies were present in the LECA, likely as single paralogues, indicating that the LECA possessed a much smaller Arf system than that observed in extant eukaryotes such as *D. discoideum, N. gruberi*, and trypanosomes.

3.3 Abbreviated materials and methods

3.3.1 ArfGAPs

Comparative genomic analyses carried out as described in section 2.2 using the genomes illustrated in Figures 3-3 and 3-4. Phylogenetic analyses were carried out as described in section 2.3. The details of each phylogenetic analysis, including: number of taxa, length of masked alignment, and model parameters for each method can be found in Table 3-1.

Table 3-1. Parameters of phylogenetic analysis, corresponding dataset, and figure number for each ArfGAP subfamily

Figure	Dataset Name	Number of taxa	Length of alignment (a.a.)	Evolutionary Model Used		
				MrBayes	PhyML	RAxML
3-7	ACGSAP	118	565	WAG+G	WAG+G	PROTCATWAG
3-10	ArfGAPC2	35	107	WAG+G	WAG+G	PROTCATWAG
3-12	ArfGAP1	28	111	WAG+G	WAG+G	PROTCATWAG
3-13	ArfGAP2/3	41	191	WAG+I+G	WAG+I+G	PROTCATWAG
3-14	ArfGAP23.holozoa.R2	21	443	mixed+I+G	JTT+I+G	PROTCATWAG
3-15	ACAP	54	212	WAG+G	WAG+G	PROTCATWAG
3-16	ACAP_holozoa	21	197	WAG+G	WAG+G	PROTCATWAG
3-17	AGFG	38	288	WAG+G	WAG+G	PROTCATWAG
3-18	AGFG_holozoa	16	254	mixed+G	JTT+G	PROTCATWAG
3-19	ADAP	20	347	WAG+G	WAG+G	PROTCATWAG
3-20	ASAP	27	658	mixed+G	JTT+G	PROTCATWAG
3-21	SMAP	53	110	WAG+I+G	WAG+I+G	PROTCATWAG
3-22	SMAP_holozoa	17	264	mixed+I+G	JTT+I+G	PROTCATWAG
3-23	AGAP_holozoa	28	661	JTT+I+G	JTT+I+G	PROTCATWAG
3-24	GIT	18	427	mixed+G	JTT+G	PROTCATWAG
3-25	ARAP	19	861	mixed+I+G	JTT+I+G	PROTCATWAG

3.3.2 ArfGEFs

Comparative genomic analyses were performed as described in section 2.2 using the genomes illustrated in Figures 3-3 and 3-4. Phylogenetic analyses were carried out as described in section 2.3. Details of each phylogenetic analysis,

including: number of taxa, length of masked alignment, and model parameters for each method can be found in Table 3-2.

Table 3-2. Parameters of phylogenetic analysis, corresponding dataset, and figure number for each ArfGEF subfamily

Figure	Dataset name	Number of taxa	Length of alignment (a.a.)	Evolutionary model used	
				Phylobayes	RAxML
3-29	GBF.r1	72	1062	LG+CAT+I+G	LG+CAT+F
3-30	BIG.euk.r1	125	1219	LG+CAT+I+G*	LG+CAT+F
3-31	BRAG.r1	36	572	LG+CAT+I+G	LG+CAT+F
3-32	CYTH.r1.euks	74	339	LG+CAT+I+G*	LG+CAT+F
3-33	CYTH.holozoa.r2	45	387	LG+CAT+I+G	LG+CAT+F
3-34	EFA6.r2	47	459	LG+CAT+I+G	LG+CAT+F
3-35	FBX8.r1	13	318	LG+CAT+I+G	LG+CAT+F

*Denotes that Phylobayes analyses did not converge are therefore not displayed.

3.4 Results

3.4.1.1 Multiple levels of stringency and validation of the comparative genomic approach

Comparative genomic analyses were carried out to assess the evolution of the ArfGAP domain-containing proteins. To identify ArfGAP homologues, homology searching using BLAST and HMMer was performed on the genomes of 38 organisms spanning the diversity of eukaryotes (Figure 3-3, 3-4), using the human and *S. cerevisiae* sequences as queries. 446 candidate sequences were identified, of which

Figure 3-3. Relative relationships of sequenced genomes used in the analysis of ArfGAP and ArfGEF evolution. 75 eukaryotic genomes were analyzed, 4 genomes are specific to the survey of ArfGAPs (marked by *) and 40 are specific to the ArfGEF study (marked by $). 31 genomes are common to both. The difference in the number of genomes sampled in both analyses is the result of substantially more genomes available during the ArfGEF study as compared to when the ArfGAPs were analyzed. Due to spatial constraints and to the large number of opisthokont genomes sampled, the relative relationships of these organisms is depicted in Figure 3-4.

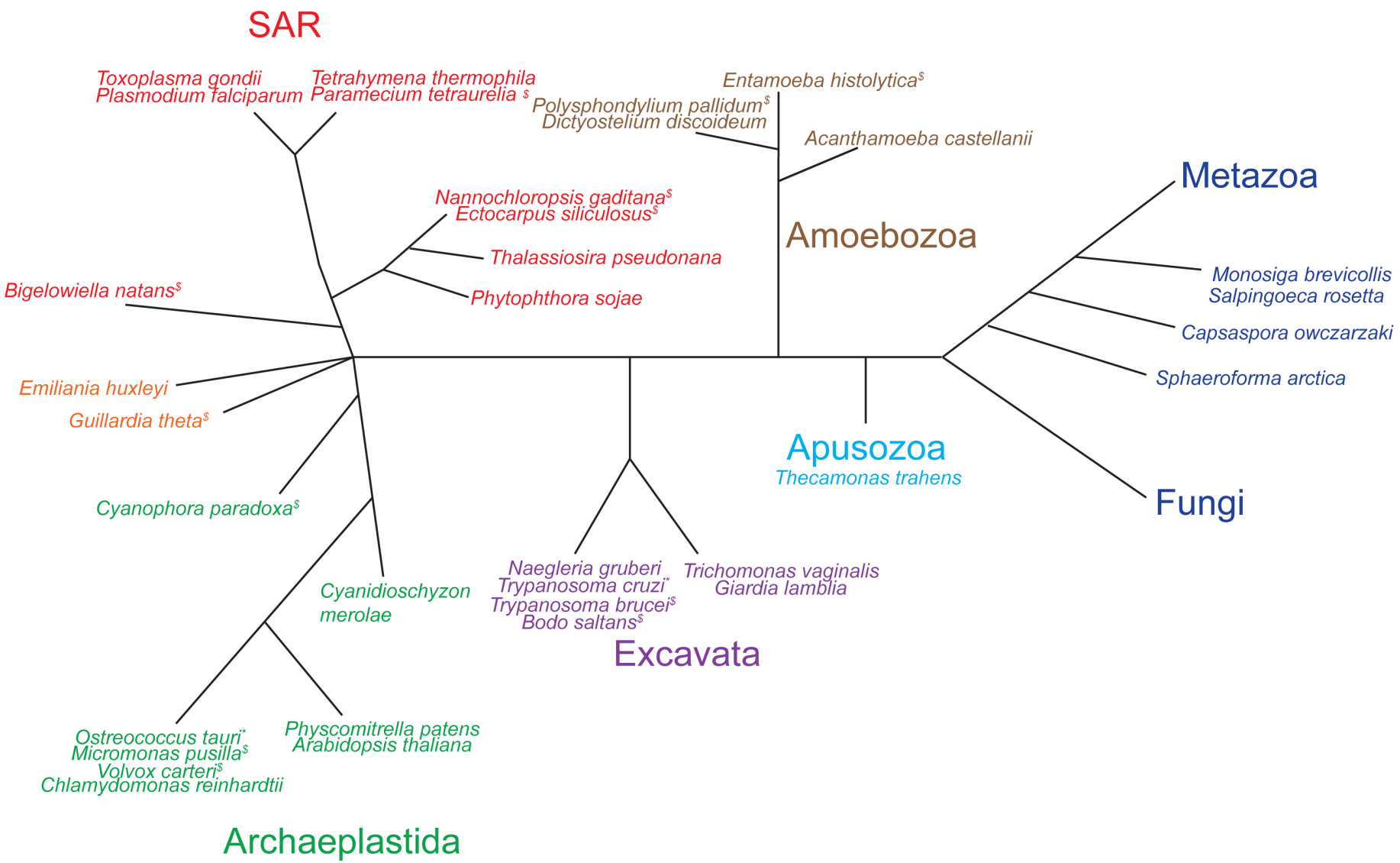
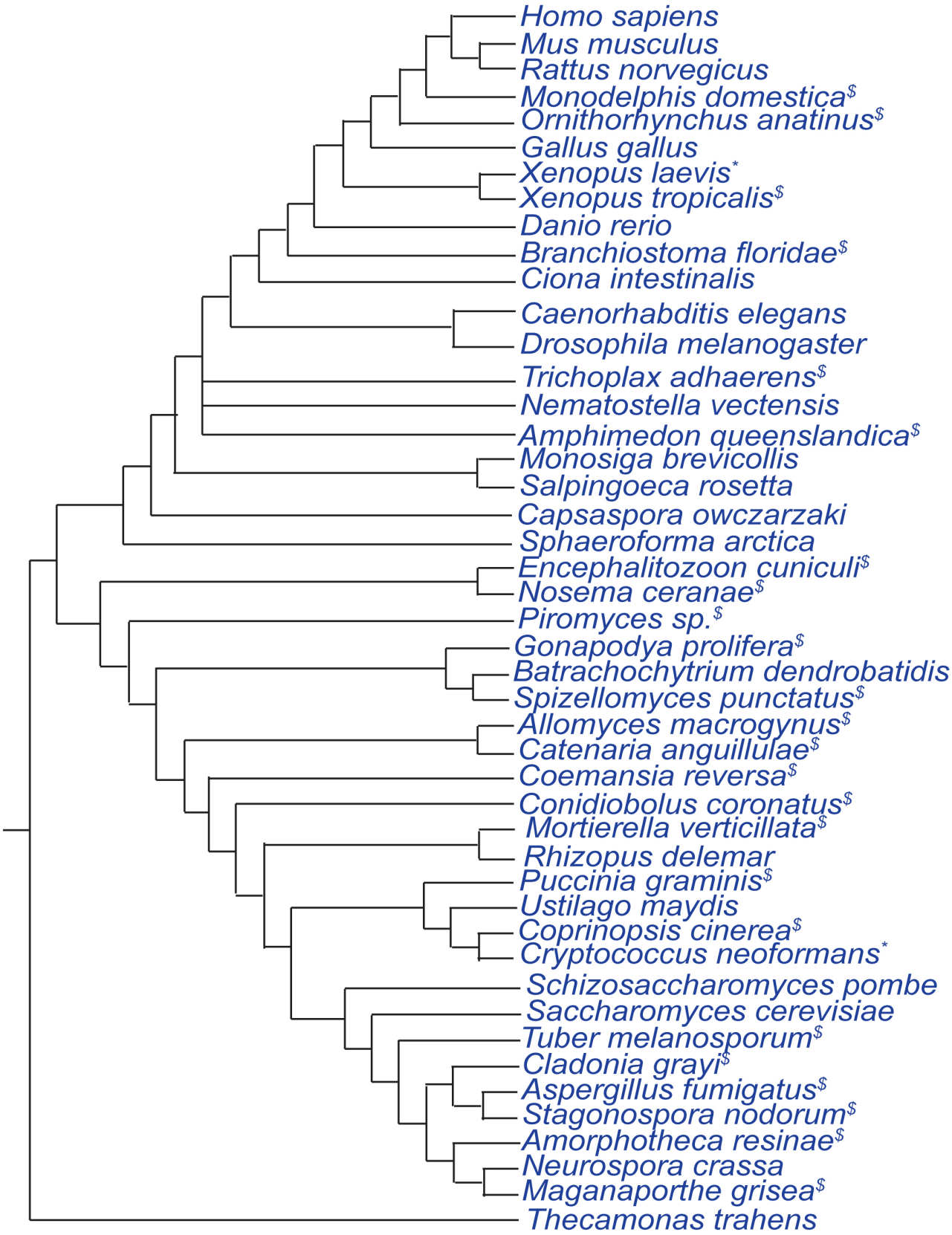

SAR
Toxoplasma gondii
Plasmodium falciparum
Tetrahymena thermophila
Paramecium tetraurelia $
Nannochloropsis gaditana$
Ectocarpus siliculosus$
Thalassiosira pseudonana
Phytophthora sojae
Bigelowiella natans$
Emiliania huxleyi
Guillardia theta$
Cyanophora paradoxa$
Cyanidioschyzon merolae
Ostreococcus tauri*
Micromonas pusilla$
Volvox carteri$
Chlamydomonas reinhardtii
Physcomitrella patens
Arabidopsis thaliana
Archaeplastida
Entamoeba histolytica$
Polysphondylium pallidum$
Dictyostelium discoideum
Acanthamoeba castellanii
Amoebozoa
Naegleria gruberi
Trypanosoma cruzi*
Trypanosoma brucei$
Bodo saltans$
Trichomonas vaginalis
Giardia lamblia
Excavata
Apusozoa
Thecamonas trahens
Metazoa
Monosiga brevicollis
Salpingoeca rosetta
Capsaspora owczarzaki
Sphaeroforma arctica
Fungi

Figure 3-4. Relative relationships of sequenced opisthokont genomes used in the analysis of ArfGAP and ArfGEF evolution. Illustration of the relationships of taxa sampled. Genomes specific to the ArfGAP analysis are denoted by an *, whereas those specific to the ArfGEF analysis are denoted by a $.

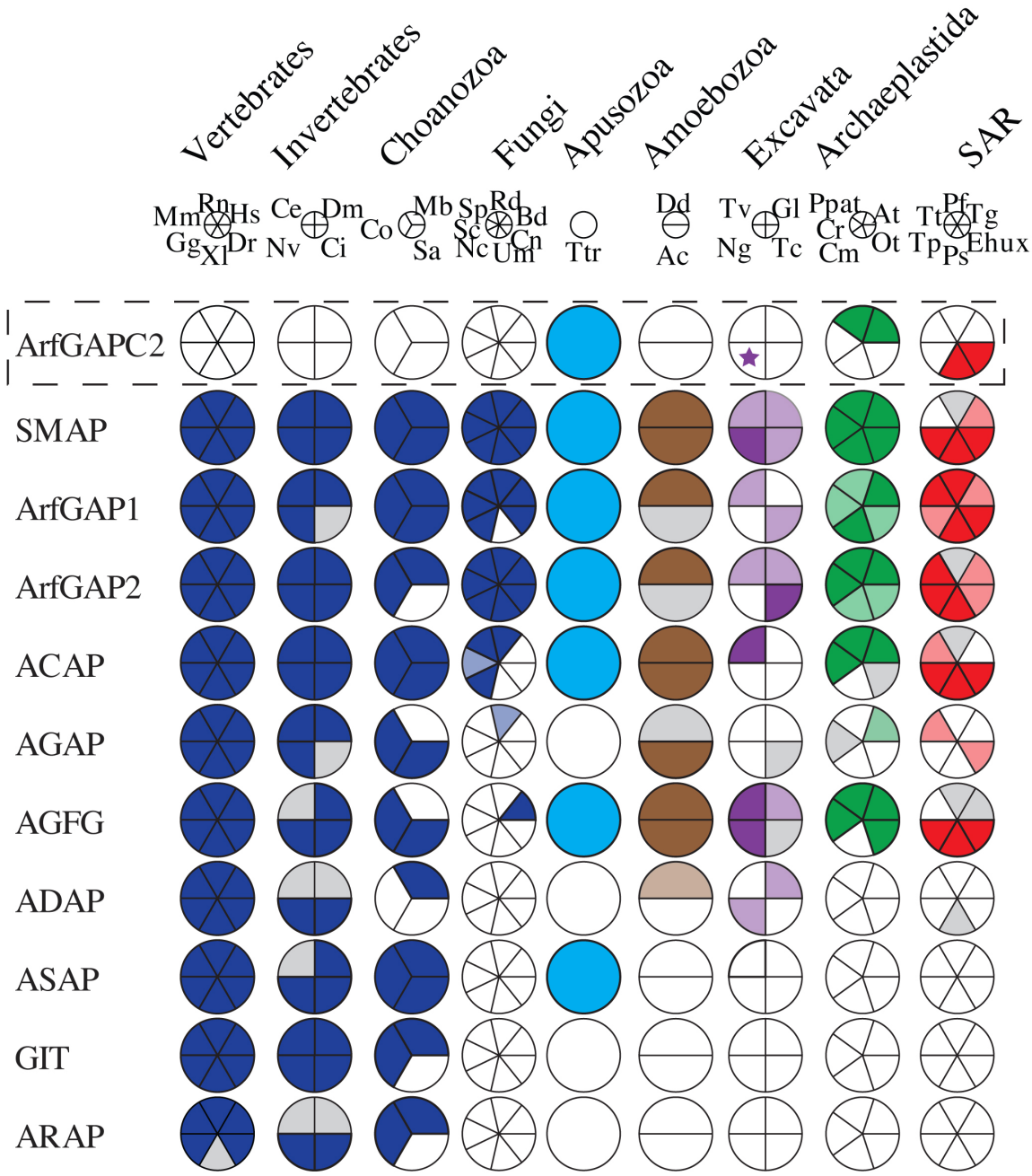

Homo sapiens
Mus musculus
Rattus norvegicus
Monodelphis domestica$
Ornithorhynchus anatinus$
Gallus gallus
Xenopus laevis*
Xenopus tropicalis$
Danio rerio
Branchiostoma floridae$
Ciona intestinalis
Caenorhabditis elegans
Drosophila melanogaster
Trichoplax adhaerens$
Nematostella vectensis
Amphimedon queenslandica$
Monosiga brevicollis
Salpingoeca rosetta
Capsaspora owczarzaki
Sphaeroforma arctica
Encephalitozoon cuniculi$
Nosema ceranae$
Piromyces sp.$
Gonapodya prolifera$
Batrachochytrium dendrobatidis
Spizellomyces punctatus$
Allomyces macrogynus$
Catenaria anguillulae$
Coemansia reversa$
Conidiobolus coronatus$
Mortierella verticillata$
Rhizopus delemar
Puccinia graminis$
Ustilago maydis
Coprinopsis cinerea$
Cryptococcus neoformans*
Schizosaccharomyces pombe
Saccharomyces cerevisiae
Tuber melanosporum$
Cladonia grayi$
Aspergillus fumigatus$
Stagonospora nodorum$
Amorphotheca resinae$
Neurospora crassa
Maganaporthe grisea$
Thecamonas trahens

410 were classified into one of the ten previously identified ArfGAP subfamilies, leaving 36 sequences unclassified (*i.e.,* rogues). Our inability to classify all of the ArfGAP domain-containing sequences may stem from the presence of additional domains in ArfGAP proteins (*e.g.,* PH domains, ankyrin repeats, *etc.*), which may increase the probability of misclassification. Therefore, additional criteria were applied to the reciprocal best hits to increase the stringency of these analyses. Assignment to a specific subfamily required that candidate sequences to retrieve the initial query with E-values 5-orders of magnitude better (*i.e.,* smaller) than those of the representatives of the next best-scoring ArfGAP subfamily, henceforth referred to as the '5-orders criterion'. At this criterion 334 sequences were unambiguously classified as a member of one of the ten ArfGAP subfamilies, leaving 112 rogue sequences (Figure 3-5).

Assignments of orthology at this criterion were regarded with high confidence and formed the basis upon which evolutionary inferences were made. Candidate sequences that did not satisfy this criterion were assessed at a less stringent 2-orders of magnitude better than the representatives of the next best subfamily (2-orders criterion), providing a set of more weakly supported hypotheses that are reported nonetheless.

Although the above criteria greatly increased the confidence of the sequence classification, the 2- and 5-orders criteria are arbitrary. Therefore, their accuracy was assessed, with the assumption that consistency of assignment corresponds to successful assignment. BLAST experiments were carried out using the ArfGAPs from primarily non-model organisms in order to assess homology. This served as a

Figure 3-5. Distribution of ArfGAP subfamilies across eukaryotic taxa. Five ArfGAP subfamilies, SMAP, ArfGAP1, ArfGAP2, ACAP, AGFG, and the newly identified ArfGAPC2, are found broadly across the diversity of eukaryotes suggesting their presence in the LECA. AGAP is likely present in the opisthokont and amoebozoan ancestor, and is ancient, if not necessarily in the LECA. GIT and ARAP are specific to the Filozoa (Holozoa without *S. arctica*), while ASAP is found in opisthokonts and apusomonads. Large taxonomic groupings are color coded, with taxonomic key below. Numbers in brackets indicate the total number of ArfGAPs identified in the corresponding genome. Sectors with solid colors indicate those homologues identified using the 5-orders criterion. The pale coloured sectors indicate those identified using the 2-orders criterion. Grey sectors indicate that no orthologue was found in the genome of the organism in question, but an orthologue was found in the genome of a closely related organism through nr-BLAST at the 2-orders criterion (see methods). Open sectors indicate that no orthologue was found using BLAST, HMMer or nr-BLASTs. For the ArfGAPC2 row (boxed), the solid colours represent the presence of at least one orthologue meeting a criterion of a bi-directional retrieval of another ArfGAPC2 orthologue at the 5-orders criterion. Purple star indicates the identification of ArfGAPC2 in *Naegleria gruberi* and in *Naegleria fowleri* after completion of this analysis (E. Herman, personal communication).



Vertebrates
Hs = *Homo sapiens (31)*
Rn = *Rattus norvegicus (21)*
Mm = *Mus musculus (24)*
Gg = *Gallus gallus (15)*
Xl = *Xenopus laevis (18)*
Dr = *Danio rerio (30)*

Invertebrates
Ce = *Caenorhabditis elegans (6)*
Dm = *Drosophila melanogaster (8)*
Ci = *Ciona intestinalis (10)*
Nv = *Nematostella vectensis (14)*

Choanozoa
Mb = *Monosiga brevicollis (12)*
Co = *Capsaspora owczarzaki (14)*
Sa = *Sphaeroforma arctica (18)*

Fungi
Rd = *Rhizopus delemar (9)*
Bd = *Batrachochytrium dendrobatidis (5)*
Cn = *Cryptococcus neoformans (5)*
Um = *Ustilago maydis (4)*
Nc = *Neurospora crassa (5)*
Sc = *Saccharomyces cerevisiae (6)*
Sp = *Schizosaccharomyces pombe (6)*

Apusozoa
Ttr = *Thecamonas trahens (15)*

Amoebozoa
Dd = *Dictyostelium discoideum (11)*
Ac = *Acanthamoeba castellanii (14)*

Excavata
Tv = *Trichomonas vaginalis (21)*
Gl = *Giardia lamblia (4)*
Tc = *Trypanosoma cruzi (7)*
Ng = *Naegleria gruberi (6)*

Archaeplastida
At = *Arabidopsis thaliana (21)*
Ot = *Ostreococcus tauri (5)*
Cm = *Cyanidioschyzon merolae (3)*
Cr = *Chlamydomonas reinhardtii (5)*
Ppat = *Physcomitrella patens (18)*

SAR
Pf = *Plasmodium falciparum (2)*
Tg = *Toxoplasma gondii (6)*
Ehux = *Emiliania huxleyi (16)*
Ps = *Phythophthora sojae (9)*
Tp = *Thalassiosira pseudonana (8)*
Tt = *Tetrahymena thermophila (6)*

control to determine whether or not the reciprocal BLAST experiments against the *H. sapiens* and *S. cerevisiae* genomes were able to correctly classify ArfGAP proteins from all eukaryotes. The assumption is that if the subfamily assignments obtained using the methodology described in section 2.2.1 is correct, then these assignments should not change if a different eukaryotic genome is used for reciprocal BLAST experiments. In order to test this assumption bidirectional BLAST searches were performed using the following pairs of organisms: *N. gruberi* and *D. discoideum*, *T. pseudonana* and *A. thaliana*, *P. sojae* and *T. vaginalis*, and *C. reinhardtii* and *A. thaliana*. Sequences meeting both the 5-orders and 2-orders criteria were used. Positive and negative results were tallied for each experiment. A positive result required that the query retrieve the correspondingly assigned orthologue in the target genome at the relevant criterion, (*e.g.* the *N. gruberi* ADAP homologue retrieved the equivalent *D. discoideum* sequence). Of the sequences identified using the 5-orders criterion, 42 of 45 ArfGAP sequences tested (93%) returned a sequence at the 5-orders criterion that was originally classified as the same subfamily using *H. sapiens*. The remaining three sequences assessed did return the appropriate orthologue, but at the 2-orders criterion. All 13 additional sequences identified at the 2-orders criterion retrieved the appropriate orthologue at that criterion. As these criteria were robust and successful for the classification of ArfGAP sequences, they were carried over to the ArfGEF analysis (see section 3.4.2.1).

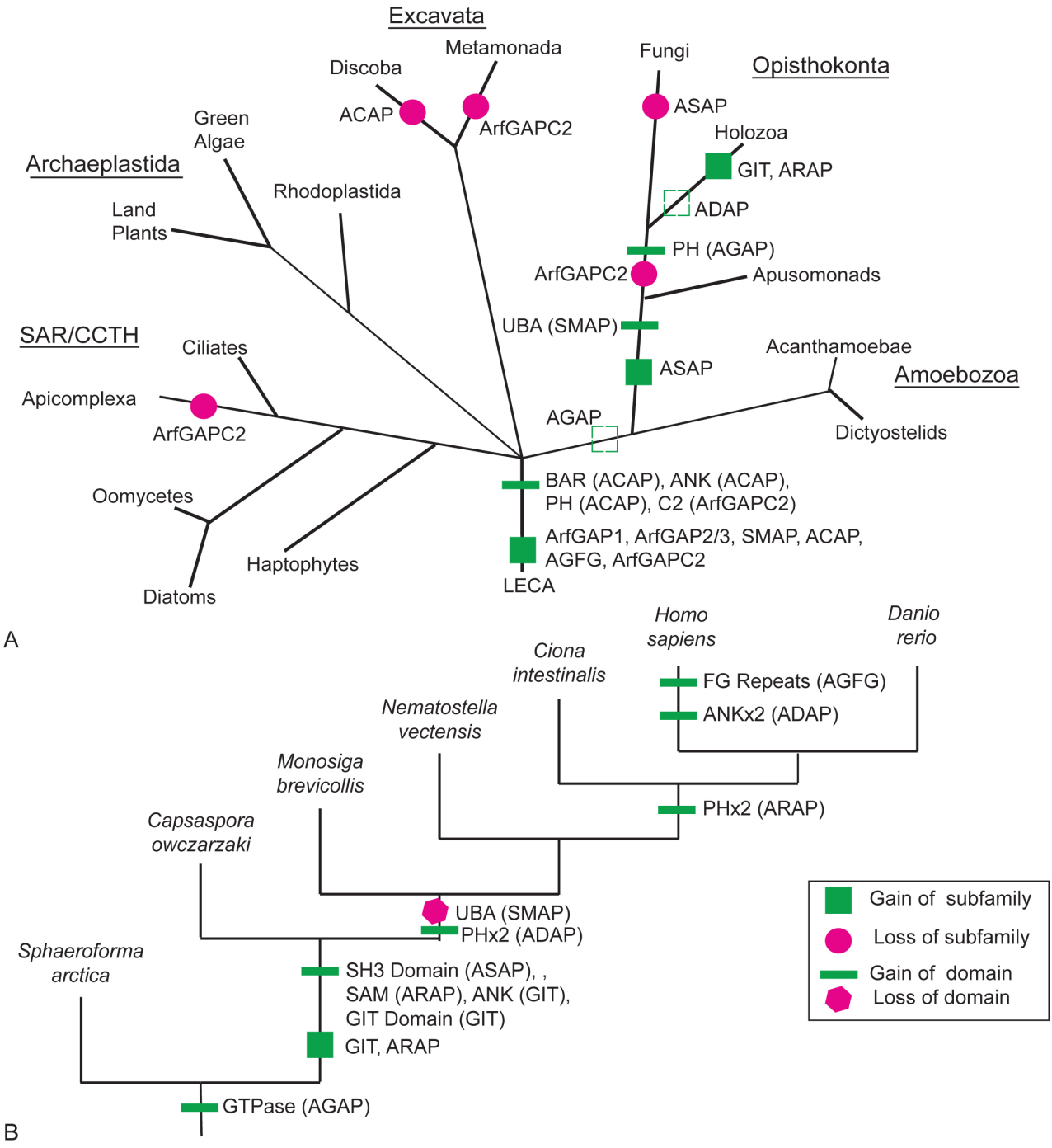
A phylogenetic analysis of all ArfGAP domain-containing proteins was undertaken in an attempt to classify the remaining sequences. No resolution was

obtained, therefore attempts to classify the rogue sequences were abandoned, and these sequences were left as unclassified.

3.4.1.2 BLAST against the NCBI non-redundant database to avoid false negatives in the search for ArfGAP homologues and the identification of five ancient ArfGAP subfamilies

Comparative genomic analysis identified the presence of five of the ten previously identified human ArfGAP subfamilies [SMAP (Small ArfGAP), ArfGAP1, ArfGAP2, ACAP (ArfGAP and coiled-coil domain-containing protein), and AGFG (ArfGAP domain and FG repeat-containing protein)] in four or more supergroups at the 5-orders criterion indicating that they were present in the LECA (Figure 3-5, Figure 3-6). To determine whether our inability to detect certain ArfGAP subfamilies in different eukaryotic genomes was the result of taxon selection, *i.e.,* whether we chose divergent genomes or genomes that have undergone accelerated loss compared to other closely related organisms, we searched for ArfGAP proteins in the NCBI non-redundant (nr)-database at GenBank. BLASTp searches were carried out using *H. sapiens* and *S. cerevisiae* sequences as queries. The NCBI nr-database was restricted to the broadest taxonomic grouping without overlapping with that of another organism already included in the study. Orthology was considered using the 2-orders criterion. By searching this database, we were able to search all of the sequenced genomes available at NCBI. In some cases, orthologues of ArfGAP subfamilies missing from taxa included in our study were found in closely related organisms of the same lineage (grey sectors, Figure 3-5). This identified orthologues of broadly distributed ArfGAP subfamilies in taxa where they were previously

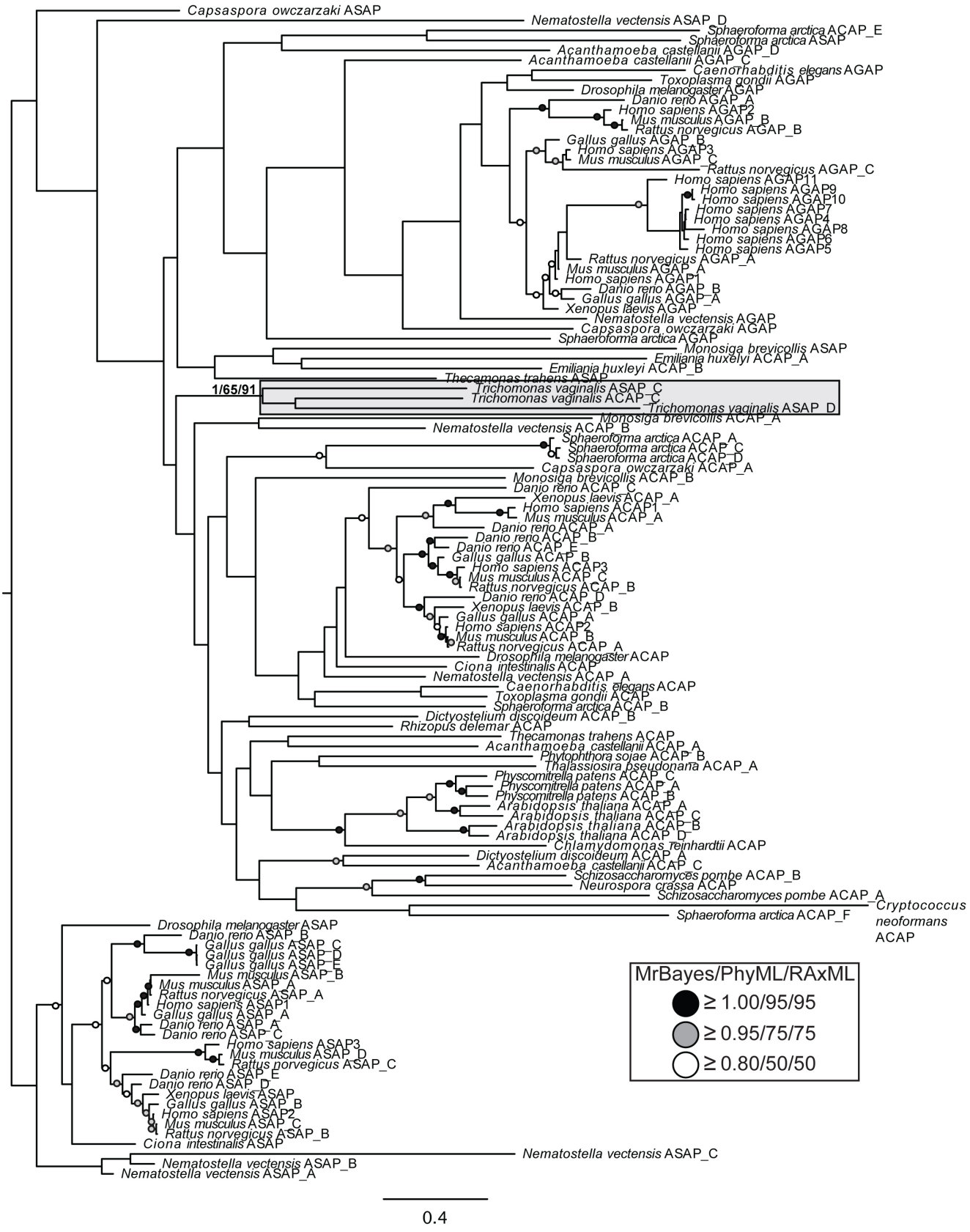
Figure 3-6. Gain and loss of ArfGAP subfamilies and domains in eukaryotes. A) Tree of eukaryotes depicting domains and ArfGAP subfamilies present in the LECA, as well as gains or losses of additional domains and subfamilies throughout eukaryotes. Losses are only proposed when the deduction is based on absence in two genomes of the relevant lineage. The origins of AGAP and ADAP are represented as boxes with dashed lines denoting the minimal distribution of these proteins as comparative genomic analysis suggests that they may be more broadly distributed, as the 2- and 5-orders criteria gave inconsistent results. B) Gain and loss of ArfGAP subfamilies and domains in Holozoa. Symbol legend for both panels is inset in B and the subfamily in which the domain was gained or lost is indicated in brackets. PH = Pleckstrin Homology domain; ANK = Ankyrin Repeat; BAR = Bin/Amphiphysin/Rvs; C2 =calcium dependent membrane-targeting domain; SAM = Sterile alpha motif; SH3 = Src homology-3 domain; GIT = G protein-coupled receptor kinase-interacting protein domain; UBA = ubiquitin associated/translation elongation factor EF1B N-terminal domain (definitions are taken from InterProScan results).


Excavata
Metamonada
Discoba
ACAP
ArfGAPC2
Fungi
Opisthokonta
ASAP
Holozoa
GIT, ARAP
ADAP
PH (AGAP)
ArfGAPC2
Apusomonads
Green
Algae
Archaeplastida
Rhodoplastida
Land
Plants
UBA (SMAP)
ASAP
SAR/CCTH
Ciliates
Acanthamoebae
Amoebozoa
Apicomplexa
ArfGAPC2
AGAP
Dictyostelids
BAR (ACAP), ANK (ACAP),
PH (ACAP), C2 (ArfGAPC2)
Oomycetes
ArfGAP1, ArfGAP2/3, SMAP, ACAP,
AGFG, ArfGAPC2
Haptophytes
Diatoms
LECA
A
Homo
sapiens
Danio
rerio
Ciona
intestinalis
FG Repeats (AGFG)
ANKx2 (ADAP)
Nematostella
vectensis
Monosiga
brevicollis
PHx2 (ARAP)
Capsaspora
owczarzaki
UBA (SMAP)
PHx2 (ADAP)
Sphaeroforma
arctica
SH3 Domain (ASAP), ,
SAM (ARAP), ANK (GIT),
GIT Domain (GIT)
GIT, ARAP
GTPase (AGAP)
B
Gain of subfamily
Loss of subfamily
Gain of domain
Loss of domain

missing, providing additional support for the broad distribution and ancient nature of the SMAP, ArfGAP1, ArfGAP2, ACAP, and AGFG. The distribution of ASAP, ARAP, and GIT did not change after searching the nr-database, indicating that ARAP and GIT are restricted to the Filozoa (*C. owczarzaki, M. brevicollis,* and Metazoa), while ASAP is found only in Holozoa (Filozoa + *S. arctica*) and the apusomonad *T. trahens*, but is absent from fungi (Figure 3-5). The narrow distribution of these three subfamilies suggests that they evolved much more recently and were not present in the LECA.

Several sequences from *T. vaginalis* retrieved ASAP as their top reciprocal BLAST hit in humans, followed by ACAP and AGAP as second and third best hits. To determine whether these sequences are *bona fide* ASAP sequences or whether they are misclassified ACAP or AGAP sequences, phylogenetic analysis of these three subfamilies was undertaken. If these *T. vaginalis* sequences are indeed members of the ASAP subfamily, then they should group with other ASAP sequences with strong support. Similarly, if these are the result of Horizontal Gene Transfer (HGT), then they should group within the ASAP clade next to ASAP sequences from the donor organism with strong support. However, in the phylogenetic analysis, these putative ASAP sequences group with the *T. vaginalis* ACAP sequences with strong support in two of the three methods (Figure 3-7). This result is inconsistent with what has previously been observed for horizontally transferred genes (Archibald et al., 2003; Bergthorsson et al., 2003; discussed in Keeling and Palmer, 2008), indicating that these are ACAP sequences that were mis-classified by BLAST. Because ACAP is an ancient subfamily, and *T. vaginalis* is known to have highly divergent protein

Figure 3-7. Phylogenetic analysis reveals *T. vaginalis* ASAP sequences are divergent ACAPs. In order to determine whether the *T. vaginalis* ASAP sequences are *bona fide* ASAPs, phylogenetic analysis of ACAP, ASAP, and AGAP was carried out. As with other phylogenetic analyses, sequences that meet the 5-orders criterion were used. The *T. vaginalis* ASAP sequences form a moderately supported clade with *T. vaginalis* ACAP. Importantly, the *T. vaginalis* sequences are also excluded from the ASAP clade. Together, this suggests that these are divergent ACAP sequences, not ASAPs. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (MrBayes)/Maximum-Likelihood bootstrap values (PhyML)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95/95, closed light circles ≥ 0.95/75/75, open circles ≥ 0.8/50/50.


Capsaspora owczarzaki ASAP
Nematostella vectensis ASAP_D
Sphaeroforma arctica ACAP_E
Sphaeroforma arctica ASAP
Acanthamoeba castellanii AGAP_D
Acanthamoeba castellanii AGAP_C
Caenorhabditis elegans AGAP
Toxoplasma gondii AGAP
Drosophila melanogaster AGAP
Danio rerio AGAP_A
Homo sapiens AGAP2
Mus musculus AGAP_B
Rattus norvegicus AGAP_B
Gallus gallus AGAP_B
Homo sapiens AGAP3
Mus musculus AGAP_C
Rattus norvegicus AGAP_C
Homo sapiens AGAP11
Homo sapiens AGAP9
Homo sapiens AGAP10
Homo sapiens AGAP7
Homo sapiens AGAP4
Homo sapiens AGAP8
Homo sapiens AGAP6
Homo sapiens AGAP5
Rattus norvegicus AGAP_A
Mus musculus AGAP_A
Homo sapiens AGAP1
Danio rerio AGAP_B
Gallus gallus AGAP_A
Xenopus laevis AGAP
Nematostella vectensis AGAP
Capsaspora owczarzaki AGAP
Sphaeroforma arctica AGAP
Monosiga brevicollis ASAP
Emiliania huxleyi ACAP_A
Emiliania huxleyi ACAP_B
Thecamonas trahens ASAP
1/65/91
Trichomonas vaginalis ASAP_C
Trichomonas vaginalis ACAP_C
Trichomonas vaginalis ASAP_D
Monosiga brevicollis ACAP_A
Nematostella vectensis ACAP_B
Sphaeroforma arctica ACAP_A
Sphaeroforma arctica ACAP_C
Sphaeroforma arctica ACAP_D
Capsaspora owczarzaki ACAP_A
Monosiga brevicollis ACAP_B
Danio rerio ACAP_C
Xenopus laevis ACAP_A
Homo sapiens ACAP1
Mus musculus ACAP_A
Danio rerio ACAP_A
Danio rerio ACAP_B
Danio rerio ACAP_E
Gallus gallus ACAP_B
Homo sapiens ACAP3
Mus musculus ACAP_C
Rattus norvegicus ACAP_B
Danio rerio ACAP_D
Xenopus laevis ACAP_B
Gallus gallus ACAP_A
Homo sapiens ACAP2
Mus musculus ACAP_B
Rattus norvegicus ACAP_A
Drosophila melanogaster ACAP
Ciona intestinalis ACAP
Nematostella vectensis ACAP_A
Caenorhabditis elegans ACAP
Toxoplasma gondii ACAP
Sphaeroforma arctica ACAP_B
Dictyostelium discoideum ACAP_B
Rhizopus delemar ACAP
Thecamonas trahens ACAP
Acanthamoeba castellanii ACAP_A
Phytophthora sojae ACAP_B
Thalassiosira pseudonana ACAP_A
Physcomitrella patens ACAP_C
Physcomitrella patens ACAP_A
Physcomitrella patens ACAP_B
Arabidopsis thaliana ACAP_A
Arabidopsis thaliana ACAP_C
Arabidopsis thaliana ACAP_B
Arabidopsis thaliana ACAP_D
Chlamydomonas reinhardtii ACAP
Dictyostelium discoideum ACAP_A
Acanthamoeba castellanii ACAP_C
Schizosaccharomyces pombe ACAP_B
Neurospora crassa ACAP
Schizosaccharomyces pombe ACAP_A
Cryptococcus neoformans ACAP
Sphaeroforma arctica ACAP_F
Drosophila melanogaster ASAP
Danio rerio ASAP_B
Gallus gallus ASAP_C
Gallus gallus ASAP_D
Gallus gallus ASAP_E
Mus musculus ASAP_B
Mus musculus ASAP_A
Rattus norvegicus ASAP_A
Homo sapiens ASAP1
Gallus gallus ASAP_A
Danio rerio ASAP_A
Danio rerio ASAP_C
Homo sapiens ASAP3
Mus musculus ASAP_D
Rattus norvegicus ASAP_C
Danio rerio ASAP_E
Danio rerio ASAP_D
Xenopus laevis ASAP
Gallus gallus ASAP_B
Homo sapiens ASAP2
Mus musculus ASAP_C
Rattus norvegicus ASAP_B
Ciona intestinalis ASAP
Nematostella vectensis ASAP_C
Nematostella vectensis ASAP_B
Nematostella vectensis ASAP_A
MrBayes/PhyML/RAxML
≥ 1.00/95/95
≥ 0.95/75/75
≥ 0.80/50/50
0.4

sequences, ASAP is presumed to be a recently diverged subfamily with a likely origin in the ancestor of apusomonads and opisthokonts.

ADAP and AGAP had distributions that prevented the proposal of specific origins. Orthologues of AGAP were identified in two supergroups (Opisthokonta and Amoebozoa), suggesting an origin in an ancient ancestor, that may or may not represent the LECA, depending on the position of the eukaryotic root. Additional sequences were identified from at least two other supergroups (Archaeplastida and SAR/CCTH), but at the 2-orders criterion (Figure 3-5). Although AGAP clearly has an ancient origin, it does not meet the requirements to be considered present in the LECA. The final subfamily, ADAP, had orthologues that satisfied the 5-orders criterion only in some members of the Opisthokonta (Figure 3-5). If the 2-orders criterion is applied, then putative ADAP sequences are also found in the Amoebozoa and the Excavata. However, there is insufficient data to confidently assign origins for the ADAP and AGAP subfamilies.

Lineage-specific loss is apparent in many ArfGAP subfamilies, and is especially prevalent in fungi, where AGFG is only present in the basal fungus *Batrachochytrium dendrobatidis*, and AGAP is only present in the zygomycete *Rhizopus delemar*. Orthologues were not detected for either ASAP or ADAP. Similarly, failure to identify ADAP homologues in *C. elegans* or *D. melanogaster* also likely reflects independent loss, even though ADAPs were identified in closely related organisms (grey sectors, Figure 3-5).

3.4.1.3 Identification of ArfGAPC2, an undescribed ancient ArfGAP subfamily

Finding that many ArfGAP sequences (*i.e.,* rogues) could not be classified into one of the ten previously described subfamilies was a surprising result. One explanation for the large number of rogue sequences could be a problem of asymmetry; some of the rogues may be representative of additional subfamilies not present in humans, and therefore have no single best hit in the human genome. To assess this possibility, each rogue sequence was used as a query to search other genomes containing at least one rogue. It would be expected that sequences belonging to an undescribed ArfGAP subfamily would preferentially retrieve one another as their top BLAST hits. For the majority of rogue ArfGAP sequences, no best hits meeting either RBH criteria were found. However, for six rogue sequences, reciprocal best hits were other rogues, satisfying the 5-orders criterion (Figure 3-8). Analysis of domain composition identified shared architectures: an ArfGAP domain followed by a Ca^{2+}-dependent membrane-targeting (C2) domain, strongly suggesting orthology (Figure 3-9). Based on the domain composition, this new subfamily was designated ArfGAPC2.

Reciprocal BLAST of some ArfGAPC2 sequences into *A. thaliana* retrieved four SMAP sequences in addition to the *A. thaliana* ArfGAPC2 sequence. Upon further examination, these SMAP sequences also possessed the C-terminal C2 domain. When taken as a group, these sequences not only meet the 5-orders criterion, but are 23 orders of magnitude better than sequences of any other subfamily. Re-examination of the domain composition of all ArfGAP sequences identified one SMAP sequence from *P. patens,* two ACAP sequences from *E. huxleyi*,

Figure 3-8. Reciprocal retrieval of putative ArfGAPC2 orthologues by BLASTp. ArfGAPC2 sequences are grouped by species. Delta E-value symbols indicate the difference in orders of magnitude between the top hit (the sequence pointed to by the arrow) and the first non-rogue (non-ArfGAPC2) sequence. Thick lines indicate a difference of five orders of magnitude or greater; the dashed lines indicate a difference of less than five orders of magnitude, but that still retrieved another ArfGAPC2 as the top hit. AGC2 = ArfGAPC2; At = *A. thaliana*; Eh = *E. huxleyi*; Pp = *P. patens*; Ps = *Phytophthora sojae*; Ttr = *T. trahens*.

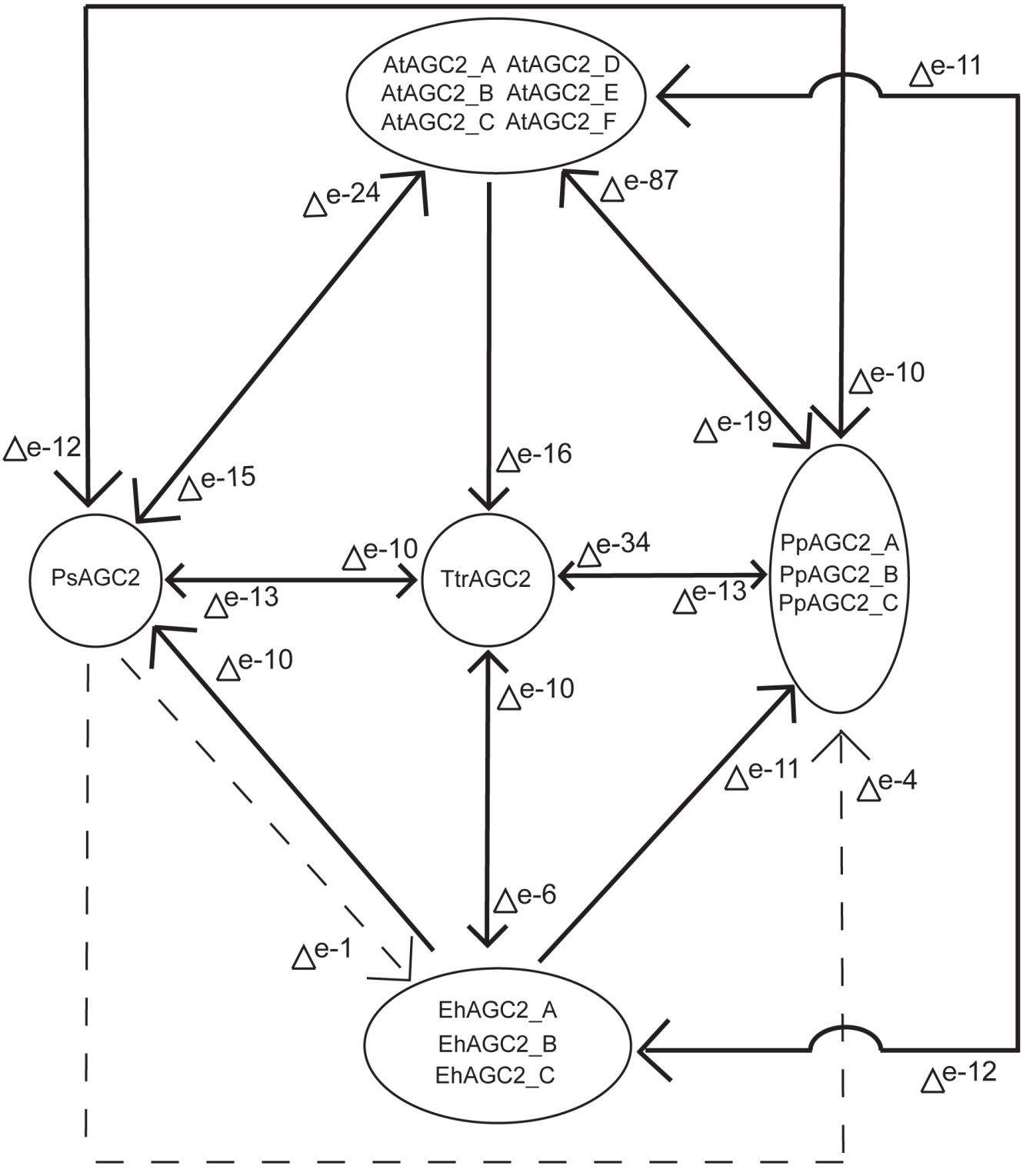
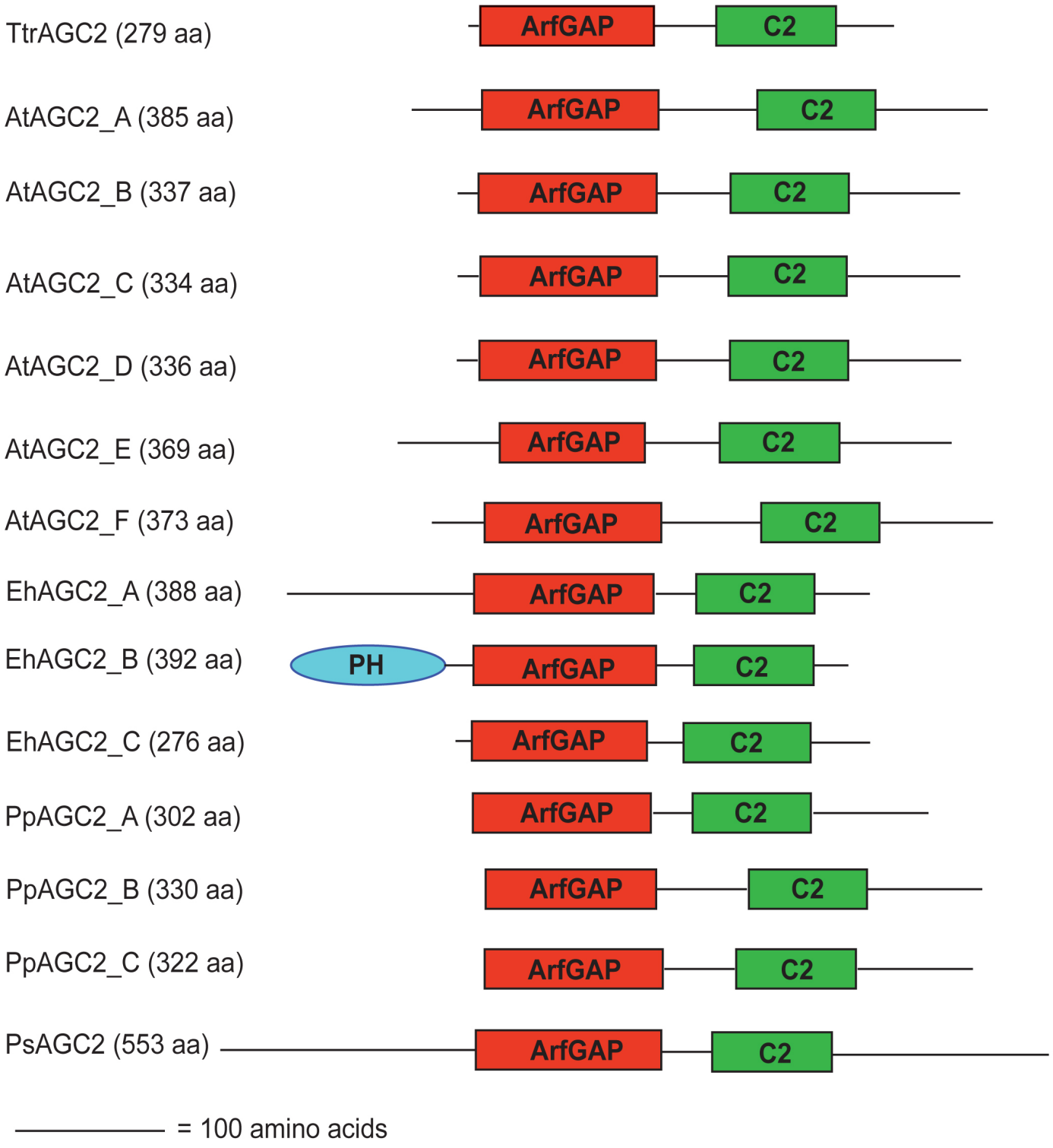

AtAGC2_A AtAGC2_D
AtAGC2_B AtAGC2_E
AtAGC2_C AtAGC2_F
PsAGC2
TtrAGC2
PpAGC2_A
PpAGC2_B
PpAGC2_C
EhAGC2_A
EhAGC2_B
EhAGC2_C
△e-11
△e-24
△e-87
△e-10
△e-19
△e-12
△e-15
△e-16
△e-10
△e-34
△e-13
△e-13
△e-10
△e-10
△e-11
△e-4
△e-6
△e-1
△e-12

Figure 3-9. Domain composition of the ArfGAPC2 subfamily. Domain organization of each ArfGAPC2 subfamily member is illustrated as determined by InterProScan. Each sequence contains an ArfGAP domain followed by a C2 domain. Sequences are drawn to scale. ArfGAP = ArfGAP domain; C2 = Calcium Dependent Membrane-Targeting; PH = Pleckstrin Homology; AGC2 = ArfGAPC2; At = *A. thaliana*; Eh = *E. huxleyi*; Pp = *P. patens*; Ps = *P. sojae*; Ttr = *T. trahens*.

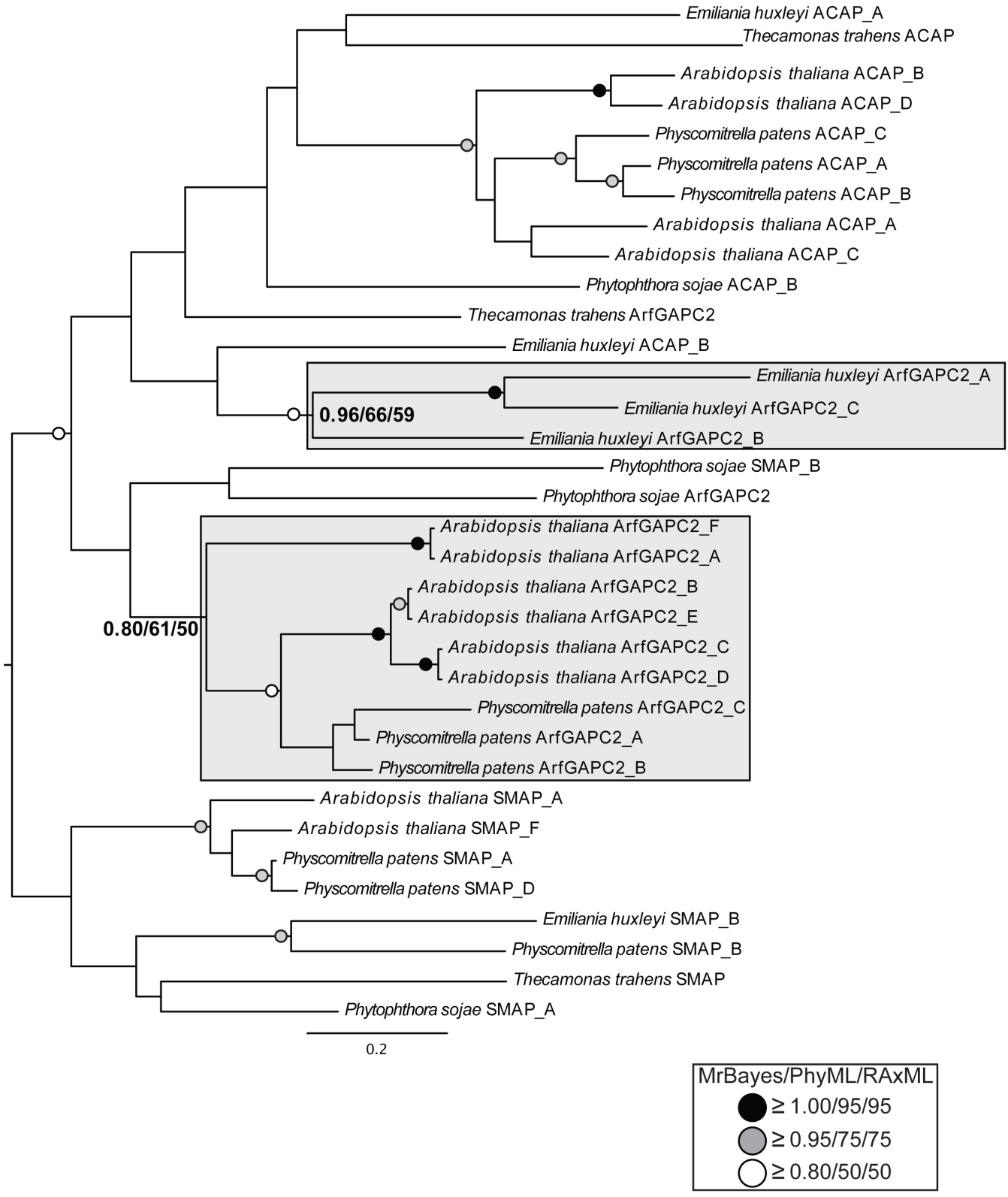

TtrAGC2 (279 aa)
ArfGAP
C2
AtAGC2_A (385 aa)
ArfGAP
C2
AtAGC2_B (337 aa)
ArfGAP
C2
AtAGC2_C (334 aa)
ArfGAP
C2
AtAGC2_D (336 aa)
ArfGAP
C2
AtAGC2_E (369 aa)
ArfGAP
C2
AtAGC2_F (373 aa)
ArfGAP
C2
EhAGC2_A (388 aa)
ArfGAP
C2
EhAGC2_B (392 aa)
PH
ArfGAP
C2
EhAGC2_C (276 aa)
ArfGAP
C2
PpAGC2_A (302 aa)
ArfGAP
C2
PpAGC2_B (330 aa)
ArfGAP
C2
PpAGC2_C (322 aa)
ArfGAP
C2
PsAGC2 (553 aa)
ArfGAP
C2
= 100 amino acids

and one additional rogue sequence from *A. thaliana* all containing the ArfGAPC2 domain architecture. BLAST analysis of these sequences confirmed that they did in fact retrieve ArfGAPC2 as their best hit at the 5-orders criterion. Moreover, phylogenetic analysis of all SMAP, ACAP, and ArfGAPC2 sequences from the taxa in question revealed moderate support for the unification of plant ArfGAPC2 clade (0.80/61/50), as well as for the clade formed by the three *E. huxleyi* sequences (0.96/66/59; Figure 3-10). The presence of sequences satisfying the 5-orders criterion present in four major lineages (archaeplastids, apusomonads, stramenopiles, and haptophytes), combined with the recent identification of ArfGAPC2 homologues in the genome of *N. gruberi* and *N. fowleri* (E. Herman, personal communication) strongly points to the presence of this subfamily in the LECA, making it the sixth ancient ArfGAP subfamily.

3.4.1.4 Domain evolution: ArfGAPs reflect plasticity and lineage-specific tailoring

In addition to the catalytic domains, ArfGAPs possess a variety of accessory domains that are important for functionality. For example, the PH domain of ASAP has been shown to be a positive regulator of the ArfGAP domain, as its deletion results in reduced GAP activity (Kam et al., 2000). The PH domains of ADAP are necessary for association with specific membrane lipids such as phosphatidylinositol (3,4,5)-trisphosphate (Venkateswarlu and Cullen, 1999; Venkateswarlu et al., 2004). The RhoGAP domain of ARAP is the basis for interaction with Rho proteins, providing crosstalk between Arf and Rho signalling cascades

Figure 3-10. ArfGAPC2 forms distinct clades from SMAP and ACAP. To determine whether all ArfGAPC2 sequences form a single subfamily, phylogenetic analysis of SMAP, ACAP, and all ArfGAP proteins containing a C2 domain was carried out. The analysis reveals that the C2 domain containing sequences from Archaeplastida (*A. thaliana* and *P. patens*) and *E. huxleyi* cluster with other ArfGAPC2 subfamily members. Although they do not form a single group to the exclusion of all other sequences, this topology cannot be ruled out because of the general lack of node support across the tree. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (MrBayes)/Maximum-Likelihood bootstrap values (PhyML)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95/95, closed light circles ≥ 0.95/75/75, open circles ≥ 0.8/50/50.

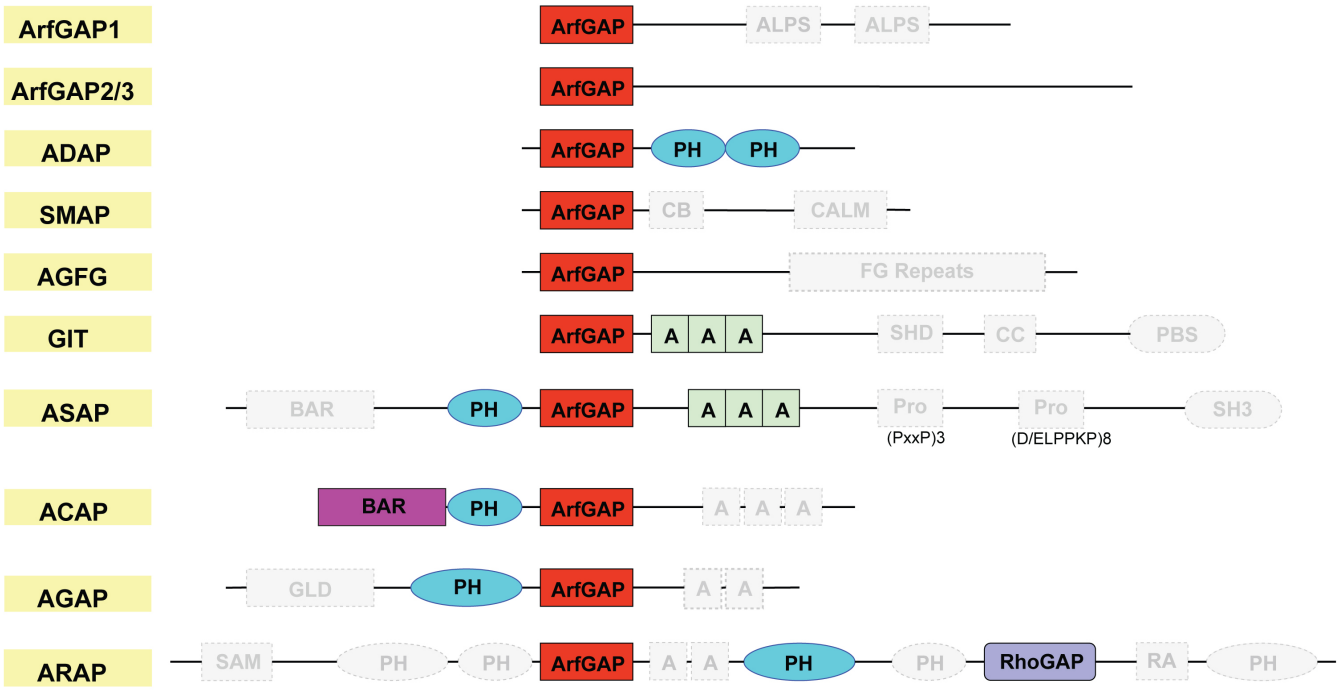

Emiliania huxleyi ACAP_A
Thecamonas trahens ACAP
Arabidopsis thaliana ACAP_B
Arabidopsis thaliana ACAP_D
Physcomitrella patens ACAP_C
Physcomitrella patens ACAP_A
Physcomitrella patens ACAP_B
Arabidopsis thaliana ACAP_A
Arabidopsis thaliana ACAP_C
Phytophthora sojae ACAP_B
Thecamonas trahens ArfGAPC2
Emiliania huxleyi ACAP_B
Emiliania huxleyi ArfGAPC2_A
0.96/66/59
Emiliania huxleyi ArfGAPC2_C
Emiliania huxleyi ArfGAPC2_B
Phytophthora sojae SMAP_B
Phytophthora sojae ArfGAPC2
Arabidopsis thaliana ArfGAPC2_F
Arabidopsis thaliana ArfGAPC2_A
Arabidopsis thaliana ArfGAPC2_B
0.80/61/50
Arabidopsis thaliana ArfGAPC2_E
Arabidopsis thaliana ArfGAPC2_C
Arabidopsis thaliana ArfGAPC2_D
Physcomitrella patens ArfGAPC2_C
Physcomitrella patens ArfGAPC2_A
Physcomitrella patens ArfGAPC2_B
Arabidopsis thaliana SMAP_A
Arabidopsis thaliana SMAP_F
Physcomitrella patens SMAP_A
Physcomitrella patens SMAP_D
Emiliania huxleyi SMAP_B
Physcomitrella patens SMAP_B
Thecamonas trahens SMAP
Phytophthora sojae SMAP_A
0.2
MrBayes/PhyML/RAxML
≥ 1.00/95/95
≥ 0.95/75/75
≥ 0.80/50/50

(Krugmann et al., 2004; Yoon et al., 2006). Overall, accessory domains are important for the function and regulation of ArfGAPs.

To assess the evolution of accessory domains in ArfGAP proteins, diagnostic domain structures were established using sequences classified at the 5-orders criterion. Domain profiles were created for each subfamily by identifying the domains present in each sequence using InterProScan (Figure 3-11; Hunter et al., 2009; Mitchell et al., 2014) with all 14 algorithms available for domain recognition selected. Presence of a domain in at least 85% of a given subfamily was set as a criterion for that domain to be conserved in that subfamily. This was found to be the most stringent criterion that still retained a complete ArfGAP domain.

After bioinformatic identification, domain profiles emerged for each subfamily. Surprisingly, only BAR, PH, ankyrin repeats, and the C2 domains are conserved across eukaryotes suggesting that these domains became associated with the ArfGAP domain prior to the LECA (Figure 3-6, 3-9, 3-11). Assessing the domain architecture of ArfGAPs at different evolutionary time points allows us to reconstruct the stepwise acquisition of domain complexity found in humans (Figure 3-6). Many of the domains that define the human subfamilies are not widely conserved, and only appear relatively recently. For example, the GTPase domain of AGAP does not appear until the divergence of Holozoa from fungi (Figure 3-6; Xia et al., 2003), even though AGAP appears in the ancestor of opisthokonts and Amoebozoa (Figure 3-5). Similarly, the SH3 domain that defines the ASAP subfamily (Brown et al., 1998) does not appear until after the divergence of *S. arctica* from the

Figure 3-11. Conservation of ArfGAP accessory domains. Conserved domains of each ArfGAP subfamily as defined by this study are shown in color. These represent the configurations likely found in the ancestral sequence of each subfamily. Domains identified in humans as defined by Kahn et al., (2008), but not conserved are shown in grey and bound by a dashed outline. Only the BAR domains, PH domains, the ArfGAP domain and ankyrin repeats are conserved across eukaryotes. The RhoGAP domain of ARAP is highly conserved in all ARAP sequences but the ARAP subfamily is only present in the Holozoa. Therefore, the RhoGAP domain is a defining feature of ARAP sequences, although it is present in only a limited set of eukaryotic organisms. ArfGAP = ArfGAP domain; ALPS = Amphipathic Lipid Packing Sensor; CB = Clathrin-Box; CALM = CALM binding domain; SHD = Spa-homology domain; CC = Coiled-coil; PBS = Paxillin Binding Site; BAR = Bin/Amphiphysin/Rvs; PH = Pleckstrin Homology; Pro = Proline rich regions (motifs and number of repeats are indicated below each region); SH3 = Src homology-3 domain; GLD = GTPase-like domain; SAM = Sterile alpha motif; RhoGAP = RhoGAP domain; RA = Ras-association. Modified from Kahn et al., (2008).


ArfGAP1
ArfGAP
ALPS
ALPS
ArfGAP2/3
ArfGAP
ADAP
ArfGAP
PH
PH
SMAP
ArfGAP
CB
CALM
AGFG
ArfGAP
FG Repeats
GIT
ArfGAP
A
A
A
SHD
CC
PBS
ASAP
BAR
PH
ArfGAP
A
A
A
Pro
(PxxP)3
Pro
(D/ELPPKP)8
SH3
ACAP
BAR
PH
ArfGAP
A
A
A
AGAP
GLD
PH
ArfGAP
A
A
ARAP
SAM
PH
PH
ArfGAP
A
A
PH
PH
RhoGAP
RA
PH

rest of the Holozoa (Figure 3-6), even though ASAP was present in the ancestor of opisthokonts and apusomonads (Figure 3-5).

3.4.1.5 Phylogenetic analysis of ArfGAP subfamilies suggests the presence of six ArfGAP domain-containing proteins in the LECA, and reveals coordinated duplication with Arf GTPase expansion

The analysis above identified the presence of at least six ArfGAP subfamilies in the LECA. However, it does not answer the question of how many paralogues of each subfamily were present, nor does it bring to light the evolutionary histories of each subfamily. In humans, most ArfGAP subfamilies possess multiple paralogues; it remains unknown exactly when these duplication occurred. To address these questions, phylogenetic analysis of each of the ten previously described ArfGAP subfamilies was carried out.

The analyses revealed expansions in non-opisthokont lineages, notably of ArfGAP2, ACAP, and AGFG in archaeplastids (see below). None of the analyses performed suggested the presence of multiple paralogues in the LECA, as judged by the presence of multiple clades of more than one supergroup. The most striking result was that, with the exception of ArfGAP1, all ArfGAP subfamilies have undergone one or more gene duplications near the vertebrate transition, giving rise to two or more paralogues. Below is a description of the results of the phylogenetic analysis for each subfamily.

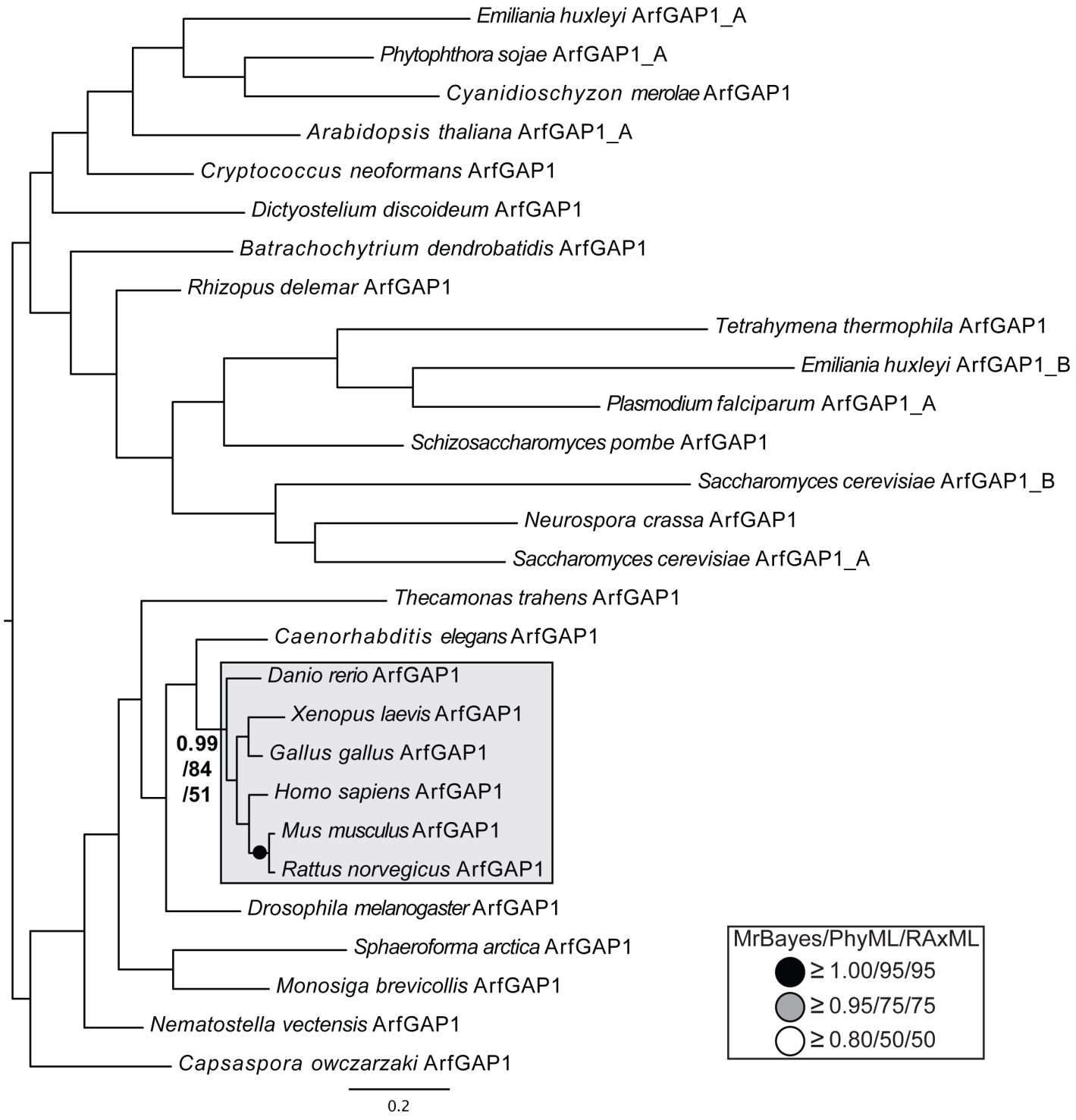
3.4.1.5.1 ArfGAP1

ArfGAP1 localizes to the Golgi apparatus and regulates Golgi-to-ER retrograde trafficking of COPI, intra-Golgi trafficking, and interacts with AP-1 and AP-2 (Rawet et al., 2010). ArfGAP1 does not possess any functional domains outside of the central ArfGAP domain (Kahn et al., 2008). ArfGAP1 possesses two ALPS motifs that recognize membrane curvature and are necessary for its recruitment to the Golgi (Bigay et al., 2005; Mesmin et al., 2007). ArfGAP1 was originally thought to be the founding member of the ArfGAP family, and therefore likely to be ubiquitously distributed (Kahn et al., 2008). A single ArfGAP1 paralogue is found in humans. The ArfGAP1 tree is likely the simplest and most straightforward to interpret: no duplications have occurred in any taxa satisfying the 5-orders criterion, indicating that only a single paralogue was likely to have been present in the LECA (Figure 3-12).

3.4.1.5.2 ArfGAP2

Two members of this subfamily, ArfGAP2 and ArfGAP3, also function in the regulation of COPI coat formation in humans (Kartberg et al., 2010). Neither ArfGAP2, nor ArfGAP3 possess any additional functional domains, and neither possesses ALPS motifs. Rather than being recruited to Golgi membranes by lipid curvature, as is the case for ArfGAP1, ArfGAP2 and ArfGAP3 are recruited via interaction with the COPI complex (Weimer et al., 2008). Although ArfGAP2 and ArfGAP3 appear to play redundant roles in COPI trafficking, ArfGAP3 has recently been shown to play a role in trafficking from early to late endosomes (Shiba et al.,

Figure 3-12. Phylogenetic analysis of ArfGAP1 identifies a single paralogue present in the LECA. A single clade from each supergroup was reconstructed, indicating that no major gene duplications have occurred for ArfGAP1 and that a single ArfGAP1 sequence was present in the LECA. No lineage-specific duplications are apparent in the sequences classified at the 5-orders criterion. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (MrBayes)/Maximum-Likelihood bootstrap values (PhyML)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95/95, closed light circles ≥ 0.95/75/75, open circles ≥ 0.8/50/50.


Emiliania huxleyi ArfGAP1_A
Phytophthora sojae ArfGAP1_A
Cyanidioschyzon merolae ArfGAP1
Arabidopsis thaliana ArfGAP1_A
Cryptococcus neoformans ArfGAP1
Dictyostelium discoideum ArfGAP1
Batrachochytrium dendrobatidis ArfGAP1
Rhizopus delemar ArfGAP1
Tetrahymena thermophila ArfGAP1
Emiliania huxleyi ArfGAP1_B
Plasmodium falciparum ArfGAP1_A
Schizosaccharomyces pombe ArfGAP1
Saccharomyces cerevisiae ArfGAP1_B
Neurospora crassa ArfGAP1
Saccharomyces cerevisiae ArfGAP1_A
Thecamonas trahens ArfGAP1
Caenorhabditis elegans ArfGAP1
Danio rerio ArfGAP1
Xenopus laevis ArfGAP1
0.99
/84
/51
Gallus gallus ArfGAP1
Homo sapiens ArfGAP1
Mus musculus ArfGAP1
Rattus norvegicus ArfGAP1
Drosophila melanogaster ArfGAP1
Sphaeroforma arctica ArfGAP1
Monosiga brevicollis ArfGAP1
Nematostella vectensis ArfGAP1
Capsaspora owczarzaki ArfGAP1
0.2
MrBayes/PhyML/RAxML
≥ 1.00/95/95
≥ 0.95/75/75
≥ 0.80/50/50

2013). Like ArfGAP1, the ArfGAP2 subfamily is predicted to be broadly distributed. Pan-eukaryotic phylogenetic analysis of ArfGAP2 sequences shows an expansion in archaeplastids (Figure 3-13), as denoted by the sequences of *P. patens* and *A. thaliana*. Two paralogues of ArfGAP2, ArfGAP2 and ArfGAP3, are present in vertebrates, indicated by a well-supported ArfGAP2 clade and a weakly supported ArfGAP3 clade, both nested within a moderately supported vertebrate clade. To increase the signal to noise ratio, the analysis was repeated only using sequences from the Filozoa. This analysis more clearly resolves the gene duplication in vertebrates that gave rise to ArfGAP2 and ArfGAP3 (Figure 3-14).

3.4.1.5.3 ACAP

Three ACAP paralogues are found in humans. ACAPs possess an N-terminal BAR domain, followed by a PH domain, the ArfGAP domain, and ankyrin repeats (Jackson et al., 2000). ACAPs were identified as plasma membrane localized Arf6 GAPs that regulate actin remodelling, cell movement, and clathrin-dependent endocytosis (Dias et al., 2013; Li et al., 2007). ACAP homologues have previously been identified in metazoans and amoebozoans, indicating that they are found in at least two supergroups (Kahn et al., 2008). Pan-eukaryotic phylogenetic analysis of ACAP sequences reveals multiple expansions of ACAP in archaeplastids (Figure 3-15). Three well-supported paralogous clades are present in vertebrates, indicating that at least two gene duplications occurred in this lineage. An additional sequence from *Danio rerio* (DrACAP_C) grouped basal to the rest of vertebrates. Removal of all non-metazoan sequences did not result in relocation of the *D. rerio* sequence, but

Figure 3-13. Phylogenetic analysis of ArfGAP2 identifies a single paralogue present in the LECA. A single clade of each supergroup is reconstructed indicating that a single ArfGAP2 sequence was present in the LECA. A well-supported vertebrate clade encompassing ArfGAP2 and ArfGAP3 indicates that ArfGAP3 is a vertebrate-specific ArfGAP (lower grey box). A moderately supported expansion in *P. patens* and *A. thaliana* is also observed (upper grey box). The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (MrBayes)/Maximum-Likelihood bootstrap values (PhyML)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95/95, closed light circles ≥ 0.95/75/75, open circles ≥ 0.8/50/50.

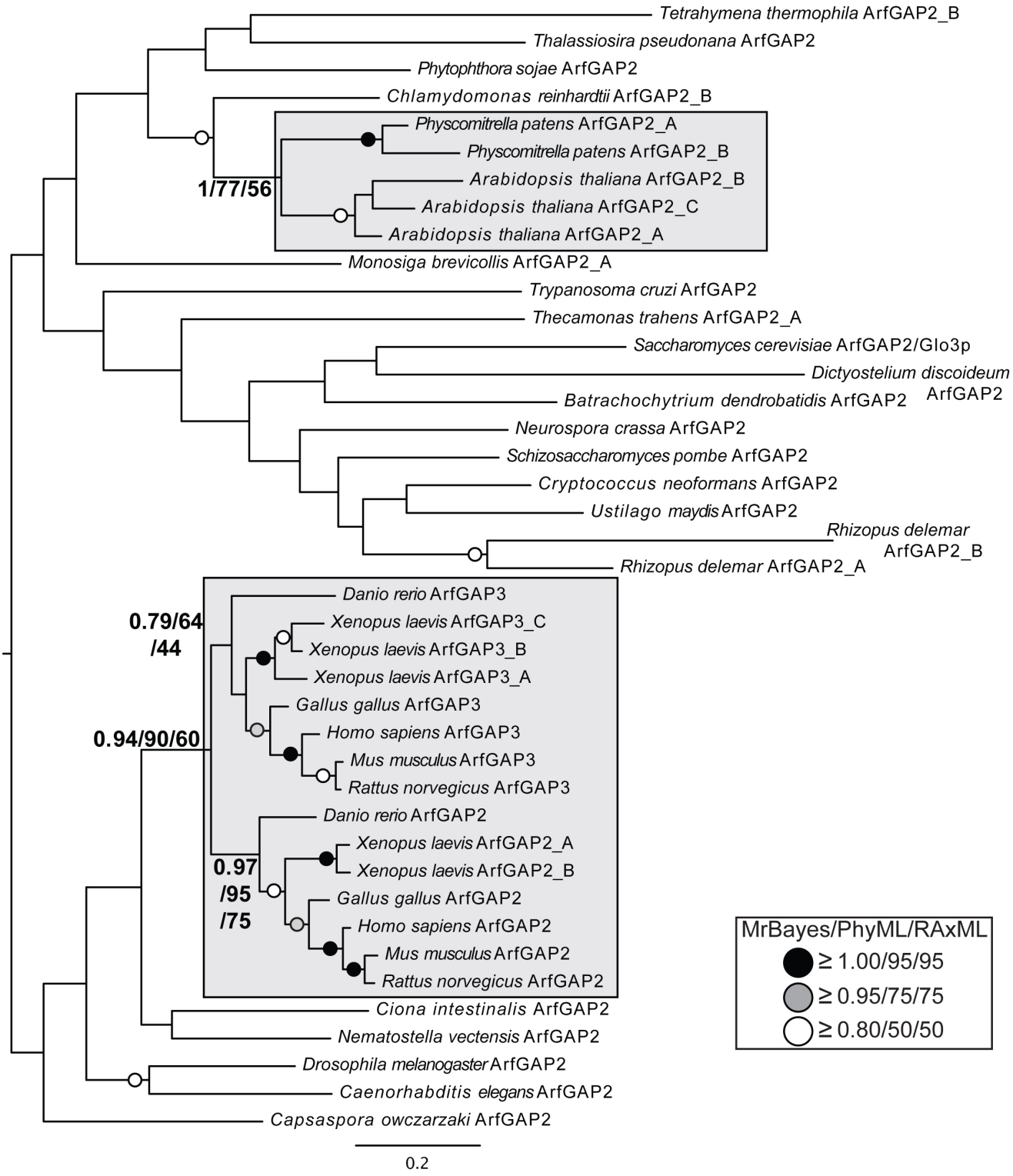

Tetrahymena thermophila ArfGAP2_B
Thalassiosira pseudonana ArfGAP2
Phytophthora sojae ArfGAP2
Chlamydomonas reinhardtii ArfGAP2_B
Physcomitrella patens ArfGAP2_A
Physcomitrella patens ArfGAP2_B
Arabidopsis thaliana ArfGAP2_B
Arabidopsis thaliana ArfGAP2_C
Arabidopsis thaliana ArfGAP2_A
1/77/56
Monosiga brevicollis ArfGAP2_A
Trypanosoma cruzi ArfGAP2
Thecamonas trahens ArfGAP2_A
Saccharomyces cerevisiae ArfGAP2/Glo3p
Dictyostelium discoideum ArfGAP2
Batrachochytrium dendrobatidis ArfGAP2
Neurospora crassa ArfGAP2
Schizosaccharomyces pombe ArfGAP2
Cryptococcus neoformans ArfGAP2
Ustilago maydis ArfGAP2
Rhizopus delemar ArfGAP2_B
Rhizopus delemar ArfGAP2_A
Danio rerio ArfGAP3
0.79/64/44
Xenopus laevis ArfGAP3_C
Xenopus laevis ArfGAP3_B
Xenopus laevis ArfGAP3_A
Gallus gallus ArfGAP3
0.94/90/60
Homo sapiens ArfGAP3
Mus musculus ArfGAP3
Rattus norvegicus ArfGAP3
Danio rerio ArfGAP2
Xenopus laevis ArfGAP2_A
0.97/95/75
Xenopus laevis ArfGAP2_B
Gallus gallus ArfGAP2
Homo sapiens ArfGAP2
Mus musculus ArfGAP2
Rattus norvegicus ArfGAP2
Ciona intestinalis ArfGAP2
Nematostella vectensis ArfGAP2
Drosophila melanogaster ArfGAP2
Caenorhabditis elegans ArfGAP2
Capsaspora owczarzaki ArfGAP2
0.2
MrBayes/PhyML/RAxML
≥ 1.00/95/95
≥ 0.95/75/75
≥ 0.80/50/50

Figure 3-14. Phylogenetic analysis of ArfGAP2 in the Filozoa reveals vertebrate origin of ArfGAP3. In order to confirm that the gene duplication producing ArfGAP2 and ArfGAP3 is vertebrate-specific, all non-filozoan sequences from Figure 3-12 were removed. Overall support increased for the vertebrate ArfGAP2 clade, the ArfGAP3 clade, and for the node grouping these to the exclusion of the other sequences. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (MrBayes)/Maximum-Likelihood bootstrap values (PhyML)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95/95, closed light circles ≥ 0.95/75/75, open circles ≥ 0.8/50/50.

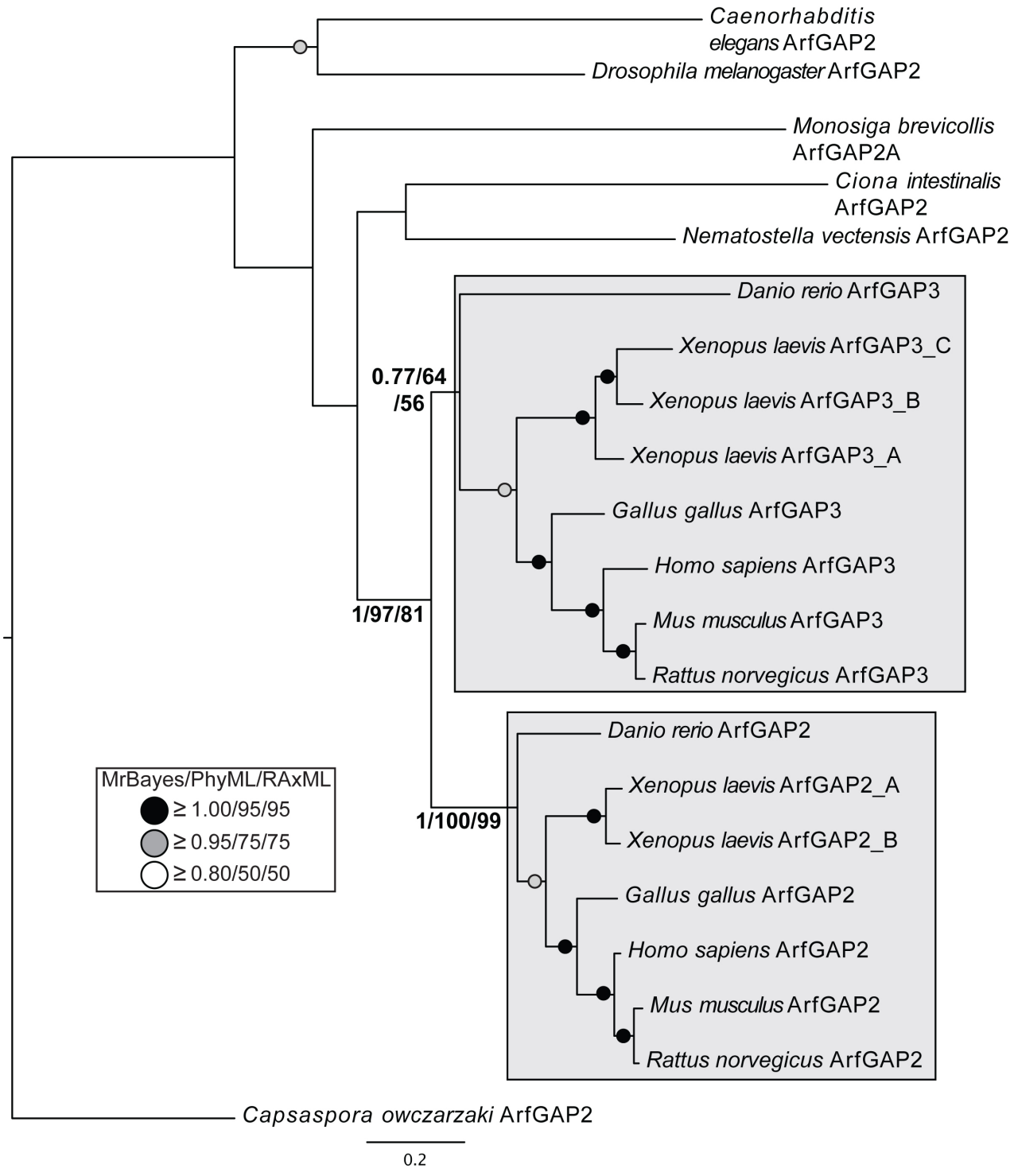
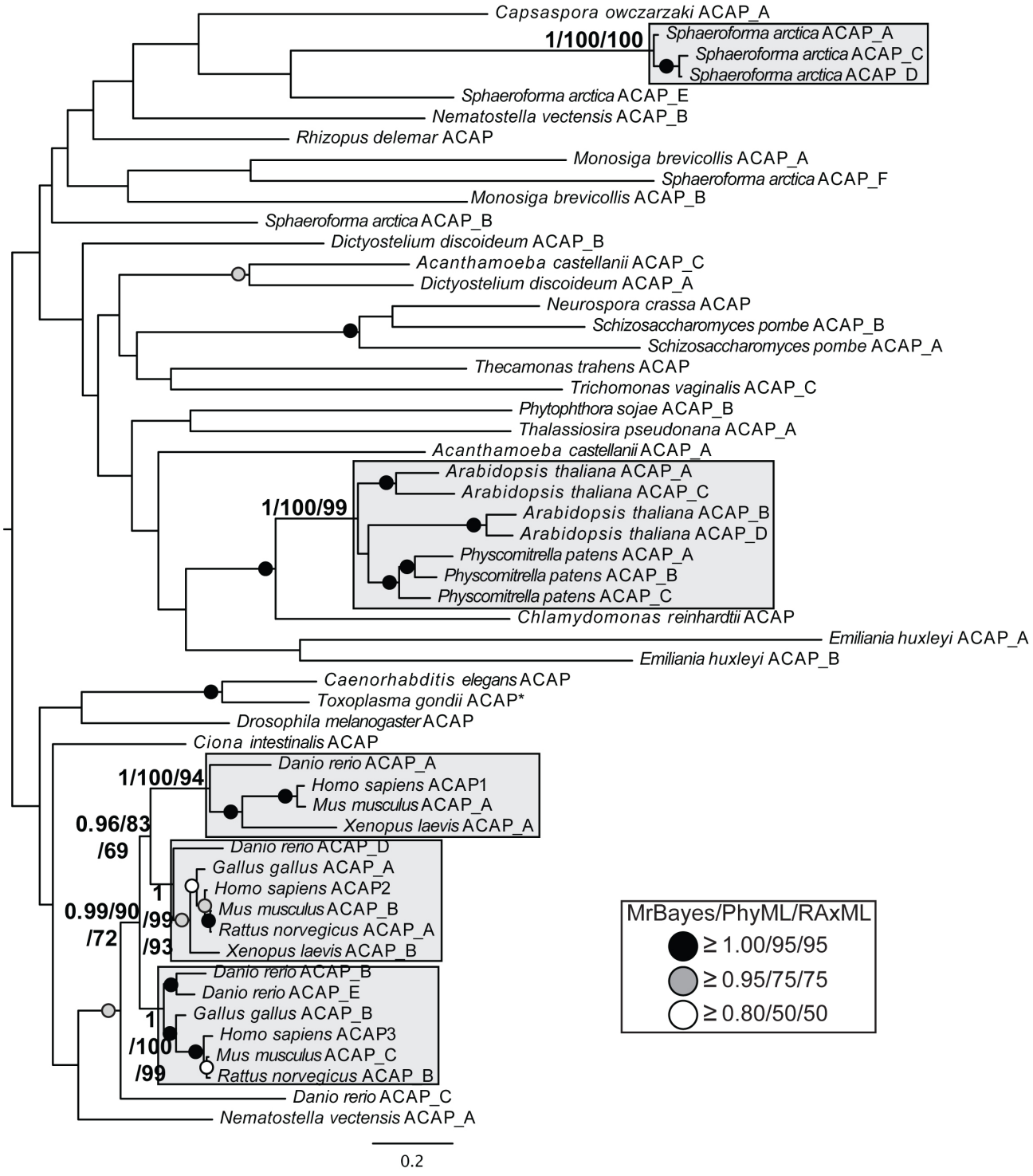

Caenorhabditis elegans ArfGAP2
Drosophila melanogaster ArfGAP2
Monosiga brevicollis ArfGAP2A
Ciona intestinalis ArfGAP2
Nematostella vectensis ArfGAP2
Danio rerio ArfGAP3
Xenopus laevis ArfGAP3_C
Xenopus laevis ArfGAP3_B
Xenopus laevis ArfGAP3_A
Gallus gallus ArfGAP3
Homo sapiens ArfGAP3
Mus musculus ArfGAP3
Rattus norvegicus ArfGAP3
0.77/64/56
1/97/81
Danio rerio ArfGAP2
Xenopus laevis ArfGAP2_A
Xenopus laevis ArfGAP2_B
Gallus gallus ArfGAP2
Homo sapiens ArfGAP2
Mus musculus ArfGAP2
Rattus norvegicus ArfGAP2
1/100/99
MrBayes/PhyML/RAxML
≥ 1.00/95/95
≥ 0.95/75/75
≥ 0.80/50/50
Capsaspora owczarzaki ArfGAP2
0.2

Figure 3-15. Phylogenetic analysis of ACAP identifies a single paralogue in the LECA. Largely, a single clade of each supergroup is reconstructed, indicating that only one ACAP sequence was present in the LECA. Three vertebrate ACAP clades are moderately supported (lower three grey boxes). Independent expansions of ACAP have occurred in Archaeplastida (middle grey box) and *S. arctica* (upper grey box). * Denotes *T. gondii* sequences that we list here but note may be database contamination due to their lack of transcriptional support and failure to be included in a contig in the EuPath database. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (MrBayes)/Maximum-Likelihood bootstrap values (PhyML)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95/95, closed light circles ≥ 0.95/75/75, open circles ≥ 0.8/50/50.


Capsaspora owczarzaki ACAP_A
1/100/100
Sphaeroforma arctica ACAP_A
Sphaeroforma arctica ACAP_C
Sphaeroforma arctica ACAP_D
Sphaeroforma arctica ACAP_E
Nematostella vectensis ACAP_B
Rhizopus delemar ACAP
Monosiga brevicollis ACAP_A
Sphaeroforma arctica ACAP_F
Monosiga brevicollis ACAP_B
Sphaeroforma arctica ACAP_B
Dictyostelium discoideum ACAP_B
Acanthamoeba castellanii ACAP_C
Dictyostelium discoideum ACAP_A
Neurospora crassa ACAP
Schizosaccharomyces pombe ACAP_B
Schizosaccharomyces pombe ACAP_A
Thecamonas trahens ACAP
Trichomonas vaginalis ACAP_C
Phytophthora sojae ACAP_B
Thalassiosira pseudonana ACAP_A
Acanthamoeba castellanii ACAP_A
Arabidopsis thaliana ACAP_A
Arabidopsis thaliana ACAP_C
1/100/99
Arabidopsis thaliana ACAP_B
Arabidopsis thaliana ACAP_D
Physcomitrella patens ACAP_A
Physcomitrella patens ACAP_B
Physcomitrella patens ACAP_C
Chlamydomonas reinhardtii ACAP
Emiliania huxleyi ACAP_A
Emiliania huxleyi ACAP_B
Caenorhabditis elegans ACAP
Toxoplasma gondii ACAP*
Drosophila melanogaster ACAP
Ciona intestinalis ACAP
Danio rerio ACAP_A
1/100/94
Homo sapiens ACAP1
Mus musculus ACAP_A
0.96/83
/69
Xenopus laevis ACAP_A
Danio rerio ACAP_D
Gallus gallus ACAP_A
1
Homo sapiens ACAP2
0.99/90
/99
Mus musculus ACAP_B
/72
/93
Rattus norvegicus ACAP_A
Xenopus laevis ACAP_B
Danio rerio ACAP_B
Danio rerio ACAP_E
Gallus gallus ACAP_B
1
Homo sapiens ACAP3
/100
Mus musculus ACAP_C
/99
Rattus norvegicus ACAP_B
Danio rerio ACAP_C
Nematostella vectensis ACAP_A
0.2
MrBayes/PhyML/RAxML
≥ 1.00/95/95
≥ 0.95/75/75
≥ 0.80/50/50

rather strengthened the support for its exclusion from the rest of the vertebrates (Figure 3-16). The simplest explanation is to disregard this paralogue as a highly divergent sequence whose exclusion from vertebrates as a phylogenetic artefact. Alternatively, this represents an additional paralogue that emerged early in vertebrates and was lost from all other taxa sampled.

3.4.1.5.4 AGFG

Two paralogues of AGFG are known in humans. AGFG, also known as Hrb, is an essential HIV cofactor, necessary for the release of HIV RNAs from the peri-nuclear region (Sánchez-Velar et al., 2004). More recently, it has also been implicated in the retrieval of VAMP7 from the plasma membrane and can therefore act as a cargo adaptor for clathrin-mediated endocytosis (Pryor et al., 2008), consistent with its localization to the plasma membrane, and interaction with AP-2 and clathrin (Chaineau et al., 2008). AGFG is named for the presence of phenylalanine-glycine repeats (Kahn et al., 2008; Pryor et al., 2008). AGFG is thought to be an ancient ArfGAP subfamily and therefore likely broadly distributed (Kahn et al., 2008). Phylogenetic reconstruction of AGFG sequences (Figure 3-17) revealed expansion in the archaeplastids, as well as two paralogues in vertebrates, the result of a single gene duplication. Notably, the clade marked by the human AGFG2 appears to have undergone loss from *Xenopus laevis* and from *Gallus gallus*. Node support for each of these clades increased when all non-filozoan sequences were removed for the analysis (Figure 3-18).

Figure 3-16. Phylogenetic analysis of metazoan ACAP sequences identifies an expansion in vertebrates. In order gain better node support to resolve the order of gene duplications giving rise to the three vertebrate ACAP clades, only the metazoan sequences from Figure 3-15 were analyzed. ACAP has undergone at least two gene duplication events resulting in at least three paralogues in vertebrates. The first duplication produced ACAP3 and the second duplication produced ACAP1 and ACAP2. An extra *D. rerio* sequence, ACAP_C, branches at the base of the vertebrate clade (see text for interpretation). The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (MrBayes)/Maximum-Likelihood bootstrap values (PhyML)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95/95, closed light circles ≥ 0.95/75/75, open circles ≥ 0.8/50/50.

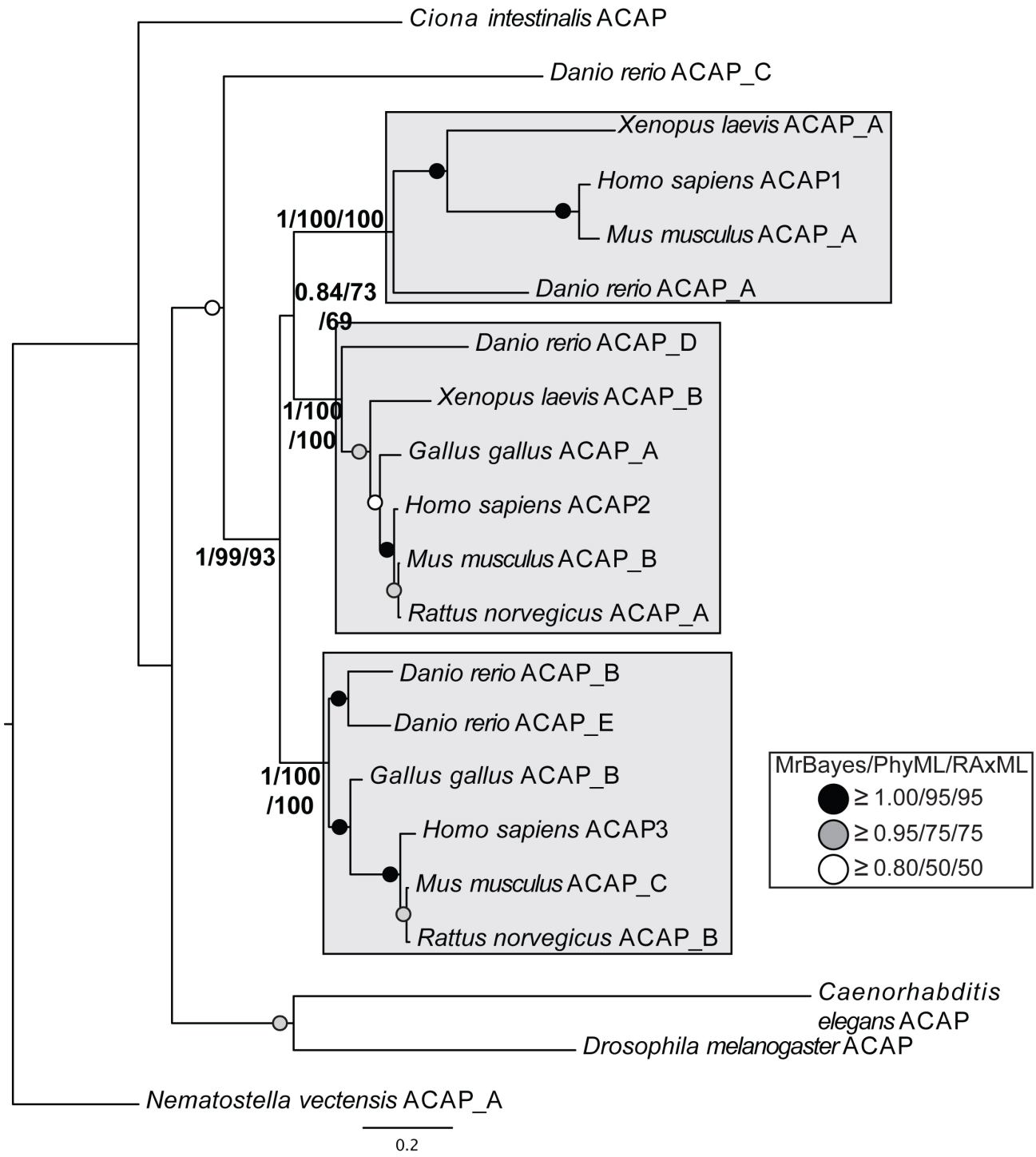

Ciona intestinalis ACAP
Danio rerio ACAP_C
Xenopus laevis ACAP_A
Homo sapiens ACAP1
Mus musculus ACAP_A
Danio rerio ACAP_A
1/100/100
0.84/73
/69
Danio rerio ACAP_D
Xenopus laevis ACAP_B
1/100
/100
Gallus gallus ACAP_A
Homo sapiens ACAP2
1/99/93
Mus musculus ACAP_B
Rattus norvegicus ACAP_A
Danio rerio ACAP_B
Danio rerio ACAP_E
1/100
/100
Gallus gallus ACAP_B
Homo sapiens ACAP3
Mus musculus ACAP_C
Rattus norvegicus ACAP_B
MrBayes/PhyML/RAxML
≥ 1.00/95/95
≥ 0.95/75/75
≥ 0.80/50/50
Caenorhabditis elegans ACAP
Drosophila melanogaster ACAP
Nematostella vectensis ACAP_A
0.2

Figure 3-17. Phylogenetic analysis of AGFG identifies a single paralogue in the LECA. Absence of support for multiple clades of supergroups indicates that a single AGFG sequence was present in the LECA. A weakly supported duplication event at the base of vertebrates has produced two clades of AGFG paralogues (lower grey boxes). Additional expansions have occurred in the Archaeplastida: one in *A. thaliana* and one in *P. patens* (upper grey box). The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (MrBayes)/Maximum-Likelihood bootstrap values (PhyML)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95/95, closed light circles ≥ 0.95/75/75, open circles ≥ 0.8/50/50.

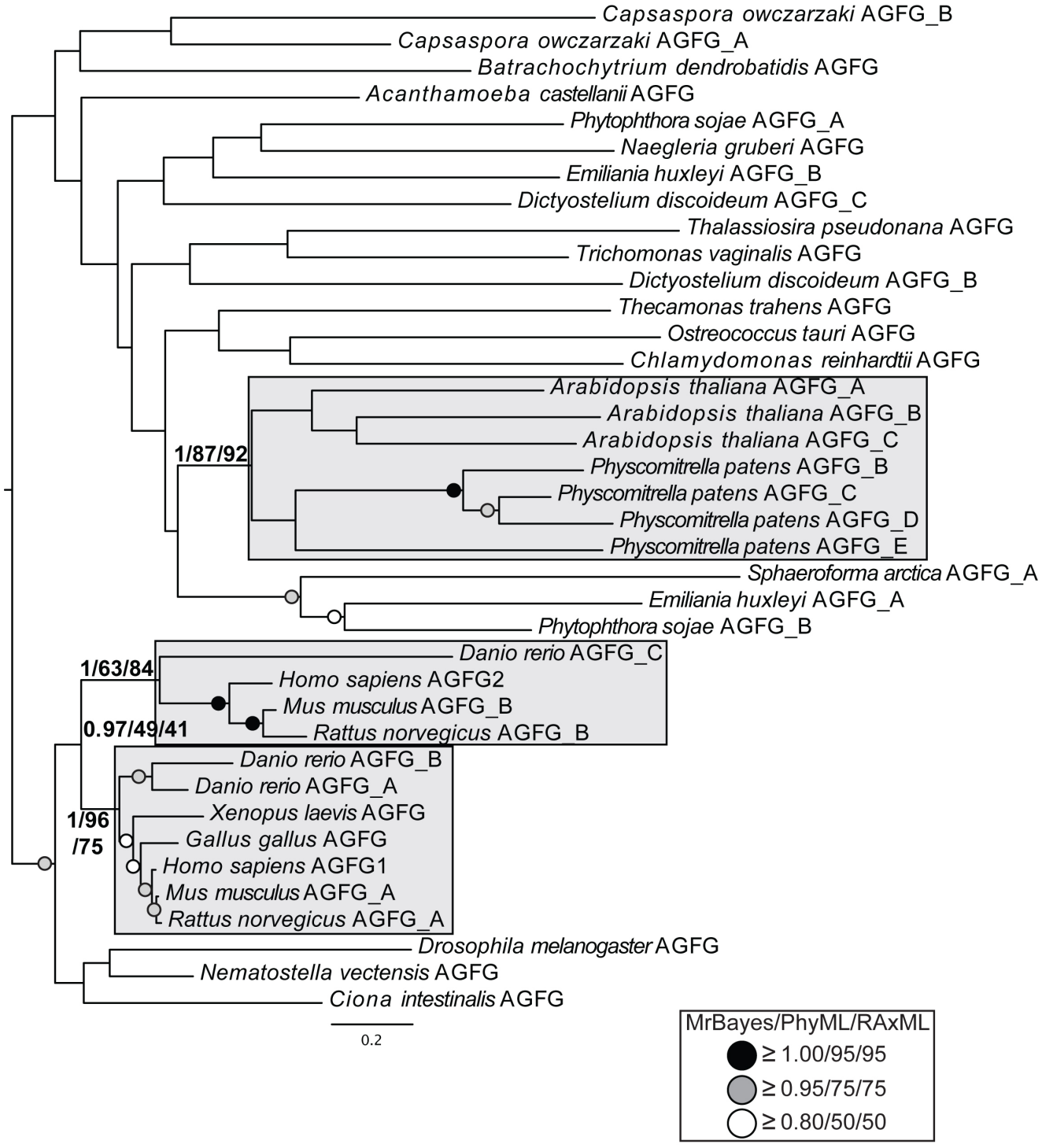
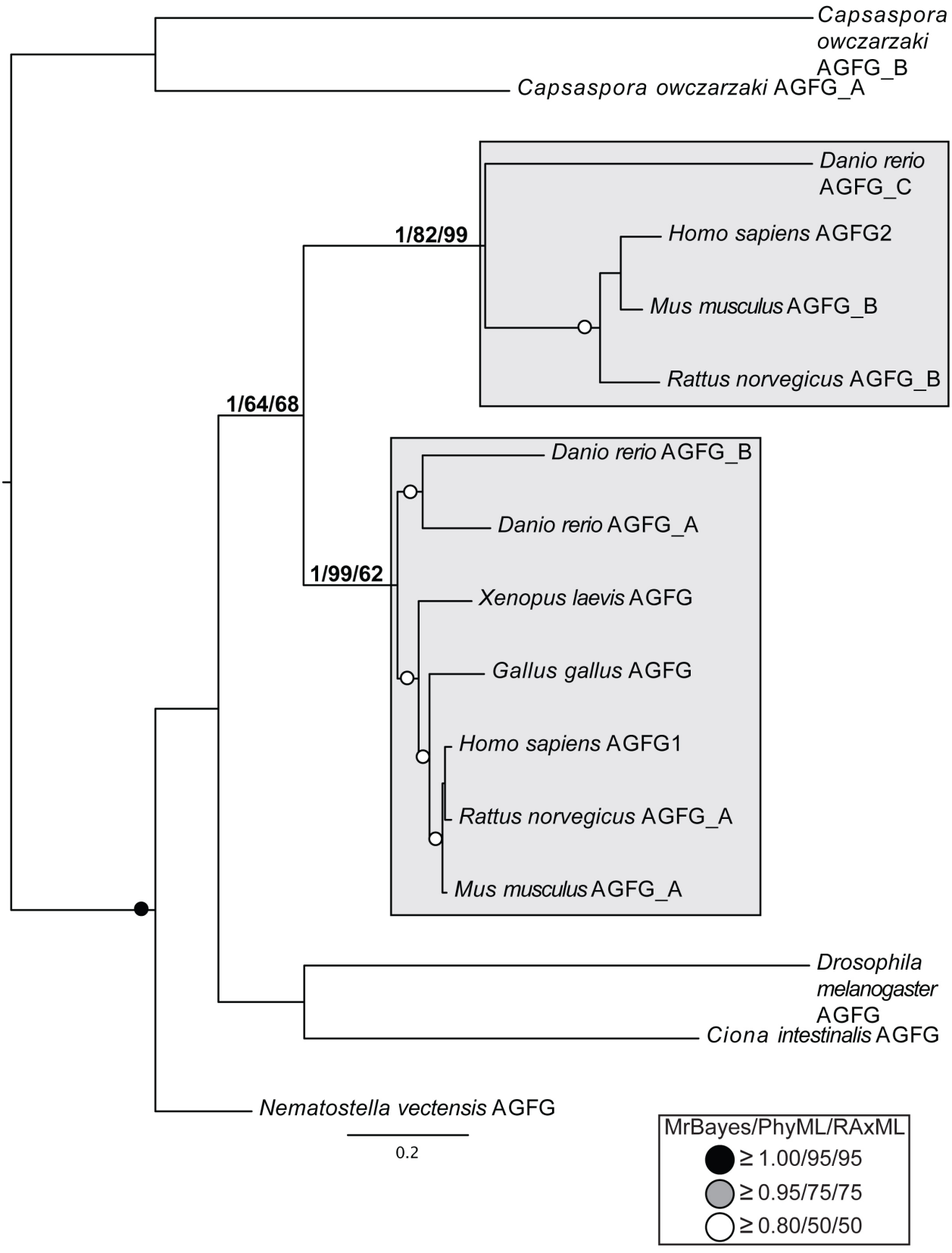

Capsaspora owczarzaki AGFG_B
Capsaspora owczarzaki AGFG_A
Batrachochytrium dendrobatidis AGFG
Acanthamoeba castellanii AGFG
Phytophthora sojae AGFG_A
Naegleria gruberi AGFG
Emiliania huxleyi AGFG_B
Dictyostelium discoideum AGFG_C
Thalassiosira pseudonana AGFG
Trichomonas vaginalis AGFG
Dictyostelium discoideum AGFG_B
Thecamonas trahens AGFG
Ostreococcus tauri AGFG
Chlamydomonas reinhardtii AGFG
Arabidopsis thaliana AGFG_A
Arabidopsis thaliana AGFG_B
Arabidopsis thaliana AGFG_C
Physcomitrella patens AGFG_B
Physcomitrella patens AGFG_C
Physcomitrella patens AGFG_D
Physcomitrella patens AGFG_E
1/87/92
Sphaeroforma arctica AGFG_A
Emiliania huxleyi AGFG_A
Phytophthora sojae AGFG_B
Danio rerio AGFG_C
1/63/84
Homo sapiens AGFG2
Mus musculus AGFG_B
Rattus norvegicus AGFG_B
0.97/49/41
Danio rerio AGFG_B
Danio rerio AGFG_A
Xenopus laevis AGFG
1/96
/75
Gallus gallus AGFG
Homo sapiens AGFG1
Mus musculus AGFG_A
Rattus norvegicus AGFG_A
Drosophila melanogaster AGFG
Nematostella vectensis AGFG
Ciona intestinalis AGFG
0.2
MrBayes/PhyML/RAxML
≥ 1.00/95/95
≥ 0.95/75/75
≥ 0.80/50/50

Figure 3-18. Phylogenetic analysis of filozoan AGFG identifies two vertebrate paralogues. In order to increase support resolving the gene duplication event at the base of vertebrates producing AGFG1 and AGFG2, all non-filozoan sequences were removed from the analysis. Two well-supported clades, AGFG1 and AGFG2, are reconstructed (grey boxes). Node support uniting these clades also increased, confirming that this duplication is vertebrate-specific. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (MrBayes)/Maximum-Likelihood bootstrap values (PhyML)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95/95, closed light circles ≥ 0.95/75/75, open circles ≥ 0.8/50/50.


Capsaspora owczarzaki AGFG_B
Capsaspora owczarzaki AGFG_A
Danio rerio AGFG_C
1/82/99
Homo sapiens AGFG2
Mus musculus AGFG_B
Rattus norvegicus AGFG_B
1/64/68
Danio rerio AGFG_B
Danio rerio AGFG_A
1/99/62
Xenopus laevis AGFG
Gallus gallus AGFG
Homo sapiens AGFG1
Rattus norvegicus AGFG_A
Mus musculus AGFG_A
Drosophila melanogaster AGFG
Ciona intestinalis AGFG
Nematostella vectensis AGFG
0.2
MrBayes/PhyML/RAxML
≥ 1.00/95/95
≥ 0.95/75/75
≥ 0.80/50/50

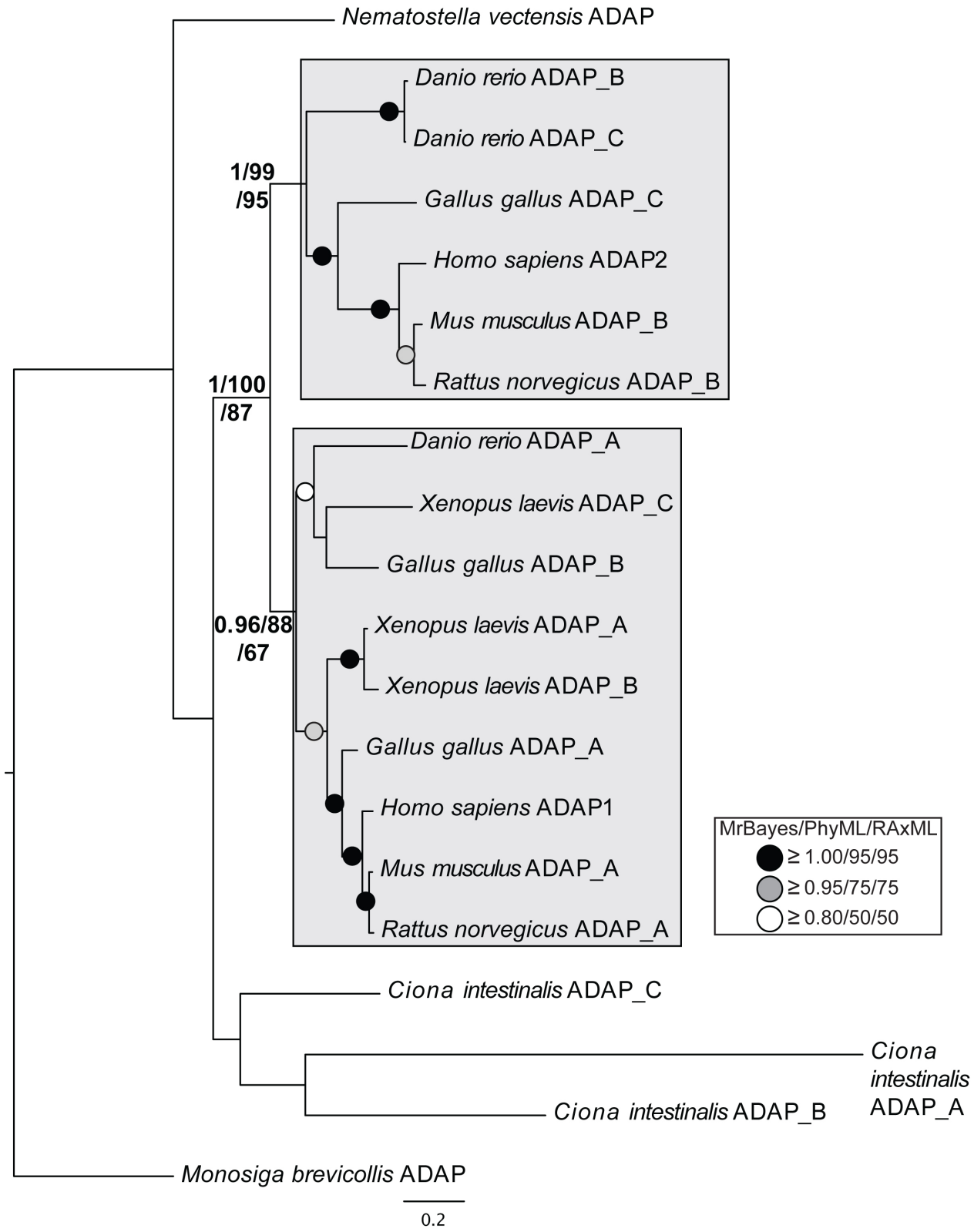
3.4.1.5.5 ADAP

ADAP1 and ADAP2 contain an ArfGAP domain and two PH domains (Venkateswarlu and Cullen, 1999), which have been shown to bind phosphoinositides PI(3,4,5)P_3 and Ins(1,3,4,5)P_4 with high affinity (Venkateswarlu et al., 2004). In humans, ADAP has been localized to dendrites, spines, and synapses of neurons and has a role in the traffic of regulated secretory vesicles (Thacker et al., 2004). Phylogenetic analysis of ADAP was limited to the Metazoa and choanoflagellates (Figure 3-19). Like ACAP, ADAP requires a more involved explanation, with at least two gene duplications and two major losses needing to be invoked. *D. rerio* would need to have lost its paralogue of ADAP1 and the entire mammalian clade would have lost the paralogue marked by DrADAP_A. It should be noted that neither the ADAP1 clade, nor the DrADAP_A clade is fully supported, and therefore, poor gene models or methodological artefact cannot be completely ruled out as causes for the observed topology.

3.4.1.5.6 ASAP

Three ASAP paralogues have been identified in humans. ASAPs possess an N-terminal BAR domain, followed by a PH domain, the ArfGAP domain, two proline-rich domains that are only present in ASAP1, and a C-terminal SH3 domain that is missing from ASAP3 (Brown et al., 1998). ASAPs localize to specialized plasma membrane structures (*e.g.,* focal adhesions; Randazzo et al., 2007), and are generally responsible for regulating endocytosis and actin remodelling (Inoue and Randazzo, 2007; Nie and Randazzo, 2006; Randazzo et al., 2007). ASAPs were previously

Figure 3-19. **Phylogenetic analysis of metazoan and *M. brevicollis* ADAP sequences identified two vertebrate paralogues.** Phylogenetic analysis reconstructs two strongly supported ADAP clades in vertebrates (grey boxes). The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (MrBayes)/Maximum-Likelihood bootstrap values (PhyML)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95/95, closed light circles ≥ 0.95/75/75, open circles ≥ 0.8/50/50.


Nematostella vectensis ADAP
Danio rerio ADAP_B
Danio rerio ADAP_C
1/99
/95
Gallus gallus ADAP_C
Homo sapiens ADAP2
Mus musculus ADAP_B
Rattus norvegicus ADAP_B
1/100
/87
Danio rerio ADAP_A
Xenopus laevis ADAP_C
Gallus gallus ADAP_B
0.96/88
/67
Xenopus laevis ADAP_A
Xenopus laevis ADAP_B
Gallus gallus ADAP_A
Homo sapiens ADAP1
Mus musculus ADAP_A
Rattus norvegicus ADAP_A
MrBayes/PhyML/RAxML
≥ 1.00/95/95
≥ 0.95/75/75
≥ 0.80/50/50
Ciona intestinalis ADAP_C
Ciona intestinalis ADAP_A
Ciona intestinalis ADAP_B
Monosiga brevicollis ADAP
0.2

thought to be metazoan-specific ArfGAPs (Kahn et al., 2008). The phylogeny of ASAP also requires further explanation, necessitating at least two duplications and loss of the clade marked by DrASAP_B from mammals, and independent loss of ASAP3 from *D. rerio* and *G. gallus* (Figure 3-20). Again, poor gene models, incomplete genomic databases, phylogenetic artefact, or insufficient taxon sampling could also explain these results.

3.4.1.5.7 SMAP

Two paralogues of SMAP are found in humans. SMAP1 is recruited to the plasma membrane and is involved in endocytosis while SMAP2 is involved in retrograde transport from early endosomes to the TGN (Natsume et al., 2006). SMAP1 and SMAP2 differ slightly in their domain compositions; both have a clathrin box, but SMAP2 additionally contains a CALM binding domain. The SMAP phylogeny (Figure 3-21) is straightforward to interpret, with one gene duplication at the base of vertebrates giving rise to two well supported paralogues. However, this duplication was only resolved in the analysis of metazoan sequences (Figure 3-22), as the analysis encompassing all sequences satisfying the 5-orders criterion did not exhibit a clear dichotomy between the two sets of paralogues. It also revealed a small, but unresolved expansion in archaeplastids possibly correlating with the evolution of multicellularity in this lineage (Figure 3-21).

Figure 3-20. Phylogenetic analysis of metazoan ASAP sequences identifies three vertebrate paralogues. Phylogenetic analysis has reconstructed three well-supported vertebrate-specific ASAP paralogues (grey boxes). The first gene duplication produced ASAP1, while the second produced ASAP2 and ASAP3. An independent expansion has also occurred in *N. vectensis.* The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (MrBayes)/Maximum-Likelihood bootstrap values (PhyML)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95/95, closed light circles ≥ 0.95/75/75, open circles ≥ 0.8/50/50.

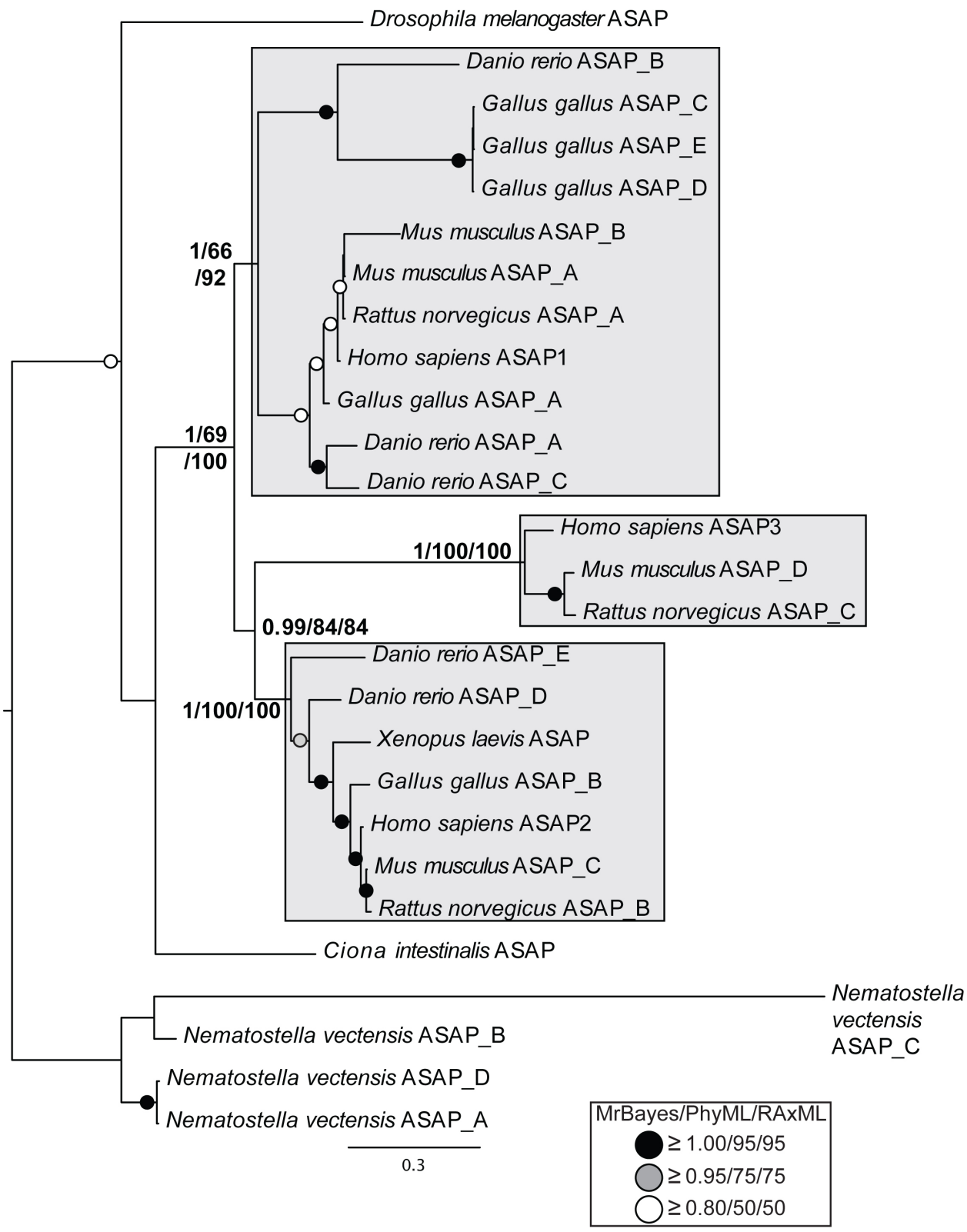

Drosophila melanogaster ASAP
Danio rerio ASAP_B
Gallus gallus ASAP_C
Gallus gallus ASAP_E
Gallus gallus ASAP_D
Mus musculus ASAP_B
1/66
/92
Mus musculus ASAP_A
Rattus norvegicus ASAP_A
Homo sapiens ASAP1
Gallus gallus ASAP_A
1/69
/100
Danio rerio ASAP_A
Danio rerio ASAP_C
Homo sapiens ASAP3
1/100/100
Mus musculus ASAP_D
Rattus norvegicus ASAP_C
0.99/84/84
Danio rerio ASAP_E
Danio rerio ASAP_D
1/100/100
Xenopus laevis ASAP
Gallus gallus ASAP_B
Homo sapiens ASAP2
Mus musculus ASAP_C
Rattus norvegicus ASAP_B
Ciona intestinalis ASAP
Nematostella vectensis ASAP_C
Nematostella vectensis ASAP_B
Nematostella vectensis ASAP_D
Nematostella vectensis ASAP_A
0.3
MrBayes/PhyML/RAxML
≥ 1.00/95/95
≥ 0.95/75/75
≥ 0.80/50/50

Figure 3-21. Phylogenetic analysis of SMAP identifies a single paralogue in the LECA. Multiple clades for some major lineages are reconstructed (*e.g.,* fungi and stramenopiles) although the phylogenetic tree is largely unsupported, indicating that this topology is no better than any other (*i.e.,* single clade of each supergroup). SMAP appears to have undergone one gene duplication event in vertebrates; however, it is unresolved using this dataset. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (MrBayes)/Maximum-Likelihood bootstrap values (PhyML)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95/95, closed light circles ≥ 0.95/75/75, open circles ≥ 0.8/50/50.

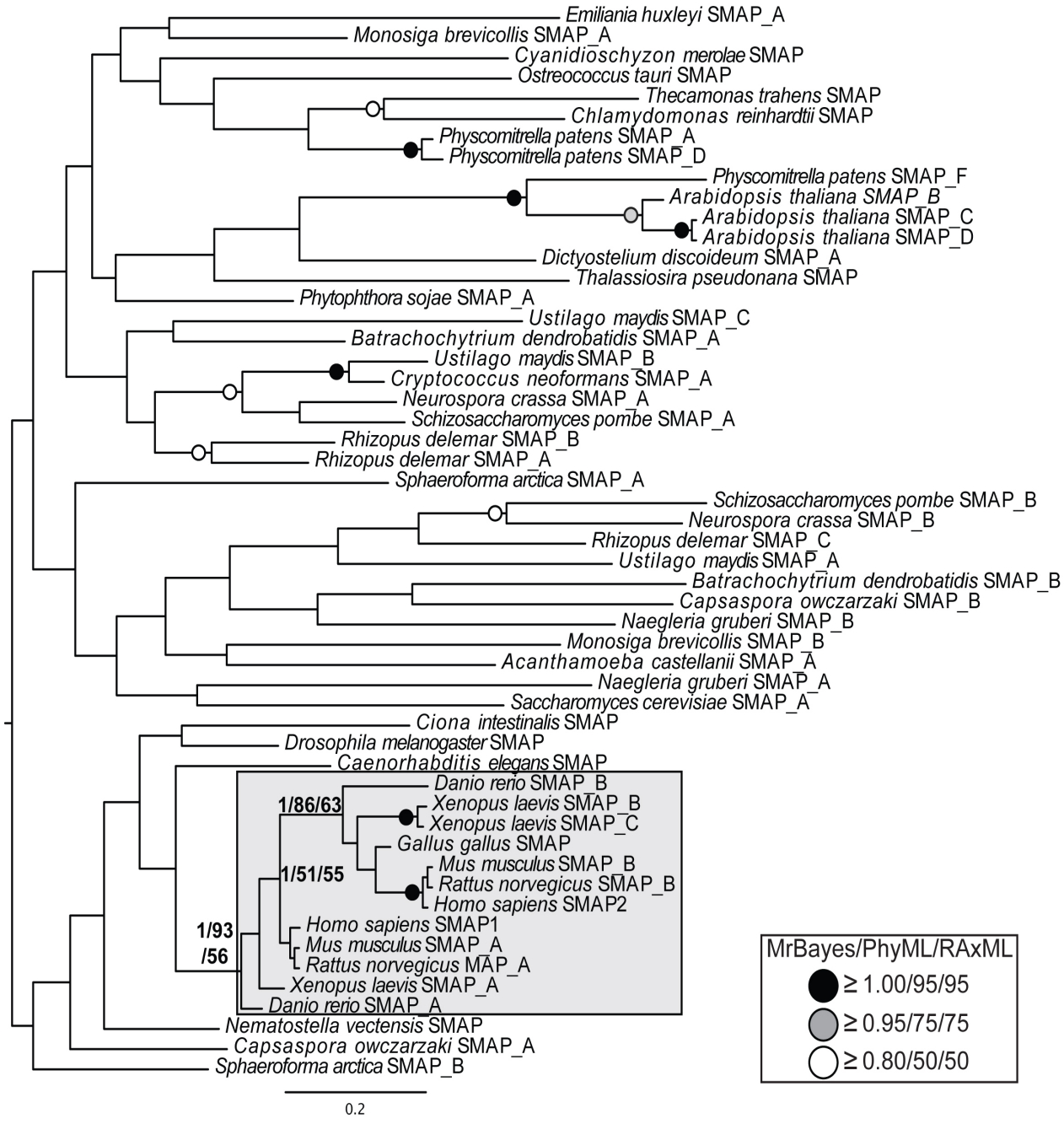
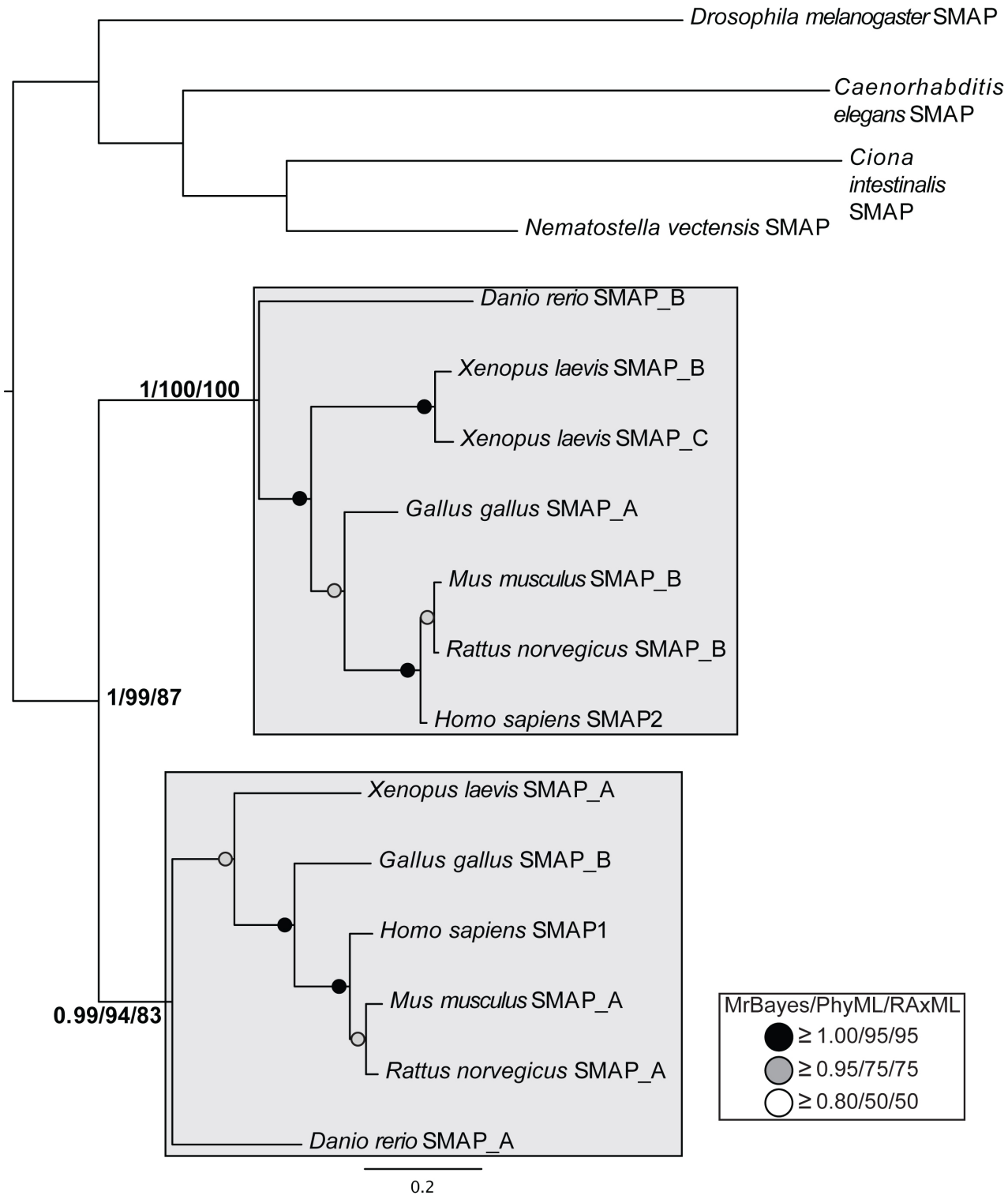

Emiliania huxleyi SMAP_A
Monosiga brevicollis SMAP_A
Cyanidioschyzon merolae SMAP
Ostreococcus tauri SMAP
Thecamonas trahens SMAP
Chlamydomonas reinhardtii SMAP
Physcomitrella patens SMAP_A
Physcomitrella patens SMAP_D
Physcomitrella patens SMAP_F
Arabidopsis thaliana SMAP_B
Arabidopsis thaliana SMAP_C
Arabidopsis thaliana SMAP_D
Dictyostelium discoideum SMAP_A
Thalassiosira pseudonana SMAP
Phytophthora sojae SMAP_A
Ustilago maydis SMAP_C
Batrachochytrium dendrobatidis SMAP_A
Ustilago maydis SMAP_B
Cryptococcus neoformans SMAP_A
Neurospora crassa SMAP_A
Schizosaccharomyces pombe SMAP_A
Rhizopus delemar SMAP_B
Rhizopus delemar SMAP_A
Sphaeroforma arctica SMAP_A
Schizosaccharomyces pombe SMAP_B
Neurospora crassa SMAP_B
Rhizopus delemar SMAP_C
Ustilago maydis SMAP_A
Batrachochytrium dendrobatidis SMAP_B
Capsaspora owczarzaki SMAP_B
Naegleria gruberi SMAP_B
Monosiga brevicollis SMAP_B
Acanthamoeba castellanii SMAP_A
Naegleria gruberi SMAP_A
Saccharomyces cerevisiae SMAP_A
Ciona intestinalis SMAP
Drosophila melanogaster SMAP
Caenorhabditis elegans SMAP
Danio rerio SMAP_B
1/86/63
Xenopus laevis SMAP_B
Xenopus laevis SMAP_C
Gallus gallus SMAP
1/51/55
Mus musculus SMAP_B
Rattus norvegicus SMAP_B
Homo sapiens SMAP2
1/93
/56
Homo sapiens SMAP1
Mus musculus SMAP_A
Rattus norvegicus MAP_A
Xenopus laevis SMAP_A
Danio rerio SMAP_A
Nematostella vectensis SMAP
Capsaspora owczarzaki SMAP_A
Sphaeroforma arctica SMAP_B
0.2
MrBayes/PhyML/RAxML
≥ 1.00/95/95
≥ 0.95/75/75
≥ 0.80/50/50

Figure 3-22. Phylogenetic analysis of metazoan SMAP sequences identifies two vertebrate paralogues. In order to resolve the potentially vertebrate gene duplication, all non-metazoan SMAP sequences were removed from the analysis. Strong support for SMAP1 and SMAP2 are recovered, and for the node unifying the two. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (MrBayes)/Maximum-Likelihood bootstrap values (PhyML)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95/95, closed light circles ≥ 0.95/75/75, open circles ≥ 0.8/50/50.


Drosophila melanogaster SMAP
Caenorhabditis elegans SMAP
Ciona intestinalis SMAP
Nematostella vectensis SMAP
Danio rerio SMAP_B
Xenopus laevis SMAP_B
Xenopus laevis SMAP_C
Gallus gallus SMAP_A
Mus musculus SMAP_B
Rattus norvegicus SMAP_B
Homo sapiens SMAP2
1/100/100
1/99/87
Xenopus laevis SMAP_A
Gallus gallus SMAP_B
Homo sapiens SMAP1
Mus musculus SMAP_A
Rattus norvegicus SMAP_A
Danio rerio SMAP_A
0.99/94/83
0.2
MrBayes/PhyML/RAxML
≥ 1.00/95/95
≥ 0.95/75/75
≥ 0.80/50/50

3.4.1.5.8 AGAP

Eleven AGAP genes have been identified in humans. AGAPs possess an N-terminal GTPase-like domain, followed by a PH domain, the ArfGAP domain, and ankyrin repeats (Nie et al., 2002; Xia et al., 2003). AGAP1 and AGAP2 both act in the endosomal system and regulate AP-3 and AP-1 trafficking pathways, respectively (Nie et al., 2003; Nie et al., 2005). Previous analysis has suggested that AGAP is only found in Metazoa (Kahn et al., 2008). Phylogenetic analysis of the holozoan AGAP sequences revealed that two duplications produced three paralogues in vertebrates (Figure 3-23). *X. laevis* appears to have lost AGAP2 along with *D. rerio,* which also appears to have lost AGAP3. This analysis also shows that AGAPs 4-11 are the result of multiple duplications of the human AGAP1 gene, and are not found in any of the other organisms sampled.

3.4.1.5.9 GIT

Two GIT paralogues are present in humans, both with N-terminal ArfGAP domains followed by ankyrin repeats, an SH domain, a coiled-coil domain, and a paxilin binding site (Mazaki et al., 2001; Premont et al., 1998; Turner et al., 1999). Unlike other ArfGAPs, GITs bind specifically to PIX/Cool proteins (Feng et al., 2002; Loo et al., 2004; Manser et al., 1998), which act as GEFs for the Rho GTPases Rac1 and Cdc42. Collectively, this complex acts as a site for signal integration at the plasma membrane (Hoefen and Berk, 2006). Like ASAPs, GITs were also predicted to be metazoan-specific (Kahn et al., 2008). The most parsimonious interpretation for the GIT phylogeny (Figure 3-24) is a gene duplication at the base of vertebrates

Figure 3-23. Phylogenetic analysis of holozoan AGAP sequences identifies three vertebrate paralogues. Phylogenetic analysis was carried out to determine whether the duplications producing the multiple mammalian AGAP paralogues occurred before or after the vertebrate transition. Two gene duplications in vertebrates produced the three paralogues as outlined by the grey boxes. AGAP2 was the first to diverge, followed by AGAP1 and AGAP3. The *H. sapiens* AGAPs 4-11 are the result of an expansion of AGAP1. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (MrBayes)/Maximum-Likelihood bootstrap values (PhyML)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95/95, closed light circles ≥ 0.95/75/75, open circles ≥ 0.8/50/50.

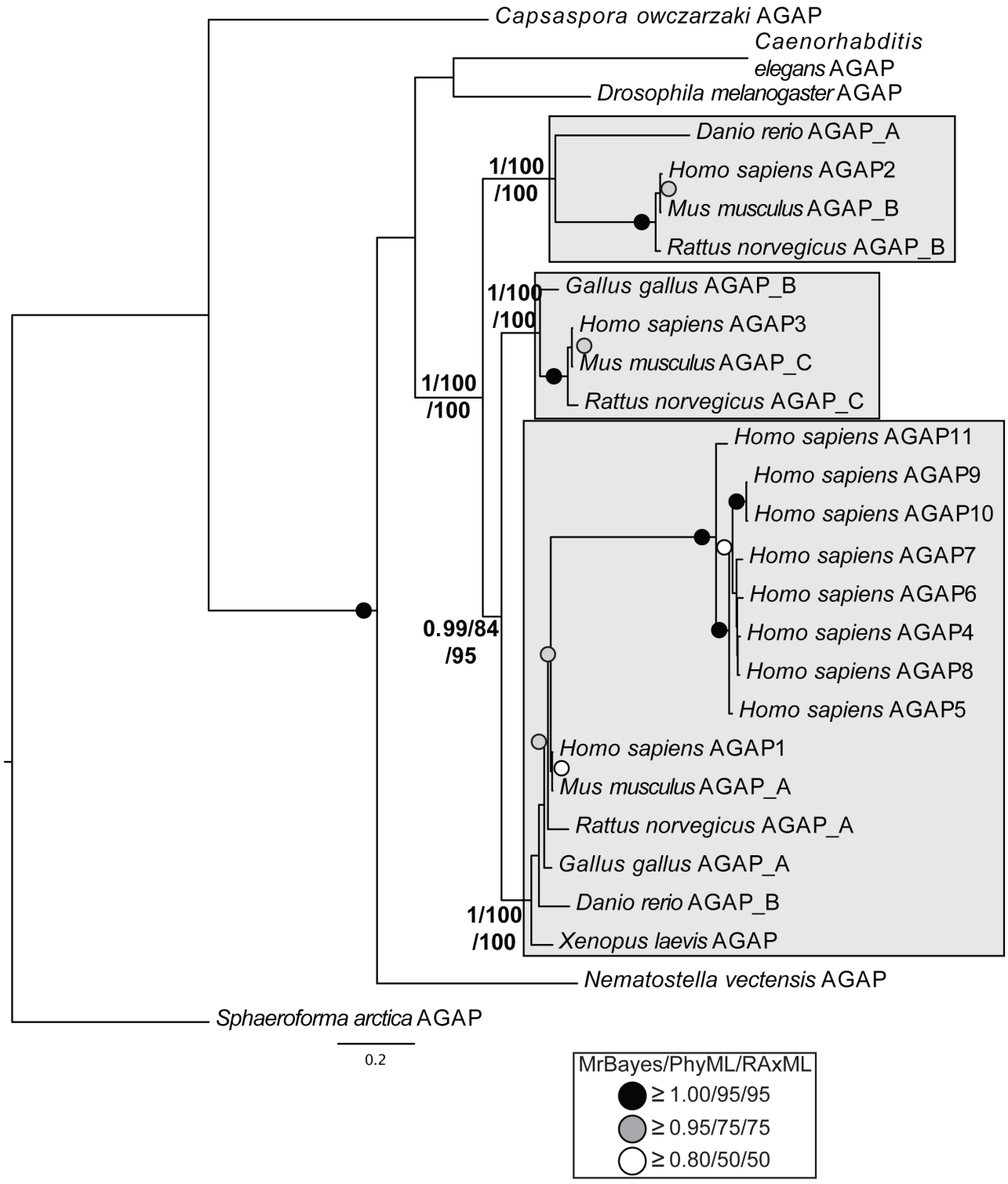
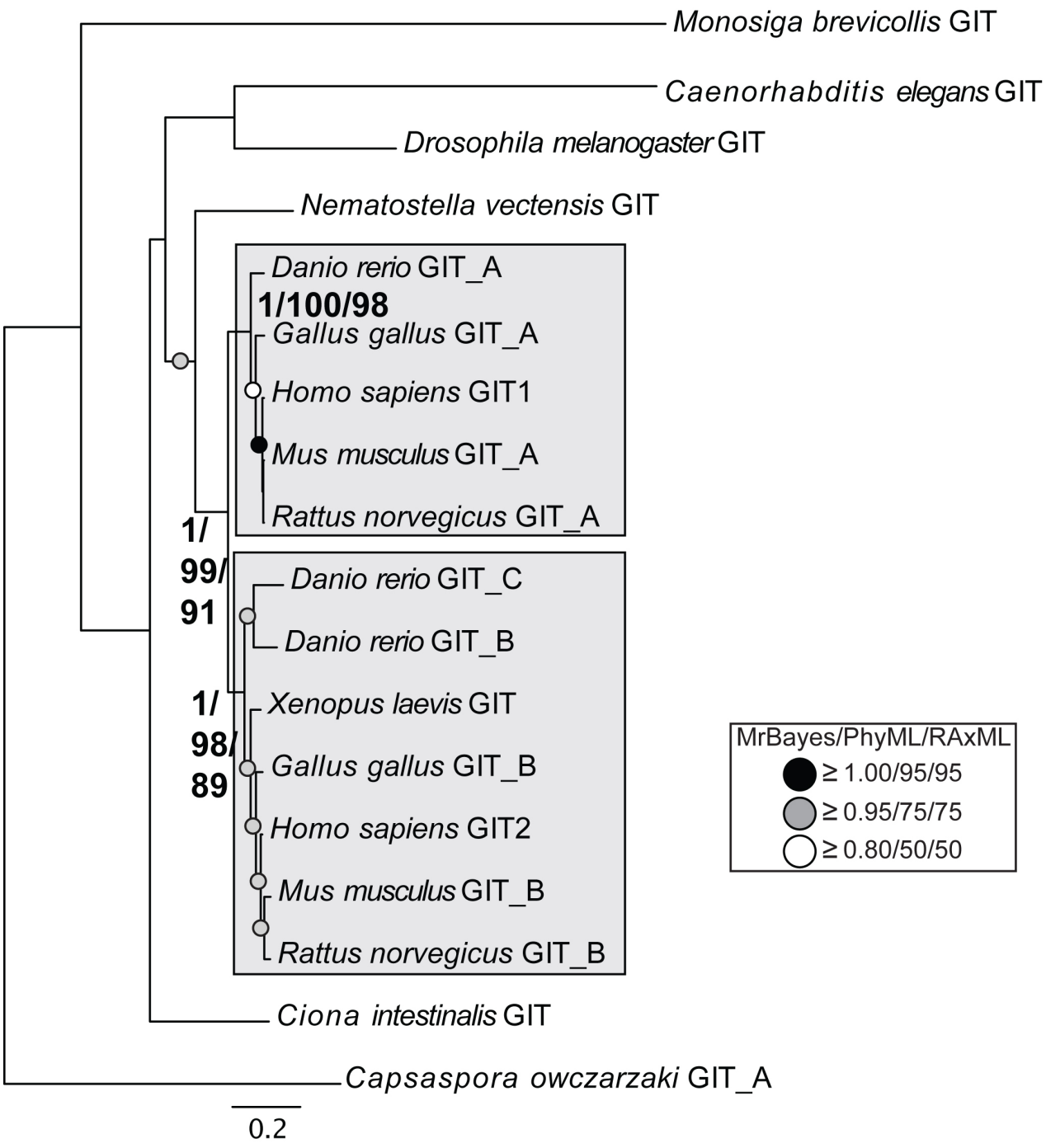

Capsaspora owczarzaki AGAP
Caenorhabditis elegans AGAP
Drosophila melanogaster AGAP
Danio rerio AGAP_A
Homo sapiens AGAP2
Mus musculus AGAP_B
Rattus norvegicus AGAP_B
1/100
/100
Gallus gallus AGAP_B
Homo sapiens AGAP3
Mus musculus AGAP_C
Rattus norvegicus AGAP_C
1/100
/100
1/100
/100
Homo sapiens AGAP11
Homo sapiens AGAP9
Homo sapiens AGAP10
Homo sapiens AGAP7
Homo sapiens AGAP6
Homo sapiens AGAP4
Homo sapiens AGAP8
Homo sapiens AGAP5
0.99/84
/95
Homo sapiens AGAP1
Mus musculus AGAP_A
Rattus norvegicus AGAP_A
Gallus gallus AGAP_A
Danio rerio AGAP_B
Xenopus laevis AGAP
1/100
/100
Nematostella vectensis AGAP
Sphaeroforma arctica AGAP
0.2
MrBayes/PhyML/RAxML
≥ 1.00/95/95
≥ 0.95/75/75
≥ 0.80/50/50

Figure 3-24. Phylogenetic analysis of GIT sequences identifies two vertebrate paralogues. Phylogenetic analysis was carried out to determine whether the gene duplication giving rise to GIT1 and GIT2 occurred before or after the vertebrate transition. Analysis strongly indicates that the duplication event occurred near the base of vertebrates. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (MrBayes)/Maximum-Likelihood bootstrap values (PhyML)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95/95, closed light circles ≥ 0.95/75/75, open circles ≥ 0.8/50/50.

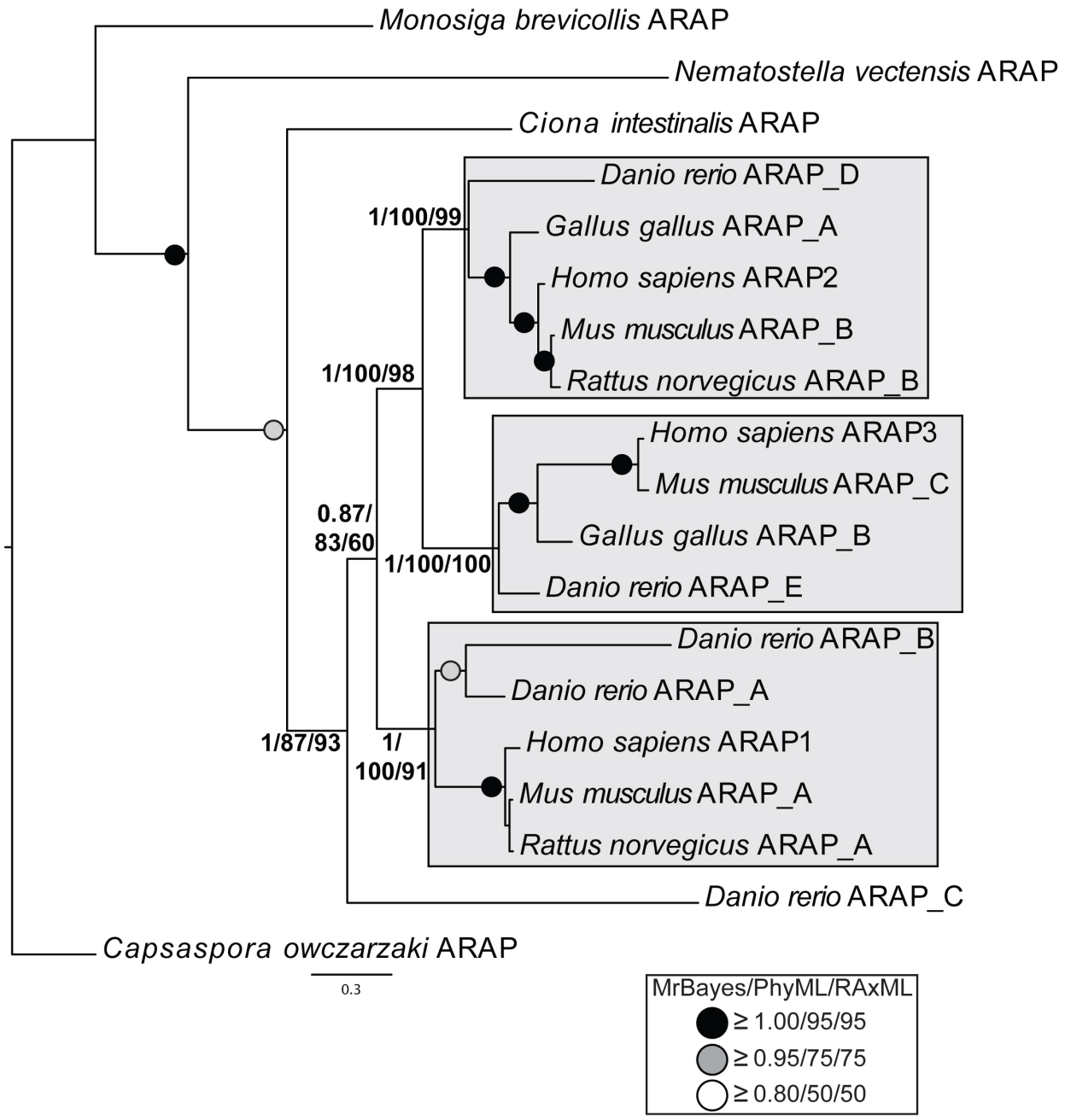

Monosiga brevicollis GIT
Caenorhabditis elegans GIT
Drosophila melanogaster GIT
Nematostella vectensis GIT
Danio rerio GIT_A
1/100/98
Gallus gallus GIT_A
Homo sapiens GIT1
Mus musculus GIT_A
Rattus norvegicus GIT_A
1/
99/
91
Danio rerio GIT_C
Danio rerio GIT_B
1/
98/
89
Xenopus laevis GIT
Gallus gallus GIT_B
Homo sapiens GIT2
Mus musculus GIT_B
Rattus norvegicus GIT_B
Ciona intestinalis GIT
Capsaspora owczarzaki GIT_A
0.2
MrBayes/PhyML/RAxML
≥ 1.00/95/95
≥ 0.95/75/75
≥ 0.80/50/50

that produced the two paralogues found in humans. However, *X. laevis* appears to have lost the GIT1 paralogue.

3.4.1.5.10 ARAP

Three ARAP paralogues are present in the human genome. ARAPs have N-terminal sterile α-motifs, followed by two PH domains, the ArfGAP domain, ankyrin repeats, two additional PH domains, a RhoGAP domain, a Ras association domain, and a C-terminal PH domain (Krugmann et al., 2002; Miura et al., 2002). ARAPs are important for signal coordination between Arf and Rho GTPase pathways (Krugmann et al., 2002; Krugmann et al., 2004; Miura et al., 2002; Stacey et al., 2004; Yoon et al., 2006) and are predicted to be chordate specific (Kahn et al., 2008). Phylogenetic analysis of ARAP suggests that two gene duplications near the base of vertebrates produced the three paralogues observed in *H. sapiens* (Figure 3-25). However, an additional *D. rerio* sequence (ARAP_C) branches basal to the rest of vertebrates with strong support. Parsimony would suggest that the position of this sequence is likely the result of phylogenetic artefact (LBA), as this sequence and all of the invertebrate sequences are quite long. Alternatively, this could represent an additional gene duplication at the base of vertebrates that was subsequently lost in all other taxa. Revisiting this question in the future with a deeper sampling of metazoan genomes should resolve this question. Nonetheless, the three human paralogues of ARAP are clearly the result of a vertebrate-specific expansion.

Figure 3-25. Phylogenetic analysis of ARAP sequences identifies three vertebrate paralogues. To determine the order of gene duplications and when these duplications occurred relative to the vertebrate transition, phylogenetic analysis was carried out. The two duplications occurred near the base of vertebrates, with ARAP1 diverging first, followed by ARAP2 and ARAP3. An extra *D. rerio* sequence, ARAP_C, branches prior to the three ARAP paralogues (see text for interpretation). The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (MrBayes)/Maximum-Likelihood bootstrap values (PhyML)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95/95, closed light circles ≥ 0.95/75/75, open circles ≥ 0.8/50/50.


Monosiga brevicollis ARAP
Nematostella vectensis ARAP
Ciona intestinalis ARAP
Danio rerio ARAP_D
1/100/99
Gallus gallus ARAP_A
Homo sapiens ARAP2
Mus musculus ARAP_B
1/100/98
Rattus norvegicus ARAP_B
Homo sapiens ARAP3
Mus musculus ARAP_C
0.87/
83/60
Gallus gallus ARAP_B
1/100/100
Danio rerio ARAP_E
Danio rerio ARAP_B
Danio rerio ARAP_A
1/87/93
1/
100/91
Homo sapiens ARAP1
Mus musculus ARAP_A
Rattus norvegicus ARAP_A
Danio rerio ARAP_C
Capsaspora owczarzaki ARAP
0.3
MrBayes/PhyML/RAxML
≥ 1.00/95/95
≥ 0.95/75/75
≥ 0.80/50/50

3.4.2.1 Comparative genomic analysis of ArfGEF proteins identifies three subfamilies present in the LECA

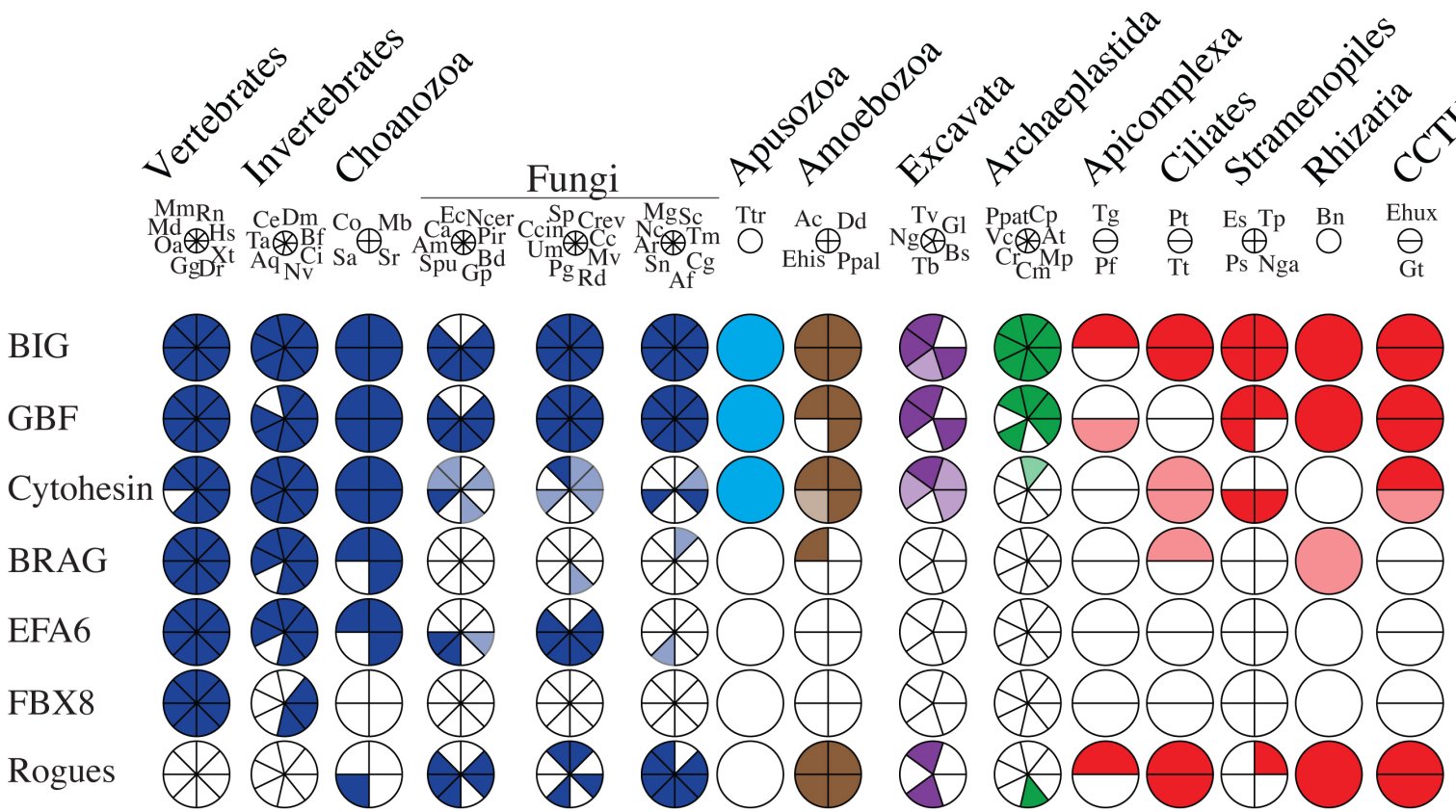
Having established the distribution of the human ArfGAP proteins across the major eukaryotic lineages, I moved on to examine their counter parts, the ArfGEFs. Similar to the GAPs, comparative genomic analysis was used to assess the distribution of the known ArfGEF subfamilies described in humans. A similar methodology was used to search 71 eukaryotic genomes (Figure 3-3, 3-4). The large difference in the number of genomes sampled between the two analyses reflects the number of newly sequenced genomes between the dates the studies were conducted. 451 Sec7 domain-containing sequences were identified, of which 353 were classified into one of the 6 human ArfGEF subfamilies at the 5-orders criterion. If the 2-orders criterion is used, an additional 43 sequences can be classified. The remaining 55 Sec7 domain-containing sequences could not be classified. BLAST of rogue sequences against genomes containing rogues (as done for ArfGAPC2) did not yield any significant results, suggesting that these rogues represent divergent members of known subfamilies, or represent lineage-specific subfamilies not present in opisthokont systems as opposed to broadly distributed unidentified subfamilies. These results could point to the acquisition of novel functions for some ArfGEFs in other eukaryotic lineages.

Although the taxon sampling of the ArfGEF analysis is nearly twice that of the ArfGAP analysis, we only identified five more GEFs than GAPs. Considering only the taxa used in the ArfGAP analysis, 264 ArfGEF sequences were identified, 218 of which were classified into an ArfGEF subfamily at the 5-orders criterion. An

additional 18 can be classified at the 2-orders criterion, leaving 28 rogue sequences. This translates to roughly 30% fewer GEFs than GAPs. While the full biological implications of this difference are not yet clear, it suggests that many more expansions of ArfGAP proteins have occurred than for the ArfGEFs. As with the ArfGAPs, phylogenetic analysis of all Sec7 domain-containing proteins did not classify any additional ArfGEF sequences.

BLASTp experiments against the nr-database were not carried out for the ArfGEFs because a larger number of genomes were used (as more had become available), therefore, redundancy was inherent in the taxon sampling of the analysis, reducing the likelihood of false negatives resulting from genome selection. This latter rationale was also carried over to the other analyses (*i.e.,* chapter 4 and chapter 5). Nonetheless, comparative genomic analysis of ArfGEFs identified three of the six subfamilies (BIG, GBF, cytohesin) present in at least four supergroups at the 5-orders criterion (Figure 3-26, Figure 3-27). BIG and GBF are broadly and frequently found in all eukaryotic lineages, whereas cytohesin is broadly but patchily distributed. If the 2-orders criterion is applied, then cytohesin is found in nearly every major lineage sampled. The broad distribution, even at the 5-orders criterion, suggests that these three ArfGEF subfamilies were present in the LECA. In contrast, EFA6 and FBX8 have narrower distributions, indicating more recent origins for these subfamilies. EFA6 is present in Holozoa and some fungi, but is absent from *S. arctica* and *Amphimedon queenslandica.* FBX8 is restricted to Metazoa and, is missing from *Trichoplax adhaerens, A. queenslandica, C. elegans,* and *D. melanogaster* (Figure 3-26). Absence from *C. elegans* and *D. melanogaster* could

Figure 3-26. Distribution of ArfGEF subfamilies across eukaryotic taxa. Three ArfGEF subfamilies are present in at least four eukaryotic supergroups at the 5-orders criterion, indicating they were present in the LECA. BRAG is present in opisthokonts and Amoebozoa, suggesting that it was present in the opisthokont and amoebozoan ancestor. Based on its presence in ciliates and *B. natans* at the 2-orders criterion, it may have been present in the LECA. EFA6 is only found in the Holozoa and Fungi, but appears to have been frequently lost from the latter group. FBX8 is specific to the Metazoa, but is missing from some early animals. Large taxonomic groups are colour coded. Sectors with solid colours indicate that at least one homologue satisfying the 5-orders criterion was identified. Sectors with pale colours indicate that all homologues identified only satisfy the 2-orders criterion. The key for species name abbreviations is boxed. The total number of Sec7 domain-containing proteins identified in each organism is indicated in brackets.



Holozoa
Hs = Homo sapiens (15)
Rn = Rattus norvegicus (13)
Mm = Mus musculus (15)
Md = Monodelphis domestica (13)
Oa = Ornithorhynchus anatinus (5)
Gg = Gallus gallus (13)
Xt = Xenopus tropicalis (14)
Dr = Danio rerio (23)
Bf = Branchiostoma floridae (11)
Ci = Ciona intestinalis (7)
Dm = Drosophila melanogaster (5)
Ce = Caenorhabditis elegans (4)
Ta = Trichoplax adhaerens (5)
Aq = Amphimedon queenslandica (4)
Nv = Nematostella vectensis (10)
Mb = Monosiga brevicollis (6)

Sr = Salpinogoeca rosetta (7)
Co = Capsapora owczarzaki (5)
Sa = Sphaeroforma arctica (8)

Fungi
Ec = Encephalitozoon cuniculi (1)
Ncer = Nosema ceranae (0)
Pir = Piromyces sp. (6)
Bd = Batrachochytrium dendrobatidis (6)
Gp = Gonapodya prolifera (4)
Spu = Spizellomyces punctatus (5)
Am = Allomyces macrogynus (15)
Ca = Catenaria anguillulae (6)
Crev = Coemansia reversa (4)
Cc = Conidiobolus coronatus (6)
Mv = Mortierella verticillata (6)
Rd = Rhizopus delemar (9)

Sp = Schizosaccharomyces pombe (6)
Sc = Saccharomyces cerevisiae (5)
Nc = Neurospora crassa (4)
Af = Aspergillus fumigatus (4)
Sn = Stagonospora nodorum (4)
Tm = Tuber melanosporum (4)
Cg = Cladonia grayi (4)
Ar = Amorphotheca resinae (4)
Mg = Magnaporthe grisea (4)
Pg = Puccinia graminis (5)
Ccin = Coprinopsis cinerea (5)
Um = Ustilago maydis (5)

Apusozoa
Ttr = Thecamonas trahens (7)

Amoebozoa
Ac = Acanthamoeba castellanii (6)
Eh = Entamoeba histolytica (9)
Ppal = Polysphondylium pallidum (7)
Dd = Dictyostelium discoideum (7)

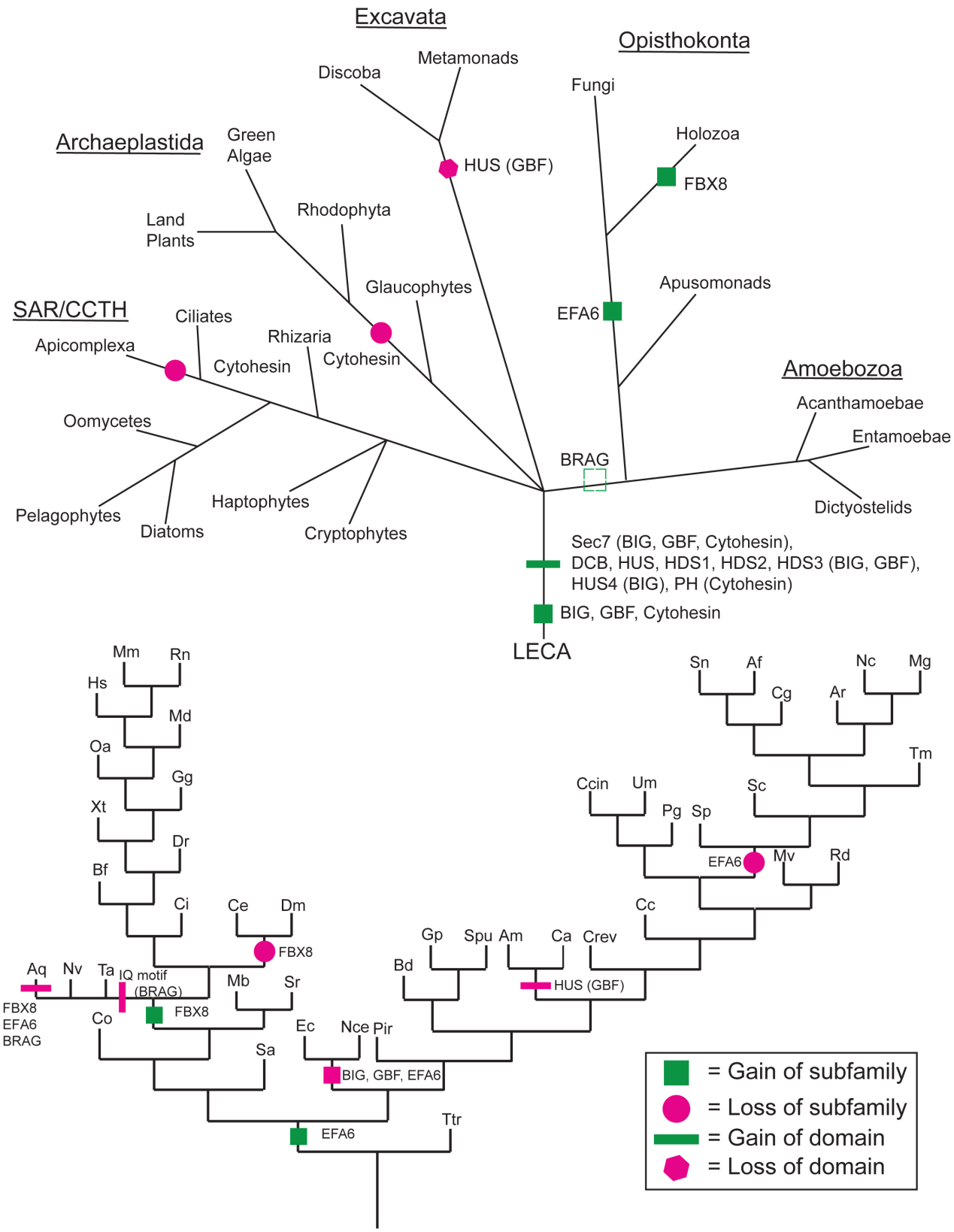
Excavata
Ng = Naegleria gruberi (4)
Tb = Trypanosoma brucei (3)
Bs = Bodo saltans (5)
Tv = Trichomonas vaginalis (18)
Gl = Giardia lamblia (2)

Archaeplastida
Cp = Cyanophora paradoxa (3)
Cm = Cyanidioschyzon merolae (2)
At = Arabidopsis thaliana (8)

Ppat = Physcomitrella patens (8)
Mp = Micromonas pusilla (3)
Vc = Volvox carteri (1)
Cr = Chlamydomonas reinhardtii (3)

SAR/CCTH
Tg = Toxoplasma gondii (4)
Pf = Plasmodium falciparum (1)
Pt = Paramecium tetraurelia (14)
Tt = Tetrahymena thermophila (7)
Ps = Phytophthora sojae (4)
Tp = Thalassiosira pseudonana (5)
Nga = Nannochloropsis gaditana (3)
Es = Ectocarpus siliculosus (6)
Bn = Bigelowiella natans (5)
Ehux = Emiliania huxleyi (9)
Gt = Guillardia theta (7)

Figure 3-27. Gain and loss of ArfGEF domains and subfamilies in eukaryotes. A) Tree of eukaryotes depicting the ArfGEF subfamilies and domains present in the LECA, as well as gains and losses of additional domains and subfamilies. To ensure confidence in the predictions, losses are only proposed when two instances of loss have occurred in the relevant lineage. The origin of BRAG is represented as box with dashed lines denoting the minimal distribution of this subfamily, as comparative genomic analysis suggests that it may be more broadly distributed. B) Gain and loss of ArfGEF subfamilies and domains in opisthokonts. Symbol legend is inset in B. DCB = Dimerization and Cyclophilin Binding; HUS = Homology Upstream of Sec7; HDS = Homology Downstream of Sec7; PH = Pleckstrin Homology. Abbreviations for taxa are as follows: Af = *Aspergillus fumigatus*, Am = *Allomyces macrogynus*, Ar = *Amorphotheca resinae*, Bd = *B. dendrobatidis*, Bf = *Branchiostoma floridae*, Ca = *Catenaria anguillulae*, Cc = *Conidiobolus coronatus*, Ccin = *Coprinopsis cinerea*, Ce = *C. elegans*, Cg = *Cladonia grayi*, Ci = *Ciona intestinalis*, Co = *C. owczarzaki*, Crev = *Coemansia reversa*, Dm = *D. melanogaster*, Dr = *D. rerio*, Ec = *E. cuniculi*, Gg = *G. gallus*, Gp = *Gonapodya prolifera*, Hs = *H. sapiens*, Mb = *M. brevicollis*, Md = *Monodelphis domestica*, Mg = *Magnaporthe grisea*, Mm = *M. musculus*, Mv = *Mortierella verticillata*, Nc = *Neurospora crassa*, Nce = *Nosema ceranae*, Nv = *Nematostella vectensis*, Oa = *Ornithorhynchus anatinus*, Pg = *Puccinia graminis*, Pir = *Piromyces sp.*, Rd = *R. delemar*, Rn = *Rattus norvegicus*, Sa = *S. arctica*, Sc = *S. cerevisiae*, Sn = *Stagonosporum nodorum*, Sp = *Schizosaccharomyces pombe*, Spu = *Spizellomyces punctatus*, Sr = *S. rosetta*, Ta = *Trichoplax adhaerens*, Tm = *Tuber melanosporum*, Ttr = *T. trahens*, Um = *Ustilago maydis*, Xt = *X. tropicalis*.


Excavata
Opisthokonta
Discoba
Metamonads
Fungi
Archaeplastida
Green Algae
HUS (GBF)
Holozoa
FBX8
Land Plants
Rhodophyta
Glaucophytes
Apusomonads
EFA6
SAR/CCTH
Ciliates
Rhizaria
Apicomplexa
Cytohesin
Cytohesin
Amoebozoa
Acanthamoebae
Oomycetes
Entamoebae
BRAG
Haptophytes
Pelagophytes
Diatoms
Cryptophytes
Dictyostelids
Sec7 (BIG, GBF, Cytohesin),
DCB, HUS, HDS1, HDS2, HDS3 (BIG, GBF),
HUS4 (BIG), PH (Cytohesin)
BIG, GBF, Cytohesin
LECA
Mm
Rn
Hs
Md
Oa
Gg
Xt
Dr
Bf
Ci
Ce
Dm
FBX8
Aq
Nv
Ta
IQ motif
(BRAG)
Mb
Sr
FBX8
FBX8
EFA6
BRAG
Co
Sa
Ec
Nce
Pir
BIG, GBF, EFA6
EFA6
Ttr
Sn
Af
Nc
Mg
Cg
Ar
Tm
Ccin
Um
Sc
Pg
Sp
EFA6
Mv
Rd
Cc
Gp
Spu
Am
Ca
Crev
Bd
HUS (GBF)
= Gain of subfamily
= Loss of subfamily
= Gain of domain
= Loss of domain

reflect instances of secondary loss, whereas absence from *T. adhaerens* could either be the product of secondary loss, or, given the contentiousness of the branching order in early animals, could represent the ancestral state depending on the phylogenetic position of *T. adhaerens* relative to *Nematostella vectensis* on the metazoan tree. Poor gene models or incomplete databases are also possible explanations for these absences.

Like the distributions of ADAP and AGAP, the distribution of BRAG is slightly more perplexing; sequences satisfying the 5-orders criterion are found in Metazoa, choanoflagellates, and *A. castellanii*, indicating that BRAG could have been present in the LECA, depending on the position of the root of the eukaryotic tree. If the 2-orders criterion is used, then homologues are also found in Fungi, ciliates, and Rhizaria, indicating that BRAG may have been present in the LECA. Nonetheless, it is clear that BRAG has an origin coinciding with the divergence of the Choanozoa at least, if not in the ancestor of opisthokonts and Amoebozoa.

Loss also appears to be quite prevalent within the ArfGEFs: cytohesin and EFA6 appear to have been lost multiple times from fungi, and BRAG is only present in fungi if the 2-orders criterion is considered. Moreover, cytohesin appears to have been nearly completely lost from Archaeplastida, and multiple losses of cytohesin and GBF appear to have occurred throughout the SAR clade (Figure 3-26).

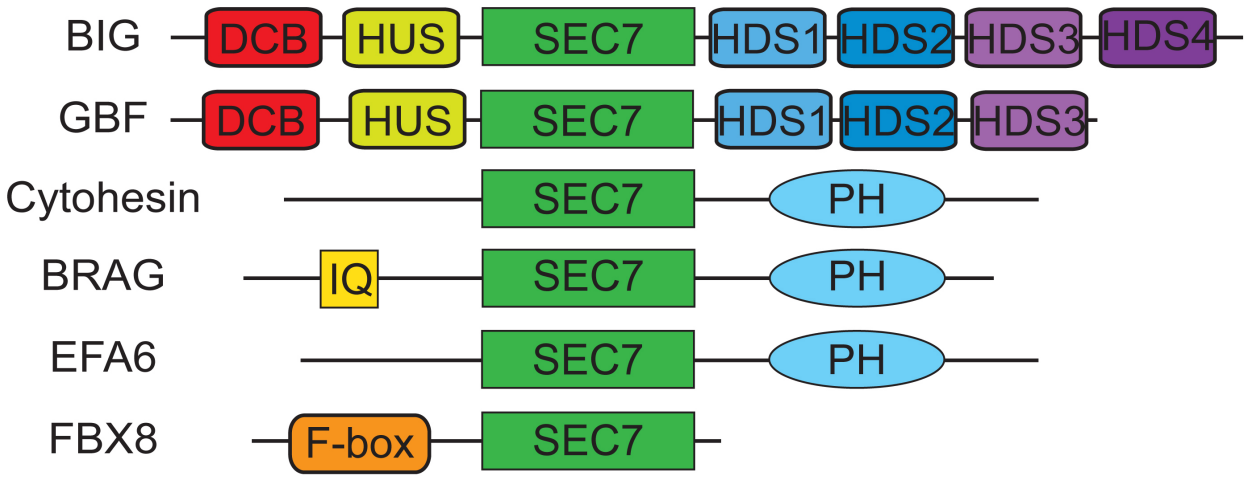
Overall, comparative genomics of ArfGAPs and GEFs has identified three patterns of conservation: i) well-conserved subfamilies that are found broadly and frequently (*e.g.,* SMAP, ArfGAP1, ArfGAP2, ACAP, AGFG, BIG, and GBF), ii) lineage-specific subfamilies, that are only found in a specific taxonomic group (*e.g.,* ARAP,

GIT, ASAP, FBX8, EFA6), and iii) patchy subfamilies, that are broadly distributed across eukaryotes, but are frequently missing (*e.g.,* AGAP, ADAP, cytohesin, and BRAG). The distributions of these subfamilies may hint at the type of processes in which they are involved. Well-conserved subfamilies are likely to be involved in ancient, fundamental cell biological process, given their broad distribution and near ubiquity. Patchy subfamilies may also be involved in ancient cellular process, but given the frequent loss of these subfamilies, these pathways may also be dispensable. In contrast, lineage-specific subfamilies are more likely to be involved in lineage-specific processes.

3.4.2.2 Domain analysis of ArfGEFs reveals a highly conserved domain complement

In contrast to the GAPs, ArfGEFs display a smaller array of accessory domains and much less domain plasticity. Domain profiles were created for each ArfGEF subfamily using the same methods and criteria used to create domain profiles for the ArfGAPs (see section 3.4.1.3). All domains characteristic of the human ArfGEF subfamilies are conserved across the distribution of their respective subfamilies (Figure 3-28). Also unlike the ArfGAPs, only a few instances of domain loss are apparent: the IQ motif of BRAG is missing from both *N. vectensis* and *T. adhaerens*, and the HUS domain has been lost from GBF in excavates and at least once in fungi. These few instances of domain loss notwithstanding, the broad conservation of the ArfGEF domain architectures suggests that the origin of these confirmations coincided with the origins of the subfamilies themselves, and that perhaps the apparent

Figure 3-28. Ancestral configuration of ArfGEF accessory domains. Conserved domains of each ArfGEF subfamily as defined by this study are shown in color. These represent the configurations likely found in the ancestral sequence of each subfamily. Unlike the ArfGAP domain configurations reported here (Figure 3-11), the human domain organization of ArfGEF proteins is conserved across the entire distribution of each subfamily and represents the domains present in the earliest ancestor of each subfamily. Less overall diversity in domain composition is observed than compared to the ArfGAP proteins. DCB = Dimerization and Cyclophilin-binding domain; HUS = Homology Upstream of Sec7; Sec7 = Sec7/ArfGEF catalytic domain; HDS1 = Homology Downstream of Sec7 1; HDS2 = Homology Downstream of Sec7 2; HDS3 = Homology Downstream of Sec7 3; PH = Pleckstrin Homology domain; IQ = IQ motif; F-box = F-box domain.


BIG
DCB
HUS
SEC7
HDS1
HDS2
HDS3
HDS4
GBF
DCB
HUS
SEC7
HDS1
HDS2
HDS3
Cytohesin
SEC7
PH
BRAG
IQ
SEC7
PH
EFA6
SEC7
PH
FBX8
F-box
SEC7

lack of plasticity is indicative of functionality integral to the function of the GEFs themselves.

3.4.2.3 Phylogenetic analysis of ArfGEFs suggests the presence of three Sec7 domain-containing proteins in the LECA, and coordinated duplications with Arfs and ArfGAPs

The comparative genomic analysis of ArfGEFs revealed that three GEF subfamilies were present in the LECA. However, it did not identify the number of paralogues from each subfamily that was present. The phylogenetic analyses presented here attempts to address this question. Largely, only one clade from each supergroup was present, suggesting that only one paralogue from each of the ancient subfamilies was present in the LECA. These analyses did identify expansions of nearly all ArfGEF subfamilies, indicating that lineage-specific evolution has occurred. Below is a description of the results of the phylogenetic analysis for each subfamily.

3.4.2.3.1 GBF

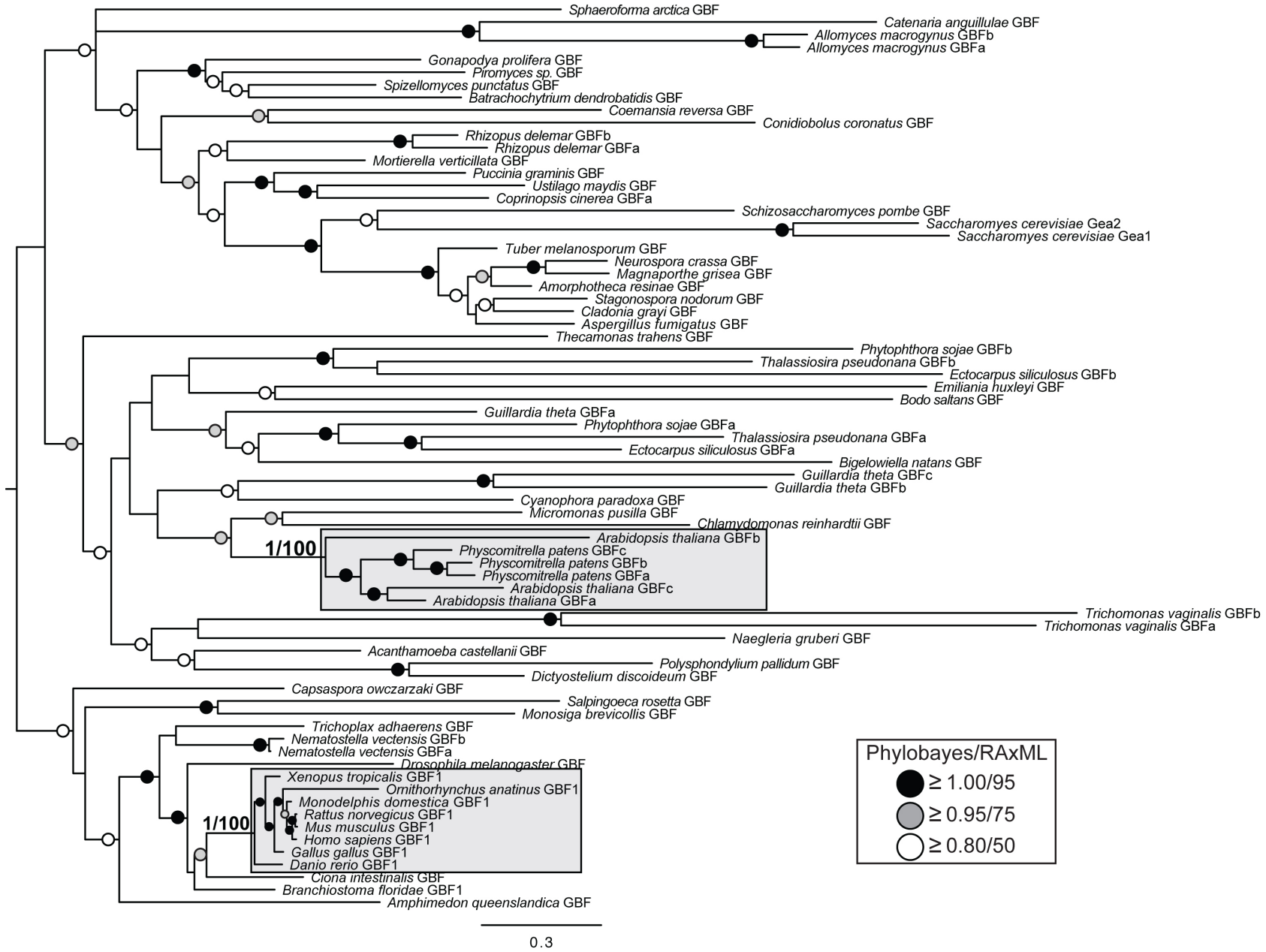
A single paralogue of GBF has been identified in humans. GBF possesses five additional domains, which it shares with BIGs (Mouratou et al., 2005). The N-terminal DCB domain is necessary for dimerization and interaction with cyclophilin in *A. thaliana* (Grebe et al., 2000). Following the DCB domain is the HUS domain (Homology Upstream of Sec7), the Sec7 domain, then the HDS1 (Homology Downstream of Sec7), HDS2, and HDS3. These domains are required for the interaction with a variety of effector proteins (García-Mata and Sztul, 2003; Saeki et

al., 2005). The HDS1 and HDS2 domains confer lipid binding to the Golgi and lipid droplets, while the DCB domain is likely important for specifically targeting GBF to the Golgi (Bouvet et al., 2013; Jackson, 2014). GBF is involved in the Arf dependent recruitment of COPI to the *cis*-Golgi and the ERGIC and is able to interact with both class I and class II Arfs (Zhao et al., 2006). In *A. thaliana*, the GBF homologue GNOM localizes to endosomes to carry out GEF activity. No such localization has been observed for human GBF. Previous phylogenetic analysis identified GBF in diverse eukaryotic lineages, suggesting its presence in the LECA (Bui et al., 2009; Cox et al., 2004). Phylogenetic analysis of GBF did not uncover any vertebrate expansions of this subfamily, with only a single paralogue identified in vertebrates (Figure 3-29). At least one small expansion is apparent in archaeplastids, in addition to multiple lineage-specific expansions.

3.4.2.3.2 BIG

Two BIG paralogues are present in humans. BIGs possess the same five accessory domains as GBF (DCB, HUS, HDS1-3), in addition to an extra HDS domain (HDS4), which has been shown to play an important regulatory role during BIG activation (McDonold and Fromme, 2014; Mouratou et al., 2005). Along with GBF, BIG is sensitive to the fungal toxin Brefeldin A (BFA; Casanova, 2007). BIGs act at the TGN and at recycling endosomes and have been implicated in regulating the trafficking of AP-1 and GGAs (Shinotsuka et al., 2002a; Shinotsuka et al., 2002b). BIGs have also been identified in diverse eukaryotic lineages, suggesting its presence in the LECA (Bui et al., 2009; Cox et al., 2004). Multiple expansions of BIG

Figure 3-29. Phylogenetic analysis of GBF sequences identifies that a single paralogue was present in the LECA. In order to determine the number of paralogues of GBF present in the LECA, phylogenetic analysis was undertaken. Generally, a single clade of each supergroup is reconstructed, suggesting that a single GBF sequence was present in the LECA. The GBF phylogeny reveals no vertebrate expansions (lower grey box). At least one expansion in archaeplastids (upper grey box) has occurred. Other lineage-specific expansions are visible (*e.g., Allomyces macrogynus, T. vaginalis*). The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (Phylobayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95, closed light circles ≥ 0.95/75, open circles ≥ 0.8/50.

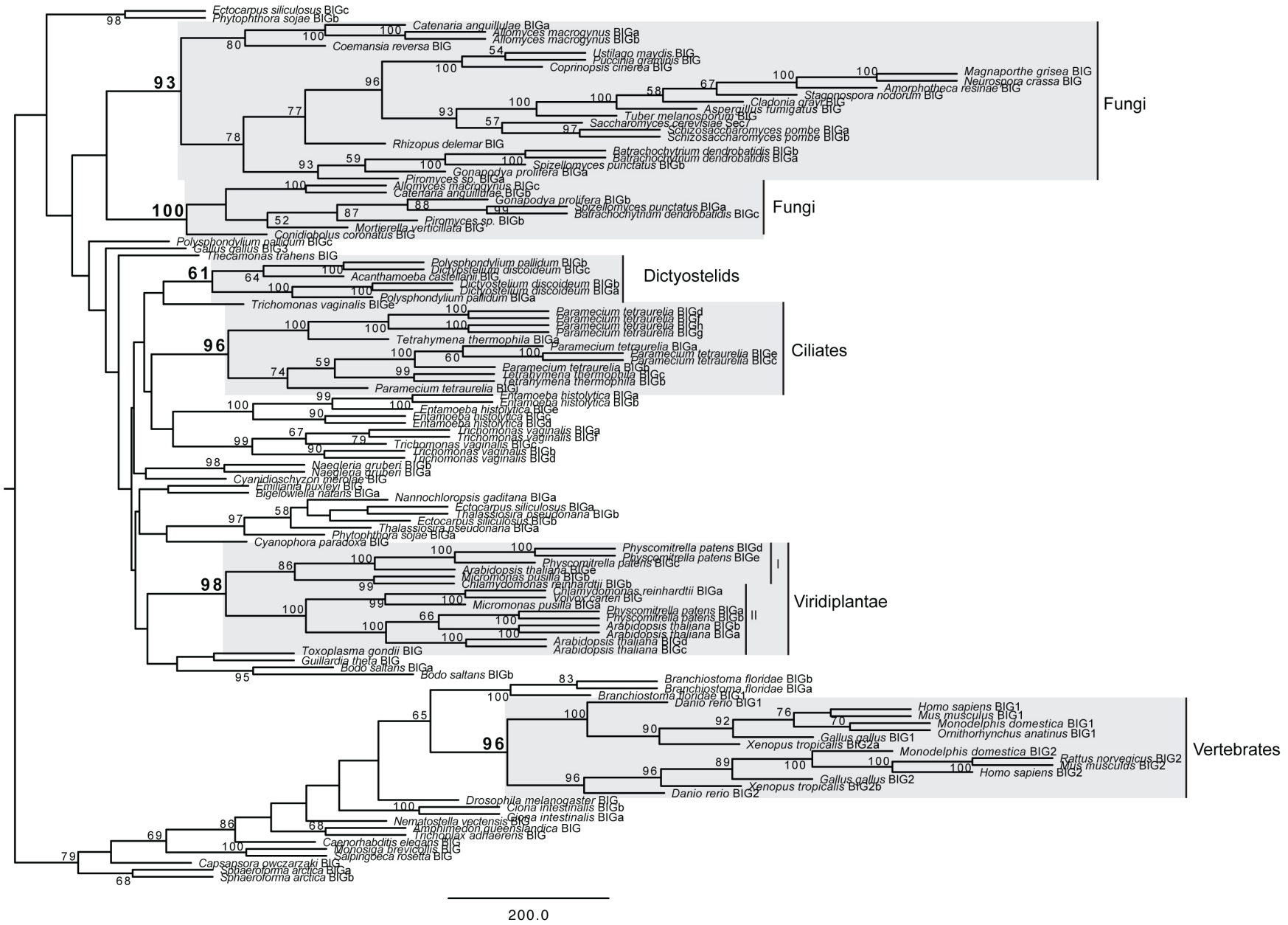

Sphaeroforma arctica GBF
Catenaria anguillulae GBF
Allomyces macrogynus GBFb
Allomyces macrogynus GBFa
Gonapodya prolifera GBF
Piromyces sp. GBF
Spizellomyces punctatus GBF
Batrachochytrium dendrobatidis GBF
Coemansia reversa GBF
Conidiobolus coronatus GBF
Rhizopus delemar GBFb
Rhizopus delemar GBFa
Mortierella verticillata GBF
Puccinia graminis GBF
Ustilago maydis GBF
Coprinopsis cinerea GBFa
Schizosaccharomyces pombe GBF
Saccharomyes cerevisiae Gea2
Saccharomyes cerevisiae Gea1
Tuber melanosporum GBF
Neurospora crassa GBF
Magnaporthe grisea GBF
Amorphotheca resinae GBF
Stagonospora nodorum GBF
Cladonia grayi GBF
Aspergillus fumigatus GBF
Thecamonas trahens GBF
Phytophthora sojae GBFb
Thalassiosira pseudonana GBFb
Ectocarpus siliculosus GBFb
Emiliania huxleyi GBF
Bodo saltans GBF
Guillardia theta GBFa
Phytophthora sojae GBFa
Thalassiosira pseudonana GBFa
Ectocarpus siliculosus GBFa
Bigelowiella natans GBF
Guillardia theta GBFc
Guillardia theta GBFb
Cyanophora paradoxa GBF
Micromonas pusilla GBF
Chlamydomonas reinhardtii GBF
1/100
Arabidopsis thaliana GBFb
Physcomitrella patens GBFc
Physcomitrella patens GBFb
Physcomitrella patens GBFa
Arabidopsis thaliana GBFc
Arabidopsis thaliana GBFa
Trichomonas vaginalis GBFb
Trichomonas vaginalis GBFa
Naegleria gruberi GBF
Acanthamoeba castellanii GBF
Polysphondylium pallidum GBF
Dictyostelium discoideum GBF
Capsaspora owczarzaki GBF
Salpingoeca rosetta GBF
Monosiga brevicollis GBF
Trichoplax adhaerens GBF
Nematostella vectensis GBFb
Nematostella vectensis GBFa
Drosophila melanogaster GBF
Xenopus tropicalis GBF1
Ornithorhynchus anatinus GBF1
Monodelphis domestica GBF1
1/100
Rattus norvegicus GBF1
Mus musculus GBF1
Homo sapiens GBF1
Gallus gallus GBF1
Danio rerio GBF1
Ciona intestinalis GBF
Branchiostoma floridae GBF1
Amphimedon queenslandica GBF
0.3
Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50

have occurred throughout eukaryotes (Figure 3-30). A gene duplication at the base of vertebrates has occurred producing the two paralogues observed in mammalian cell systems. Five BIGs have previously been identified in multicellular plants (Cox et al., 2004; Mouratou et al., 2005). Consistent with this finding, both multicellular plants included in this analysis, *A. thaliana* and *P. patens*, possess five BIG proteins (Figure 3-30). However, this is likely due to convergent evolution. A single gene duplication event in the ancestor of Viridiplantae appears to have produced two BIG paralogues. One lineage (labeled 'I' in Figure 3-30) contains a single *A. thaliana* BIG paralogue and three *P. patens* BIG paralogues. In the second lineage (labeled 'II' in Figure 3-30) there are four *A. thaliana* paralogues and two *P. patens* paralogues. In both lineages, it would seem that increases in paralogue number is likely the result of lineage-specific expansion, not multiple ancient duplications again, suggesting convergent evolution. This will be important to bear in mind when translating functional information between species, as closely related sequences may possess different functions, resulting from multiple instances of sub- or neo-functionalization of paralogues in these two species. Other lineage-specific expansions have also produced multiple paralogues for example, ciliates (with subsequent expansions in *Paramecium tetraurelia*), dictyostelids, and fungi. This analysis nonetheless suggests that a single BIG paralogue was present in the LECA.

3.4.2.3.3 BRAG

Three BRAG paralogues are found in humans, each containing an IQ-motif, a Sec7 domain, a PH domain, and at least one coiled-coil domain (Someya et al., 2001).

Figure 3-30. Phylogenetic analysis of BIGs identifies a single paralogue present in the LECA. Phylogenetic analysis of BIG sequences from all supergroups was carried out to determine the number of BIG paralogues in the LECA. This analysis reconstructed a single clade for each supergroup, indicating that only one BIG sequence was present in the LECA. Multiple lineage-specific expansions were also reconstructed (grey boxes), indicating that lineage-specific expansion has occurred multiple times. Notably, major expansions have occurred in Fungi, Dictyostelids, Ciliates, and Vertebrates. Duplication in Viridiplantae has produced two clades labelled I and II that have further expanded, giving rise to the five BIGs found in multicellular plants. Only the RAxML tree is shown, as the Phylobayes analysis did not reach convergence. All nodes with bootstrap values of at least 50% are shown.

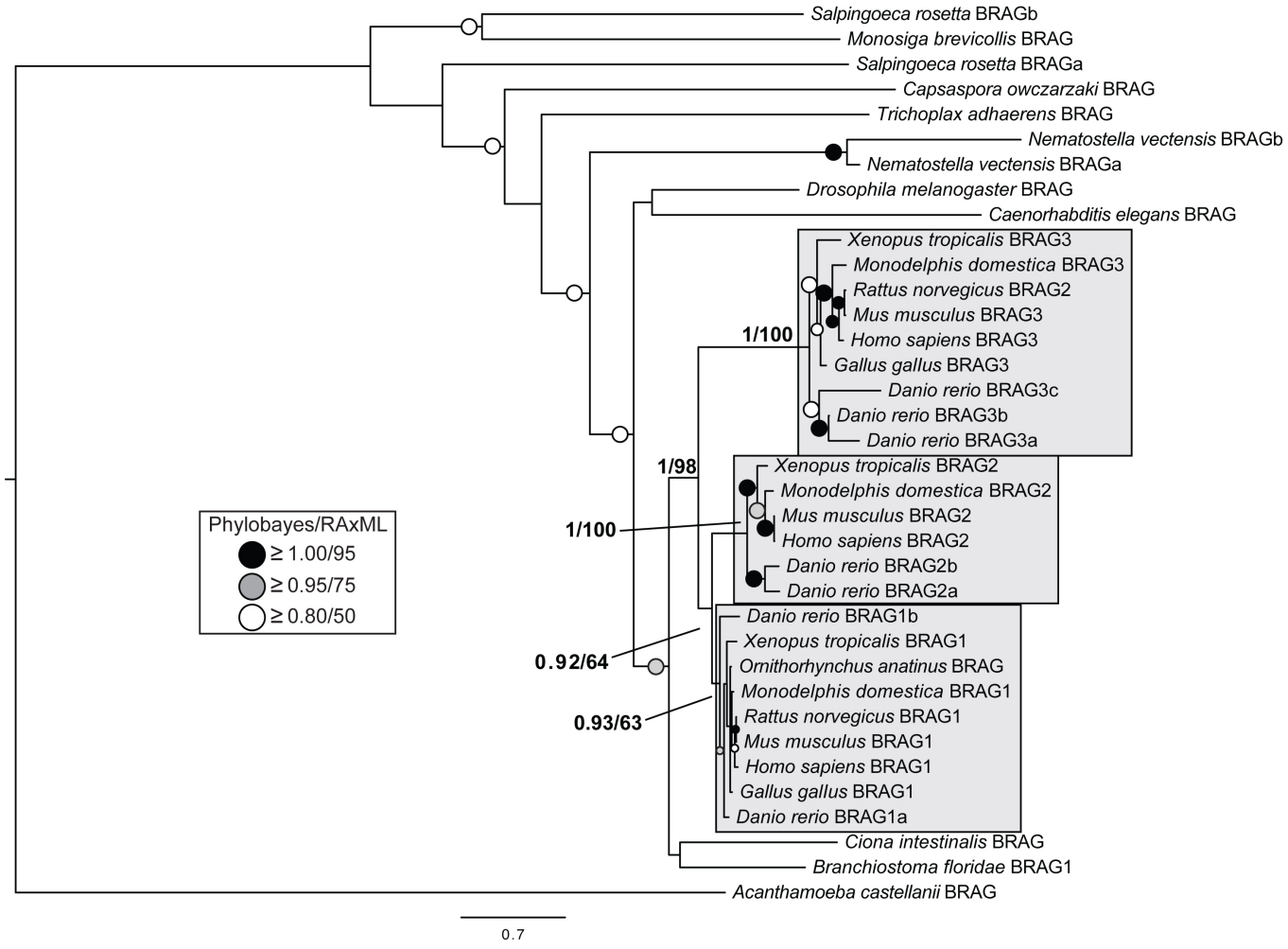

Ectocarpus siliculosus BIGc
Phytophthora sojae BIGb
Catenaria anguillulae BIGa
Allomyces macrogynus BIGa
Allomyces macrogynus BIGb
Coemansia reversa BIG
Ustilago maydis BIG
Puccinia graminis BIG
Coprinopsis cinerea BIG
Magnaporthe grisea BIG
Neurospora crassa BIG
Amorphotheca resinae BIG
Stagonospora nodorum BIG
Cladonia grayi BIG
Aspergillus fumigatus BIG
Tuber melanosporum BIG
Saccharomyces cerevisiae SecJ
Schizosaccharomyces pombe BIGa
Schizosaccharomyces pombe BIGb
Rhizopus delemar BIG
Batrachochytrium dendrobatidis BIGb
Batrachochytrium dendrobatidis BIGa
Spizellomyces punctatus BIGb
Gonapodya prolifera BIGa
Piromyces sp. BIGa
Allomyces macrogynus BIGc
Catenaria anguillulae BIGb
Gonapodya prolifera BIGb
Spizellomyces punctatus BIGa
Batrachochytrium dendrobatidis BIGc
Piromyces sp. BIGb
Mortierella verticillata BIG
Conidiobolus coronatus BIG
Fungi
Fungi
Polysphondylium pallidum BIGc
Gallus gallus BIG3
Thecamonas trahens BIG
Polysphondylium pallidum BIGb
Dictyostelium discoideum BIGc
Acanthamoeba castellanii BIG
Dictyostelium discoideum BIGb
Dictyostelium discoideum BIGa
Polysphondylium pallidum BIGa
Dictyostelids
Trichomonas vaginalis BIGe
Paramecium tetraurelia BIGd
Paramecium tetraurelia BIGf
Paramecium tetraurelia BIGh
Paramecium tetraurelia BIGg
Tetrahymena thermophila BIGa
Paramecium tetraurelia BIGa
Paramecium tetraurelia BIGe
Paramecium tetraurelia BIGc
Paramecium tetraurelia BIGb
Tetrahymena thermophila BIGc
Tetrahymena thermophila BIGb
Paramecium tetraurelia BIGj
Ciliates
Entamoeba histolytica BIGa
Entamoeba histolytica BIGb
Entamoeba histolytica BIGe
Entamoeba histolytica BIGd
Trichomonas vaginalis BIGa
Trichomonas vaginalis BIGf
Trichomonas vaginalis BIGc
Trichomonas vaginalis BIGb
Trichomonas vaginalis BIGd
Naegleria gruberi BIGb
Naegleria gruberi BIGa
Cyanidioschyzon merolae BIG
Emiliania huxleyi BIG
Bigelowiella natans BIGa
Nannochloropsis gaditana BIGa
Ectocarpus siliculosus BIGa
Thalassiosira pseudonana BIGb
Ectocarpus siliculosus BIGb
Thalassiosira pseudonana BIGa
Phytophthora sojae BIGa
Cyanophora paradoxa BIG
Physcomitrella patens BIGd
Physcomitrella patens BIGe
Physcomitrella patens BIGc
I
Arabidopsis thaliana BIGe
Micromonas pusilla BIGb
Chlamydomonas reinhardtii BIGb
Chlamydomonas reinhardtii BIGa
Volvox carteri BIG
Micromonas pusilla BIGa
Physcomitrella patens BIGa
Physcomitrella patens BIGb
Arabidopsis thaliana BIGb
Arabidopsis thaliana BIGa
II
Arabidopsis thaliana BIGd
Arabidopsis thaliana BIGc
Viridiplantae
Toxoplasma gondii BIG
Guillardia theta BIG
Bodo saltans BIGa
Bodo saltans BIGb
Branchiostoma floridae BIGb
Branchiostoma floridae BIGa
Branchiostoma floridae BIG1
Danio rerio BIG1
Homo sapiens BIG1
Mus musculus BIG1
Monodelphis domestica BIG1
Ornithorhynchus anatinus BIG1
Gallus gallus BIG1
Xenopus tropicalis BIG2a
Monodelphis domestica BIG2
Rattus norvegicus BIG2
Mus musculus BIG2
Homo sapiens BIG2
Gallus gallus BIG2
Xenopus tropicalis BIG2b
Danio rerio BIG2
Vertebrates
Drosophila melanogaster BIG
Ciona intestinalis BIGb
Ciona intestinalis BIGa
Nematostella vectensis BIG
Amphimedon queenslandica BIG
Trichoplax adhaerens BIG
Caenorhabditis elegans BIG
Monosiga brevicollis BIG
Salpingoeca rosetta BIG
Capsaspora owczarzaki BIG
Sphaeroforma arctica BIGa
Sphaeroforma arctica BIGb
200.0

Similar to EFA6 and cytohesin, BRAGs are BFA resistant GEFs and have preferential activity towards Arf6 (Franco et al., 1999; Meacci et al., 1997; Someya et al., 2001). BRAGs are located at the cell periphery, at post-synaptic densities in neuronal cells, and at the plasma membrane in non-neuronal cells where they selectively regulate the endocytosis of specific cargoes (*e.g.,* β-integrins; Dunphy et al., 2006). To date BRAGs have only been identified in animals (Cox et al., 2004). The BRAG phylogeny revealed an expansion of this subfamily in vertebrates; at least two gene duplications have produced three vertebrate paralogues (Figure 3-31). However, *Ornithorhynchus anatinus* appears to have lost BRAG2 and BRAG3. *R. norvegicus* and *G. gallus* also appear to have lost BRAG2. Independent duplications can also be seen for *N. vectensis* and *S. rosetta*. The duplication in *S. rosetta* did not resolve, likely due to the long branch of the *A. castellanii* BRAG homologue.

3.4.2.3.4 Cytohesin

Four cytohesin paralogues are present in humans, each with a coiled-coil domain, a Sec7 domain, and a PH domain (Ogasawara et al., 2000). Cytohesins are primarily located at the cell periphery, where they are recruited to the plasma membrane by the interaction of their PH domain with either $PI(3,4,5)P_3$ or $PI(4,5)P_2$, where they act preferentially as Arf6 GEFs (Klarlund et al., 1997). Cytohesins have been shown to regulate both the docking and fusion of secretory granules, and the endocytosis of G-protein coupled receptors (Claing et al., 2001; Liu et al., 2005a). They also play an important role in integrin-mediated adhesion and

Figure 3-31. Phylogenetic analysis of BRAG identifies an expansion in vertebrates. Phylogenetic analysis of BRAG was carried out to determine when the duplication events producing the three vertebrate paralogues occurred. The analysis reveals two duplications at the base of vertebrates, resulting in three paralogues with BRAG3 diverging first, followed by BRAG1 and BRAG2 (grey boxes). Duplications have also occurred in *N. vectensis* and *S. rosetta*. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (Phylobayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95, closed light circles ≥ 0.95/75, open circles ≥ 0.8/50.


Salpingoeca rosetta BRAGb
Monosiga brevicollis BRAG
Salpingoeca rosetta BRAGa
Capsaspora owczarzaki BRAG
Trichoplax adhaerens BRAG
Nematostella vectensis BRAGb
Nematostella vectensis BRAGa
Drosophila melanogaster BRAG
Caenorhabditis elegans BRAG
Xenopus tropicalis BRAG3
Monodelphis domestica BRAG3
Rattus norvegicus BRAG2
Mus musculus BRAG3
Homo sapiens BRAG3
Gallus gallus BRAG3
Danio rerio BRAG3c
Danio rerio BRAG3b
Danio rerio BRAG3a
Xenopus tropicalis BRAG2
Monodelphis domestica BRAG2
Mus musculus BRAG2
Homo sapiens BRAG2
Danio rerio BRAG2b
Danio rerio BRAG2a
Danio rerio BRAG1b
Xenopus tropicalis BRAG1
Ornithorhynchus anatinus BRAG
Monodelphis domestica BRAG1
Rattus norvegicus BRAG1
Mus musculus BRAG1
Homo sapiens BRAG1
Gallus gallus BRAG1
Danio rerio BRAG1a
Ciona intestinalis BRAG
Branchiostoma floridae BRAG1
Acanthamoeba castellanii BRAG
1/100
1/98
1/100
0.92/64
0.93/63
Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50
0.7

cell movement (Geiger et al., 2000). Cytohesins were previously identified as metazoan-specific ArfGEFs (Cox et al., 2004). Phylogenetic analysis of cytohesin identified multiple expansions including the Amoebozoa (Figure 3-32) and in vertebrates (Figure 3-32, 3-33). Secondary loss does seem to have occurred multiple times as *O. anatinus* is missing all but cytohesin1. *Monodelphis domestica* and *G. gallus* are also missing cytohesin2.

3.4.2.3.5 EFA6

Four EFA6 paralogues are present in humans, each possessing a Sec7 domain, followed by a PH domain, and are reported to possess a coiled-coil domain C-terminal to the PH domain, along with proline rich regions distributed throughout the protein (Franco et al., 1999). EFA6 is a plasma membrane localized ArfGEF that has been shown to interact with both Arf1 and Arf6 (Macia et al., 2001; Padovani et al., 2014). EFA6 is involved in the coordination of endocytosis, actin and microtubule dynamics, and the maintenance of cellular junctions (Franco et al., 1999; Klein et al., 2008). Previous analyses have proposed that EFA6 is present only in animals (Cox et al., 2004), although the yeast proteins Syt1 and Syt2 are quite similar to EFA6. Both possess a central Sec7 domain followed by a PH domain and a region of high sequence similarity to EFA6 (Cox et al., 2004; Gillingham and Munro, 2007a), suggesting that these subfamilies are fungal-specific expansions of the EFA6 subfamily. Syt1p is an Arf2p GEF and has been shown to possess GEF activity towards the Arf-like protein Arl1p both *in vitro* and *in vivo* (Chen et al., 2010). Yel1p is an Arf3p (*S. cerevisiae* orthologue of Arf6) GEF that targets the GTPase to the

Figure 3-32. Phylogenetic analysis of cytohesin identifies a single paralogue in the LECA. Phylogenetic analysis of cytohesin sequences from all eukaryotic supergroups was carried out to determine the number of paralogues present in the LECA. The analysis reconstructed a single clade for each supergroup, indicating that a single paralogue was present in the LECA. The vertebrate expansion is weakly supported here (lower four grey boxes). An unsupported expansion has occurred in Amoebozoa (upper grey box). Additional genome-specific expansions have also occurred (*e.g., T. trahens* and *A. macrogynus*). Only the RAxML tree is shown, as the Phylobayes analysis did not reach convergence. All nodes with bootstrap values of at least 50% are shown.

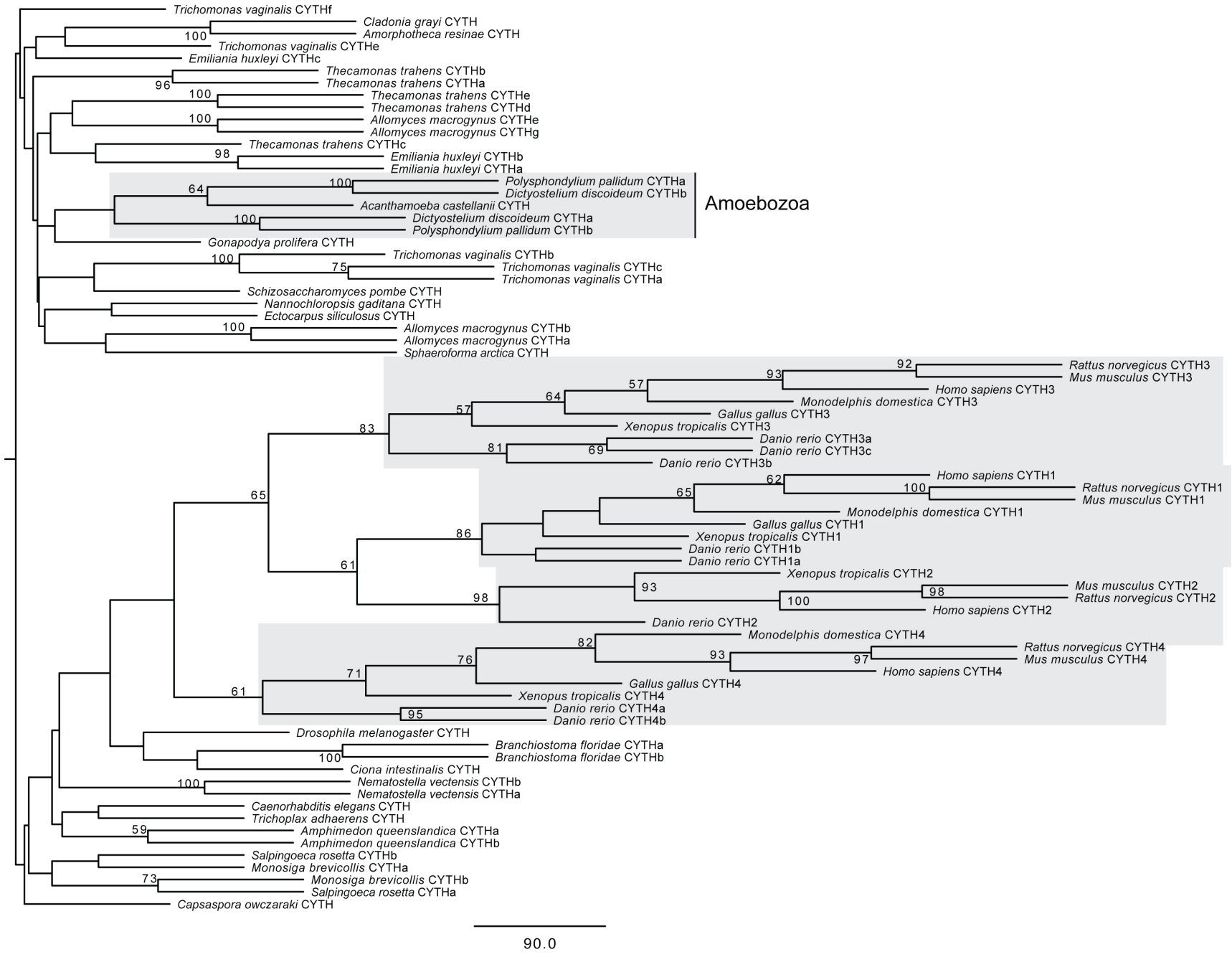
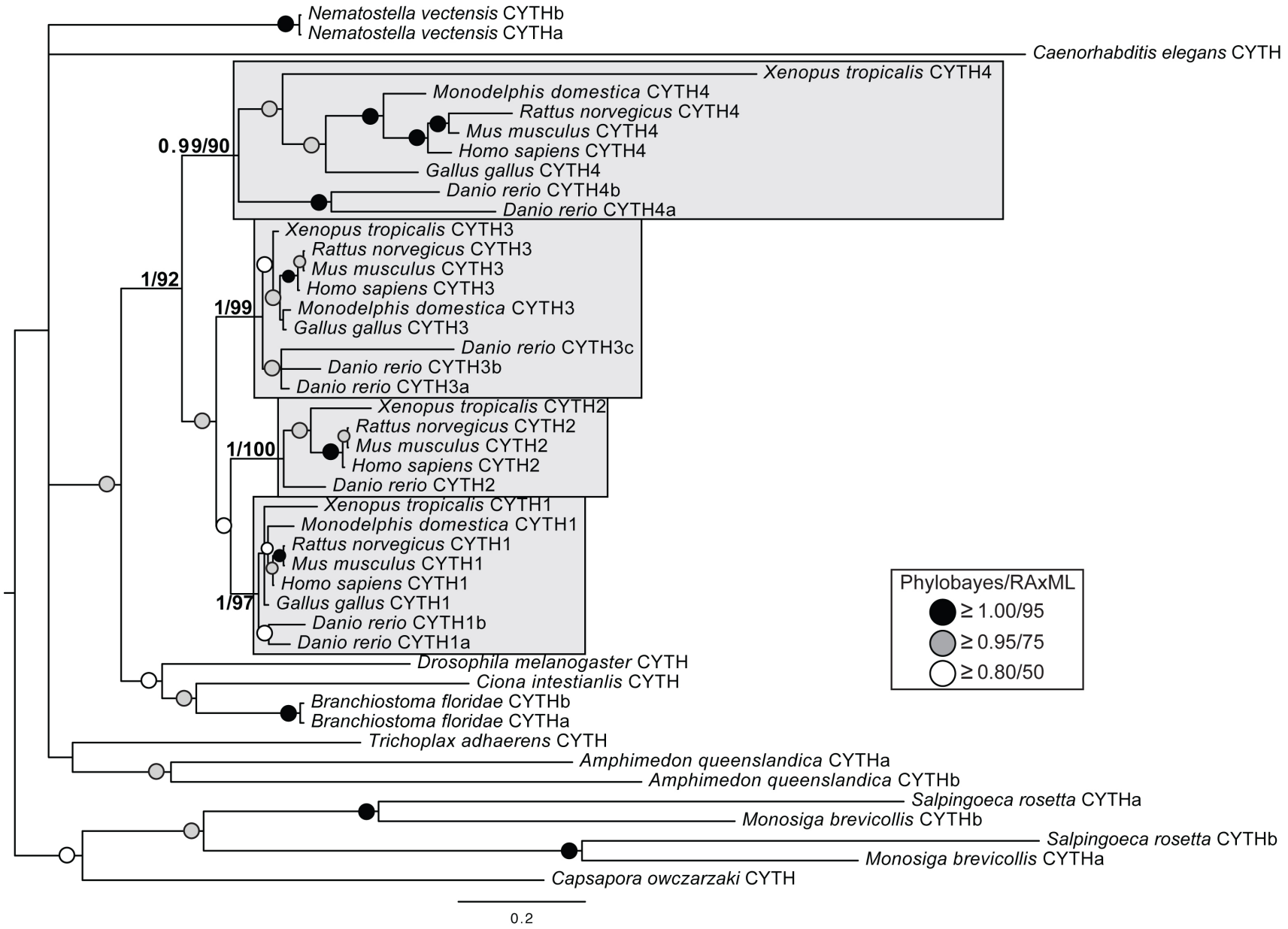

Trichomonas vaginalis CYTHf
Cladonia grayi CYTH
Amorphotheca resinae CYTH
Trichomonas vaginalis CYTHe
Emiliania huxleyi CYTHc
Thecamonas trahens CYTHb
Thecamonas trahens CYTHa
Thecamonas trahens CYTHe
Thecamonas trahens CYTHd
Allomyces macrogynus CYTHe
Allomyces macrogynus CYTHg
Thecamonas trahens CYTHc
Emiliania huxleyi CYTHb
Emiliania huxleyi CYTHa
Polysphondylium pallidum CYTHa
Dictyostelium discoideum CYTHb
Acanthamoeba castellanii CYTH
Dictyostelium discoideum CYTHa
Polysphondylium pallidum CYTHb
Amoebozoa
Gonapodya prolifera CYTH
Trichomonas vaginalis CYTHb
Trichomonas vaginalis CYTHc
Trichomonas vaginalis CYTHa
Schizosaccharomyces pombe CYTH
Nannochloropsis gaditana CYTH
Ectocarpus siliculosus CYTH
Allomyces macrogynus CYTHb
Allomyces macrogynus CYTHa
Sphaeroforma arctica CYTH
Rattus norvegicus CYTH3
Mus musculus CYTH3
Homo sapiens CYTH3
Monodelphis domestica CYTH3
Gallus gallus CYTH3
Xenopus tropicalis CYTH3
Danio rerio CYTH3a
Danio rerio CYTH3c
Danio rerio CYTH3b
Homo sapiens CYTH1
Rattus norvegicus CYTH1
Mus musculus CYTH1
Monodelphis domestica CYTH1
Gallus gallus CYTH1
Xenopus tropicalis CYTH1
Danio rerio CYTH1b
Danio rerio CYTH1a
Xenopus tropicalis CYTH2
Mus musculus CYTH2
Rattus norvegicus CYTH2
Homo sapiens CYTH2
Danio rerio CYTH2
Monodelphis domestica CYTH4
Rattus norvegicus CYTH4
Mus musculus CYTH4
Homo sapiens CYTH4
Gallus gallus CYTH4
Xenopus tropicalis CYTH4
Danio rerio CYTH4a
Danio rerio CYTH4b
Drosophila melanogaster CYTH
Branchiostoma floridae CYTHa
Branchiostoma floridae CYTHb
Ciona intestinalis CYTH
Nematostella vectensis CYTHb
Nematostella vectensis CYTHa
Caenorhabditis elegans CYTH
Trichoplax adhaerens CYTH
Amphimedon queenslandica CYTHa
Amphimedon queenslandica CYTHb
Salpingoeca rosetta CYTHb
Monosiga brevicollis CYTHa
Monosiga brevicollis CYTHb
Salpingoeca rosetta CYTHa
Capsaspora owczaraki CYTH
90.0

Figure 3-33. Phylogenetic analysis of cytohesin identifies four paralogues in vertebrates. Phylogenetic analysis of filozoan cytohesin sequences was carried out to determine the order and the timing of paralogue divergence relative to the vertebrate transition. The analysis supports three gene duplications producing four paralogues in vertebrates (grey boxes). Cytohesin4 diverged first, then cytohesin3, followed by cytohesin1 and cytohesin2. Lineage-specific expansions in other organisms (*N. vectensis, A. queenslandica,* choanoflagellates, and *B. floridae*) have also occurred. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (Phylobayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95, closed light circles ≥ 0.95/75, open circles ≥ 0.8/50.


Nematostella vectensis CYTHb
Nematostella vectensis CYTHa
Caenorhabditis elegans CYTH
Xenopus tropicalis CYTH4
Monodelphis domestica CYTH4
Rattus norvegicus CYTH4
Mus musculus CYTH4
Homo sapiens CYTH4
Gallus gallus CYTH4
Danio rerio CYTH4b
Danio rerio CYTH4a
0.99/90
Xenopus tropicalis CYTH3
Rattus norvegicus CYTH3
Mus musculus CYTH3
Homo sapiens CYTH3
Monodelphis domestica CYTH3
Gallus gallus CYTH3
Danio rerio CYTH3c
Danio rerio CYTH3b
Danio rerio CYTH3a
1/92
1/99
Xenopus tropicalis CYTH2
Rattus norvegicus CYTH2
Mus musculus CYTH2
Homo sapiens CYTH2
Danio rerio CYTH2
1/100
Xenopus tropicalis CYTH1
Monodelphis domestica CYTH1
Rattus norvegicus CYTH1
Mus musculus CYTH1
Homo sapiens CYTH1
Gallus gallus CYTH1
Danio rerio CYTH1b
Danio rerio CYTH1a
1/97
Drosophila melanogaster CYTH
Ciona intestianlis CYTH
Branchiostoma floridae CYTHb
Branchiostoma floridae CYTHa
Trichoplax adhaerens CYTH
Amphimedon queenslandica CYTHa
Amphimedon queenslandica CYTHb
Salpingoeca rosetta CYTHa
Monosiga brevicollis CYTHb
Salpingoeca rosetta CYTHb
Monosiga brevicollis CYTHa
Capsapora owczarzaki CYTH
0.2
Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50

plasma membrane where it regulates the polymerization of actin patches and polarized cell growth (Gillingham and Munro, 2007a). Phylogenetic analysis of the EFA6 subfamily revealed a major expansion in vertebrates resulting in four paralogues (Figure 3-34); however, secondary loss of some of these paralogues appears to have occurred, particularly in *O. anatinus*, which only possess EFA6D. Additionally, *Xenopus tropicalis* is missing EFA6C, and both *G. gallus* and *M. domestica* lack EFA6B.

3.4.2.3.6 FBX8

Little is known about FBX8; humans possess only one paralogue that contains an N-terminal F-box domain and a C-terminal Sec7 domain (Cox et al., 2004). F-box domains incorporate proteins into multisubunit ubiquitin-ligase complexes, resulting in their degradation (Kipreos and Pagano, 2000). FBX8 is thought to suppress the activity of Arf6 thorough ubiquitination (Yano et al., 2008). FBX8 has previously only been found in vertebrates (Cox et al., 2004). The phylogeny for FBX8 is much simpler than that of the other subfamilies as no vertebrate duplications have occurred. However, independent duplications are observed for *N. vectensis* and for *Branchiostoma floridae* (Figure 3-35).

3.5 Discussion

The analyses presented in this chapter revealed the evolutionary patterns of the previously identified ArfGAP and ArfGEF proteins. These analyses updated the timing of the origin for four of ten ArfGAP subfamilies (Kahn et al., 2008) and

Figure 3-34. Phylogenetic analysis of EFA6 subfamily identifies four paralogues in vertebrates. Phylogenetic analysis of EFA6 was carried out to determine the order and timing of gene duplication relative to the vertebrate transition. The EFA6 phylogeny reveals three gene duplications resulting in four vertebrate paralogues (grey boxes) with EFA6B diverging first, followed by EFA6A, then by EFA6C and EFA6D. Other lineage-specific duplications have also occurred, such as in *R. delemar* and *A. macrogynus*. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (Phylobayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95, closed light circles ≥ 0.95/75, open circles ≥ 0.8/50.

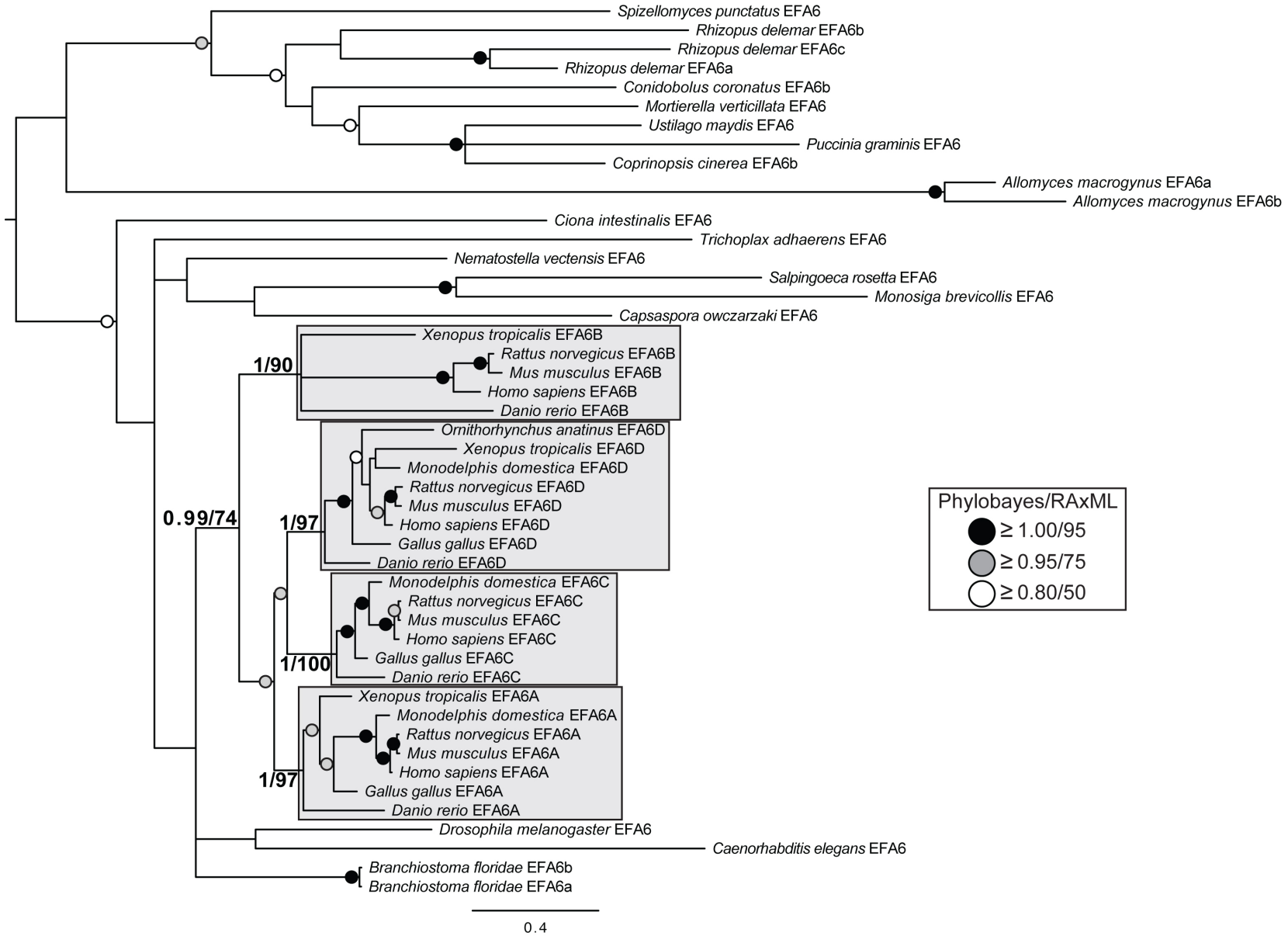
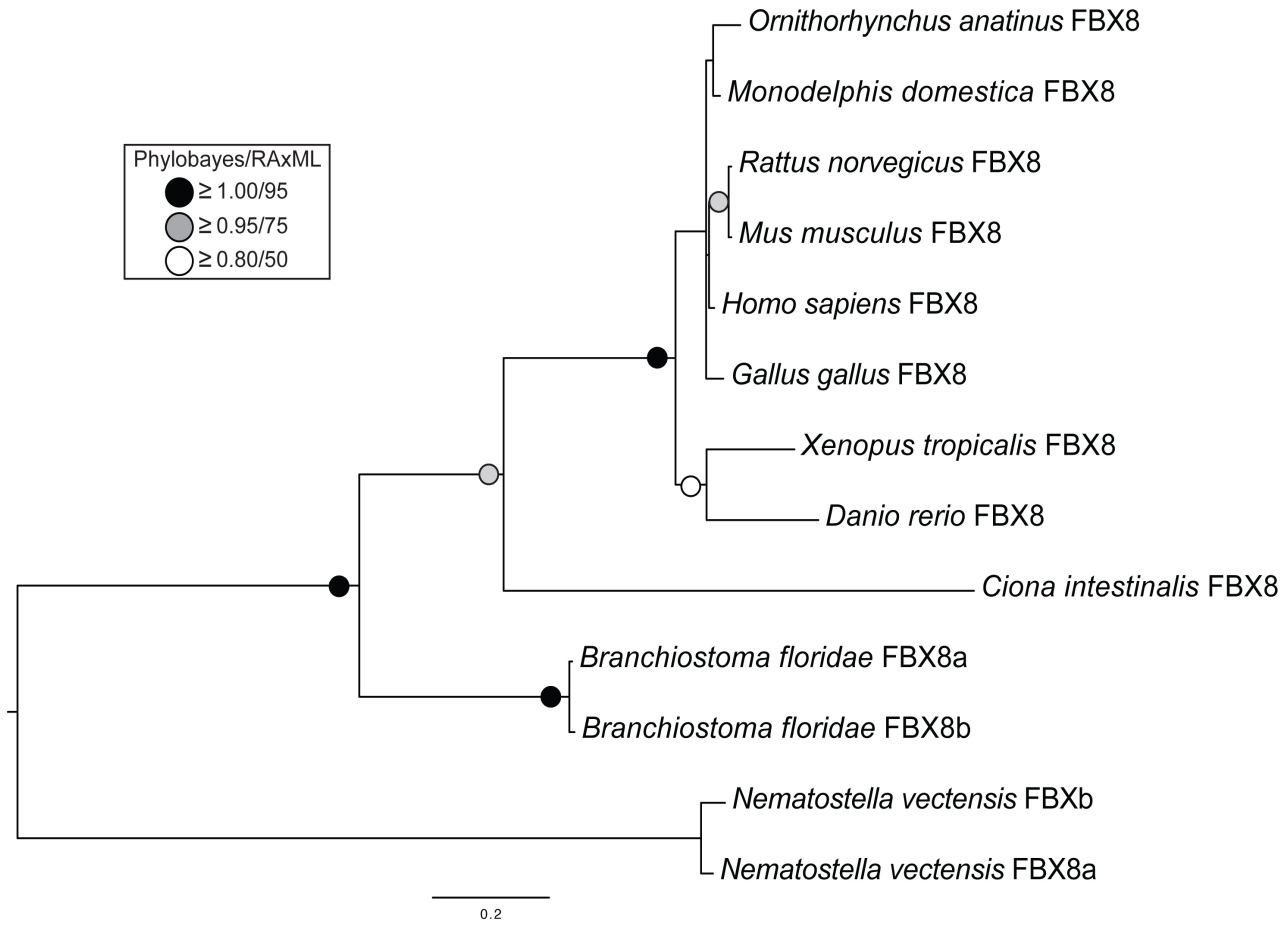

Spizellomyces punctatus EFA6
Rhizopus delemar EFA6b
Rhizopus delemar EFA6c
Rhizopus delemar EFA6a
Conidobolus coronatus EFA6b
Mortierella verticillata EFA6
Ustilago maydis EFA6
Puccinia graminis EFA6
Coprinopsis cinerea EFA6b
Allomyces macrogynus EFA6a
Allomyces macrogynus EFA6b
Ciona intestinalis EFA6
Trichoplax adhaerens EFA6
Nematostella vectensis EFA6
Salpingoeca rosetta EFA6
Monosiga brevicollis EFA6
Capsaspora owczarzaki EFA6
Xenopus tropicalis EFA6B
Rattus norvegicus EFA6B
Mus musculus EFA6B
Homo sapiens EFA6B
Danio rerio EFA6B
1/90
Ornithorhynchus anatinus EFA6D
Xenopus tropicalis EFA6D
Monodelphis domestica EFA6D
Rattus norvegicus EFA6D
Mus musculus EFA6D
Homo sapiens EFA6D
Gallus gallus EFA6D
Danio rerio EFA6D
0.99/74
1/97
Monodelphis domestica EFA6C
Rattus norvegicus EFA6C
Mus musculus EFA6C
Homo sapiens EFA6C
Gallus gallus EFA6C
Danio rerio EFA6C
1/100
Xenopus tropicalis EFA6A
Monodelphis domestica EFA6A
Rattus norvegicus EFA6A
Mus musculus EFA6A
Homo sapiens EFA6A
Gallus gallus EFA6A
Danio rerio EFA6A
1/97
Drosophila melanogaster EFA6
Caenorhabditis elegans EFA6
Branchiostoma floridae EFA6b
Branchiostoma floridae EFA6a
0.4
Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50

Figure 3-35. Phylogenetic analysis of FBX8 reveals a single paralogue in vertebrates. Phylogenetic analysis of FBX8 was carried out to determine whether or not any gene duplications had occurred. The FBX8 phylogeny reveals no vertebrate expansions; however, independent duplications in *B. floridae* and *N. vectensis* were observed. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (Phylobayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95, closed light circles ≥ 0.95/75, open circles ≥ 0.8/50.


Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50
Ornithorhynchus anatinus FBX8
Monodelphis domestica FBX8
Rattus norvegicus FBX8
Mus musculus FBX8
Homo sapiens FBX8
Gallus gallus FBX8
Xenopus tropicalis FBX8
Danio rerio FBX8
Ciona intestinalis FBX8
Branchiostoma floridae FBX8a
Branchiostoma floridae FBX8b
Nematostella vectensis FBXb
Nematostella vectensis FBX8a
0.2

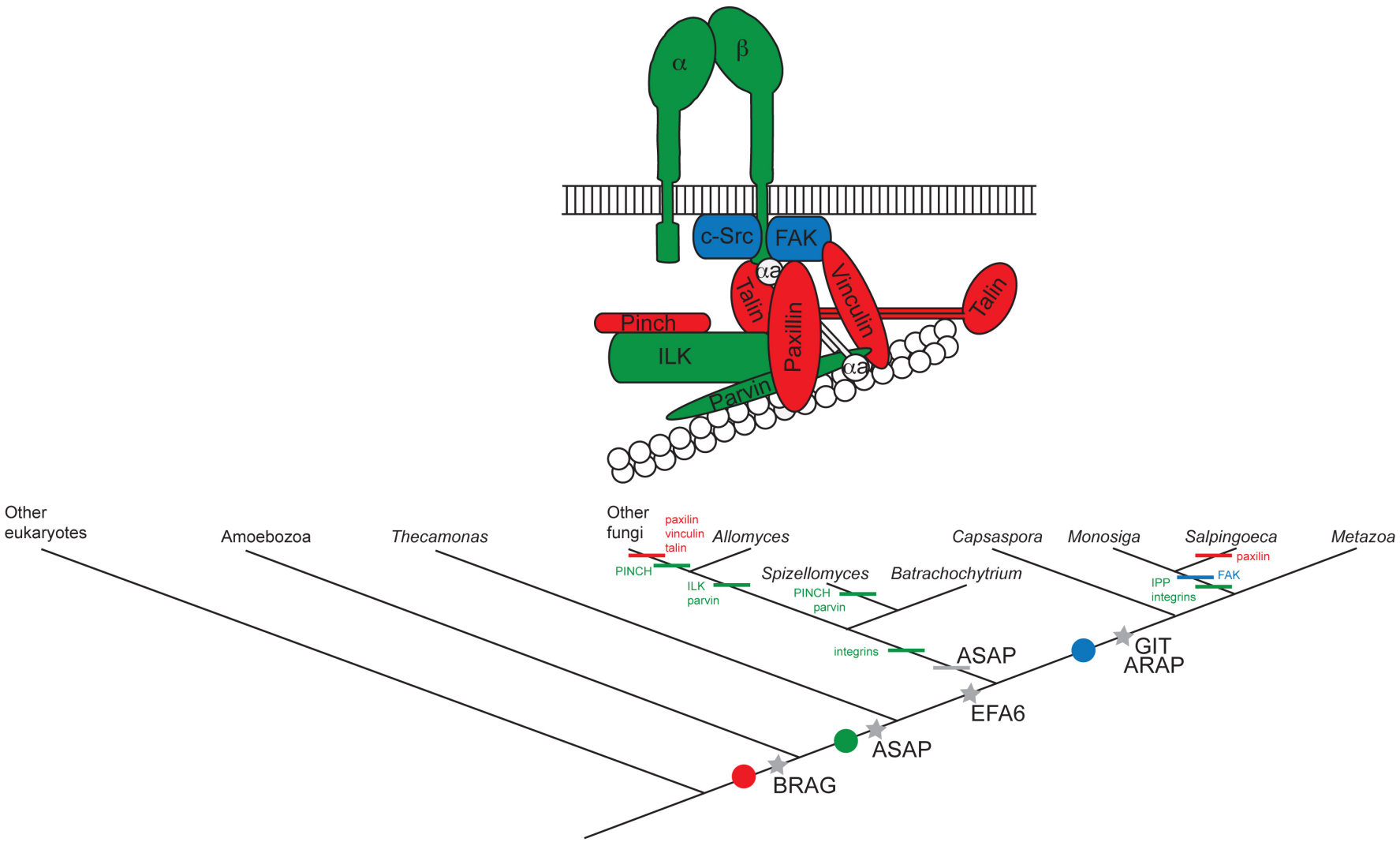
identified a previously undescribed ArfGAP subfamily, ArfGAPC2, which is broadly distributed across eukaryotes. Similarly, novel hypotheses concerning the timing of the origin for four of six human ArfGEF subfamilies (Cox et al., 2004) were generated. These results indicate that at least six ArfGAPs and three ArfGEFs were present in the LECA, with additional subfamilies emerging in various lineages.

The discovery of ArfGAPC2 increases the total number of ArfGAP subfamilies to eleven. Members of this subfamily fulfilled the 5-orders RBH criterion, and share common domain architectures: a central ArfGAP domain and a C-terminal C2 domain (Figure 3-9). Extensive *in vivo* characterization of this subfamily will be required in multiple model systems in order to thoroughly dissect the role of this novel subfamily in membrane trafficking. However, it is predicted that ArfGAPC2 does possess GAP activity; we showed previously that that 18 residues are highly conserved in all ArfGAP domains of all members of all subfamilies (Schlacht et al., 2013). Of these 18 residues, ArfGAPC2 possess 17, including the catalytic arginine. Functional characterization of ArfGAPC2 will be an important step in understanding the functional diversity of ArfGAPs, as ArfGAPC2 may be involved in processes not found in other model organisms, or may act at locations not typical of other ArfGAP proteins. Comparative genomic and phylogenetic analyses indicated that this is an ancient subfamily, present in the LECA, but that has been frequently lost (*i.e.,* patchy), yet another example of this novel class of proteins.

Five of the six recently arising ArfGAP and GEF subfamilies (ASAP, ARAP, GIT, BRAG, and EFA6) are regulators of the cytoskeleton, cell-cell communication, or cell adhesion (see Casanova, 2007; Kahn et al., 2008 and *inter alia*). The increase in the

number of GAP and GEF proteins involved in adhesion and motility processes is highly correlated with the evolution of cellular adhesion in opisthokonts. GIT controls cell migration and focal adhesion dynamics through interactions with PIX and paxillin, respectively (Hoefen and Berk, 2006; Mazaki et al., 2001; Premont et al., 2004), while ARAP regulates focal adhesion dynamics and lamellipodia formation (Inoue and Randazzo, 2007; Krugmann et al., 2006; Miura et al., 2002). The distribution of ASAP, a regulator of actin remodelling and invadopodia formation (Brown et al., 1998; Liu et al., 2002; Randazzo et al., 2007), is even more intriguing as it is found in Holozoa and the apusomonad *T. trahens*, but appears to have been lost from Fungi. EFA6, present in opisthokonts but not *T. trahens,* is also involved in regulating the actin cytoskeleton, and like ASAP is located at the plasma membrane. Both ASAP and EFA6 display preference towards class III Arfs; it would not be surprising if these two subfamilies regulate the activation and termination of the same Arf-dependent process. Additionally, *M. brevicollis* possesses the ability to attach to substrate via extracellular matrix proteins homologous to those found in humans (*e.g.,* laminin, reeler, and ependymin domains; King et al., 2008) and *C. owczarzaki* is able to form cellular aggregates using integrin-mediated adhesion, previously thought to only be present in metazoans (Sebé-pedrós et al., 2010; Suga et al., 2013). A comparative genomic analysis assessing the distribution of core integrin adhesion machinery identified the stepwise acquisition of components that produced the adhesion complexes seen in mammalian cells (Figure 3-36; Sebé-pedrós et al., 2010). They defined the core machinery to include: α and β integrins (Hynes, 2002), α-actinin (Sjöblom et al., 2008), talin (Wegener et al., 2007), paxillin

Figure 3-36. Correlated evolution of the integrin adhesion complex and ArfGAP and ArfGEF proteins. Upper: metazoan-type integrin adhesion complex. Colours correspond to the acquisitions and losses in the eukaryotic tree (lower). Circles and stars represent the gain of integrin adhesion machinery and ArfGAPs and ArfGEFs, respectively. Colour coding of circles and dashes correspond to the integrin adhesion machinery in the upper panel. Proteins in red (Pinch, Talin, Paxillin, and Vinculin) in the ancestor of the Amoebozoa, apusomonads, and Opisthokonta. Proteins in green (ILK, Parvin, and integrins α and β) evolved in the ancestor of apusomonads and Opisthokonta. Proteins in blue (c-Src and FAK) evolved in the ancestor of the Holozoa. Dashes indicate losses of the indicated machinery. Redrawn and modified from Sébé-Pedros et al. (2010).


α
β
c-Src
FAK
αa
Talin
Paxillin
Vinculin
Talin
Pinch
ILK
αa
Parvin
Other eukaryotes
Amoebozoa
Thecamonas
Other fungi
paxilin
vinculin
talin
Allomyces
Capsaspora
Monosiga
Salpingoeca
Metazoa
paxilin
PINCH
ILK
parvin
Spizellomyces
Batrachochytrium
IPP
FAK
integrins
PINCH
parvin
integrins
ASAP
GIT
ARAP
EFA6
ASAP
BRAG

(Deakin and Turner, 2008), vinculin (Ziegler et al., 2006), the IPP complex [ILK, (integrin linked kinase), PINCH (particularly interesting Cys-His-rich protein), and parvin (Legate et al., 2006; Nikolopoulos and Turner, 2001)], and the kinases c-Src (Arias-Salgado et al., 2003) and FAK (Parsons et al., 2000). These proteins appear to have evolved over the same time frame as the opisthokont-specific ArfGAPs and GEFs (Figure 3-36). PINCH, talin, paxillin, and vinculin appear to have evolved in the ancestor of Amoebozoa and Opisthokonta, while integrins α and β, ILK, and parvin evolved in the ancestor of Opisthokonta and apusomonads, whereas c-Src and FAK evolved in the ancestor of the Holozoa. α-actinin was found throughout eukaryotes. GIT evolved at roughly the same time as c-Src and FAK (Figure 3-36). GIT was originally identified as regulatory protein involved in the endocytosis of the β_2-adrenergic receptor (Premont et al., 1998). GIT is also often found in complex with PIX proteins which are GEFs for Rac, Cdc42, and Rho (Feng et al., 2002; Loo et al., 2004; Manser et al., 1998), and is phosphorylated by FAK, which results in its recruitment to focal adhesions and its interaction with paxillin (Turner et al., 1999). This association allows GIT to negatively regulate Rac1, which is able to regulate actin polymerization and the formation of lamellipodia during cell movement in an ArfGAP-dependent manner, although the mechanism through which this occurs is as yet unclear (Nishiya et al., 2005; West et al., 2001). ASAP appears to have evolved around the same time as integrins, ILK, and parvin; however, ASAP has been reported to interact with paxillin, c-Src, FAK (Brown et al., 1998; Kondo et al., 2000; Liu et al., 2002). Paxillin localization to focal adhesions was shown to be Arf1

dependent (Norman et al., 1998). Similarly, overexpression of ASAP results in a reduction of paxillin and FAK at focal adhesions and reduced cellular motility (Kondo et al., 2000; Liu et al., 2002). This suggests that ASAP was involved in regulating the localization of paxillin and perhaps extended this function to other components as they evolved. BRAG is involved in the endocytosis of β-integrins (Dunphy et al., 2006; Manavski et al., 2014; Moravec et al., 2012). Knockdown of BRAG reduces the mobility of endothelial cells, suggesting increased adhesion to substrate (Manavski et al., 2014). This is somewhat surprising since BRAG evolved prior to integrins (Figure 3-36). This may suggest that BRAG was originally involved in endocytosis and was subsequently recruited to focal adhesions after the evolution of integrins. This would be consistent with reports that BRAG interacts with AP-2 and clathrin during endocytosis (Moravec et al., 2012). By contrast, EFA6 and ARAP do not seem to be directly involved in regulating or interacting with members of the integrin adhesion complex. However, they both contribute to the regulation of the actin cytoskeleton, and therefore, may indirectly impact the stability of focal adhesions (Kanamarlapudi, 2014; Yoon et al., 2006). The overlap in timing between the evolution of integrins and their associated machinery, and the expansions of the ArfGAP and GEF proteins is highly suggestive of co-evolution between these two systems, and perhaps preadaptive to a role in multicellularity.

No additional ArfGEF subfamilies were identified, despite finding many rogue sequences (Figure 3-26), suggesting that either these rogues represent highly divergent sequences or, more likely and perhaps more interestingly, that these sequences represent multiple subfamilies that are not broadly distributed, that are

the result of lineage-specific expansion and customization in these groups. For example, in the Amoebozoa, *D. discoideum, P. pallidum,* and *A. castellanii* all possess at least one rogue sequence containing 6-8 ankyrin repeats, a Sec7 domain, a PH domain and a variable C-terminal domain. This common domain structure is suggestive of common ancestry, and perhaps functions in a process not present in other eukaryotes. However, phylogenetic tools more powerful than those used here, and classification methods not dependent on a reference genome (*i.e.,* not BLASTing against human or yeast) will be required in order to elucidate this and any other cryptic subfamilies that may lie within the ArfGAP and GEF rogue sequences.

Although comparative genomics identified the distribution of each subfamily and revealed the stepwise acquisition of most GAPs and GEFs, the point of origin of some subfamilies was more ambiguous. The restricted distribution at the 5-orders criterion for ADAP, AGAP, and BRAG would suggest that they were not present in the LECA; however, their sparse, but broad distribution at the 2-orders criterion would suggest that they are ancient. If these subfamilies were present in the LECA, then their sparse distribution is the result of frequent secondary loss (Figure 3-5, 3-6, 3-26, 3-27). Why would these subfamilies be lost? These subfamilies share at least partially overlapping functions with other subfamilies: AGAP regulates trafficking between the TGN and endosomes (Nie and Randazzo, 2006), which is also partially carried out by SMAP (Natsume et al., 2006); ADAP is involved in cytoskeletal regulation (Venkateswarlu et al., 2004), as is ACAP; BRAGs are involved in endocytosis, a function also carried out by cytohesin (Claing et al., 2001; Dunphy et al., 2006). Even if they are ancient, these patchy subfamilies evolved more recently

than other ArfGAPs/GEF subfamilies and were therefore redundant and less fixed into the mechanistic landscape of cellular function, meaning that loss during various stages of eukaryotic diversification was less detrimental than losing the other more entrenched subfamily.

In addition to their catalytic domains, ArfGAPs and ArfGEFs are both characterized by a broad array of accessory domains. These domains are predicted to regulate the protein's cellular functions by controlling catalytic activity, cellular location, and interaction with other proteins (Jian et al., 2009). The comparative genomic analysis and re-definition of the domain structure of each GAP and GEF subfamily provided a clear path of domain evolution from a simpler ArfGAP/GEF toolkit to the complex set of human ArfGAP and GEF proteins, particularly within the GAPs. Addition of peripheral domains would have increased the ability of GAPs to act as effectors in parallel biological pathways, by increasing the potential for these regulators to receive signals, aiding in the integration of cellular systems as eukaryotes explored novel ecological niches or evolved into more complex forms (*i.e.,* multicellular). In stark contrast to the GAPs, the domain composition of the human GEF proteins is identical to what is expected to have been present in the ancestor of those sequences; that is, the domain complement of the ArfGEFs has not changed since their origin, indicating that addition of accessory domains is not a driver of complexity for these proteins, although individual cases of domain acquisition can be observed by some paralogues in some lineages. The drastic change in ArfGAP domain composition may suggest that perhaps the GAPs have evolved towards signal integration and crosstalk with other signalling pathways. By

contrast, the consistency in domain composition of the GEFs suggests that their integration with other signalling pathways occurred early in ArfGEF evolution. Simpler complements of accessory domains in the LECA have also been observed for other GTPase regulatory families. Analysis of the RasGAP family identified unexpected complexity in the LECA, but with a restricted complement of domains limited to the RasGAP_C, CH, and C2 domains (van Dam et al., 2011), indicating that the acquisition of additional domains as a mechanism of increasing signalling complexity is not limited to the ArfGAPs.

Despite the finding that the total domain complement present in the Arf regulators was smaller in the LECA than seen in the human complement, it consisted of several functionally distinctive modules including the catalytic domains (ArfGAP and Sec7), ANK repeats, BAR, PH, and C2 domains for the ArfGAPs (Figure 3-9, 3-11), and PH domains, DCB, HUS, HDS1, HDS2, and HDS3 domains for the ArfGEFs (Figure 3-28). These domains may provide insight into the functionality of the ancient ArfGAP and GEF proteins. ANK repeats function in protein-protein interactions (Bork, 1993), thus the acquisition of ANK repeats would have greatly increased the number of binding partners for ArfGAPs, providing potential for novel mechanisms of regulation or localization. PH domains are more commonly involved in membrane association by binding specific phosphoinositides. This is certainly the case for some ArfGEFs; targeting of cytohesin to the plasma membrane is dependent on the presence of its PH domain (Macia et al., 2001). The same has been shown for the PH domain of EFA6 (Franco et al., 1999). PH domains are also important for the ArfGAPs, in some cases reducing GAP activity when removed (Kam et al., 2000). BAR

domains have been shown to play a role in sensing and producing membrane curvature (Field et al., 2011, *inter alia*; Masuda and Mochizuki, 2010, *inter alia*), and in some ArfGAPs suggests a role in sensing or contributing to membrane curvature during vesicle formation, similar to the ALPS motifs of *H. sapiens* ArfGAP1 (Bigay et al., 2005; Mesmin et al., 2007). C2 domains are Ca^{2+} dependent lipid binding domains, suggesting the interaction of GAPs with specific lipid subdomains (Davletov and Sudhof, 1993). The PH domain of cytohesin aside, the only ArfGEF accessory domains found in the LECA are the DCB, HUS and HDS domains found in BIGs and GBFs, and have been shown to interact with a variety of partners including p115, Exo70, myosinIXb (Gillingham and Munro, 2007b; Mouratou et al., 2005; Wright et al., 2014), and have recently been shown to act as regulatory regions, controlling Sec7 GEF activity (McDonold and Fromme, 2014).

3.5.1 Evolution of ArfGAPs and ArfGEFs in vertebrates mimics the evolution of Arfs

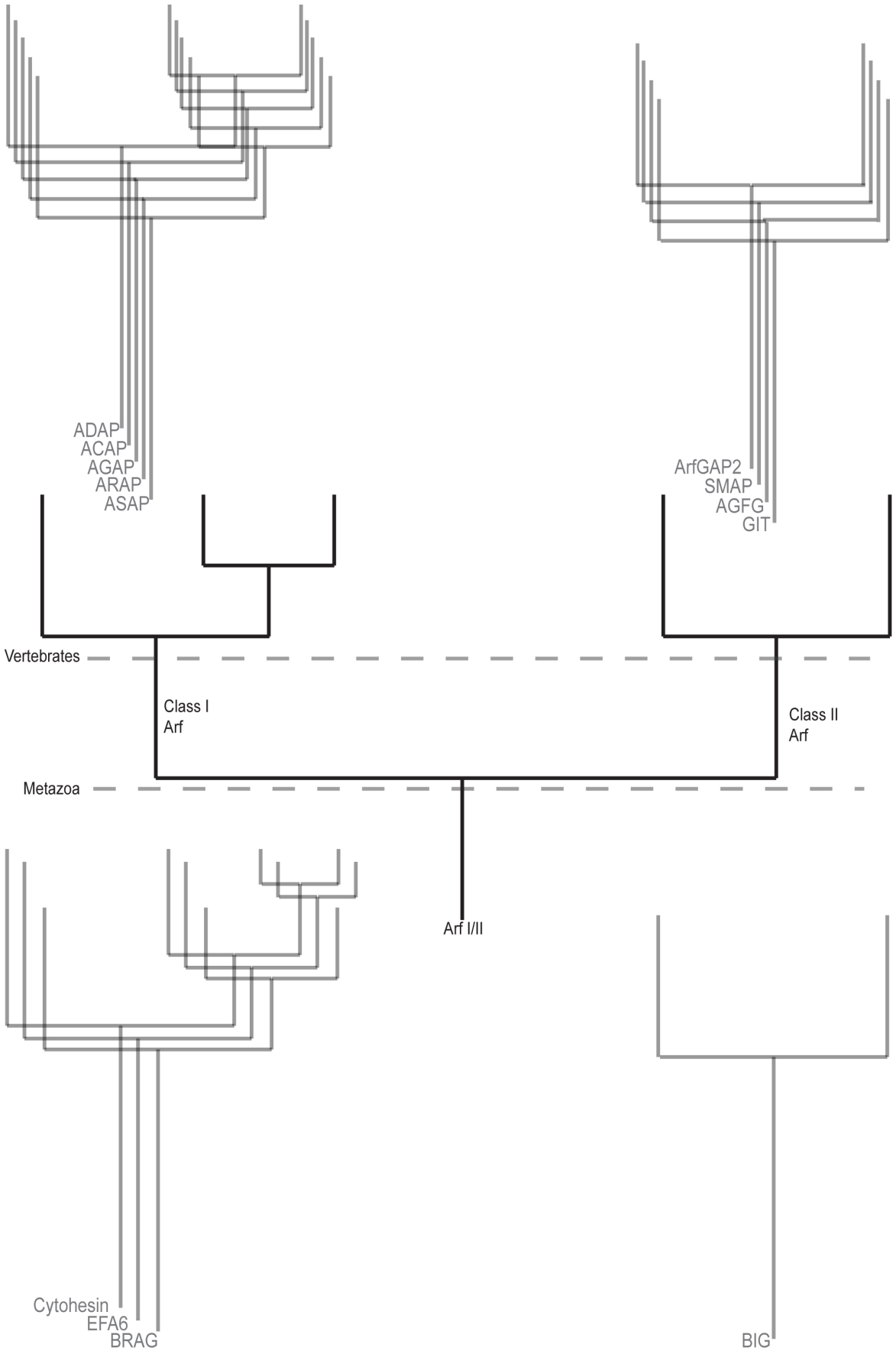
The patterns of ArfGAP and ArfGEF duplication, observed in vertebrates are quite similar to the pattern of duplications seen for the Arfs. Manolea et al., (2010) demonstrated that the ancestral opisthokont possessed two Arf proteins, progenitors of the class I/II Arfs and a class III Arf (Manolea et al., 2010). The class I/II progenitor duplicated near the ancestor of Metazoa and choanoflagellates to produce distinct class I and II Arfs. Near the base of vertebrates, these Arfs duplicated again; the class I Arf duplicated twice, producing Arfs 1-3 and while the class II Arf underwent a single duplication, producing Arfs 4 and 5. Although the pattern of ArfGAP subfamilies is more complex, the overall patterns are strikingly

similar, with all subfamilies, except for ArfGAP1, undergoing one or two gene duplications in the vertebrate ancestor (Figure 3-37). Although the correlation is less strong, a similar pattern is observed for the ArfGEFs: BRAG has undergone two duplications at the base of vertebrates, EFA6 and cytohesin have undergone three gene duplications, and BIG has duplicated only once in vertebrates (Figure 3-37). While, in the case of the GEFs, this correlation does not align perfectly with experimentally established substrate preferences, it should be remembered that the majority of Arf-ArfGEF interactions are analyzed from the perspective of their interactions with Arf1 and Arf6. The exception would be the interaction of GBF with Arf4 and Arf5 at the ERGIC; however, this may be a special case, as GBF has also been shown to localize to the TGN through interactions with Arf1 (Wright et al., 2014).

It is curious to note that, while some ArfGAP subfamilies (ACAP and ARAP) possess additional *D. rerio* sequences that appeared to branch basal to the other vertebrate sequences, no ArfGEF subfamilies displayed such a pattern. However, cytohesin and EFA6 both have four paralogues in vertebrates, unlike the two or three paralogues found in most other GAP and GEF subfamilies. These patterns correlate with the identification of additional class II Arfs in *D. rerio* (P. Melançon, unpublished). Although these subfamilies typically act on other Arf isoforms, it is an interesting correlation to pursue *in vivo*.

The above statements are not meant to suggest that functional relationships have been identified, but rather, should be considered as hypotheses for further dissection of the interactions of this regulatory system. ArfGAPs and ArfGEFs may

Figure 3-37. The evolution of vertebrate ArfGAPs and ArfGEFs mimics the evolution of class I and class II Arfs. Arfs (black line): the pre-duplicate of the class I and class II Arfs duplicated once near the base of vertebrates, producing the class I and class II Arfs (also see Figure 3-1). Each of these lineages duplicated near the base of vertebrates, the class I Arf duplicating twice, producing Arfs 1, 2, and 3, whereas the class II Arf duplicated once, producing Arf 4 and 5. Upper: Some ArfGAPs display the same evolutionary patterns as the class I and class II Arfs. ADAP, ACAP, AGAP, ARAP, and ASAP have undergone two gene duplications near the base of vertebrates, mimicking the class I Arfs. ArfGAP2, SMAP, AGFG, and GIT have undergone two gene duplications near the base of vertebrates, mimicking the class II Arfs. Lower: Some ArfGEFs display similar evolutionary patterns to the class I and class II Arfs. Cytohesin, EFA6, and BRAG have each undergone at least two gene duplications, similar to class I Arfs. BIG is the only ArfGEF to have undergone a single gene duplication. ArfGAP1 and GBF are not shown, as neither has undergone any gene duplications in Metazoa.

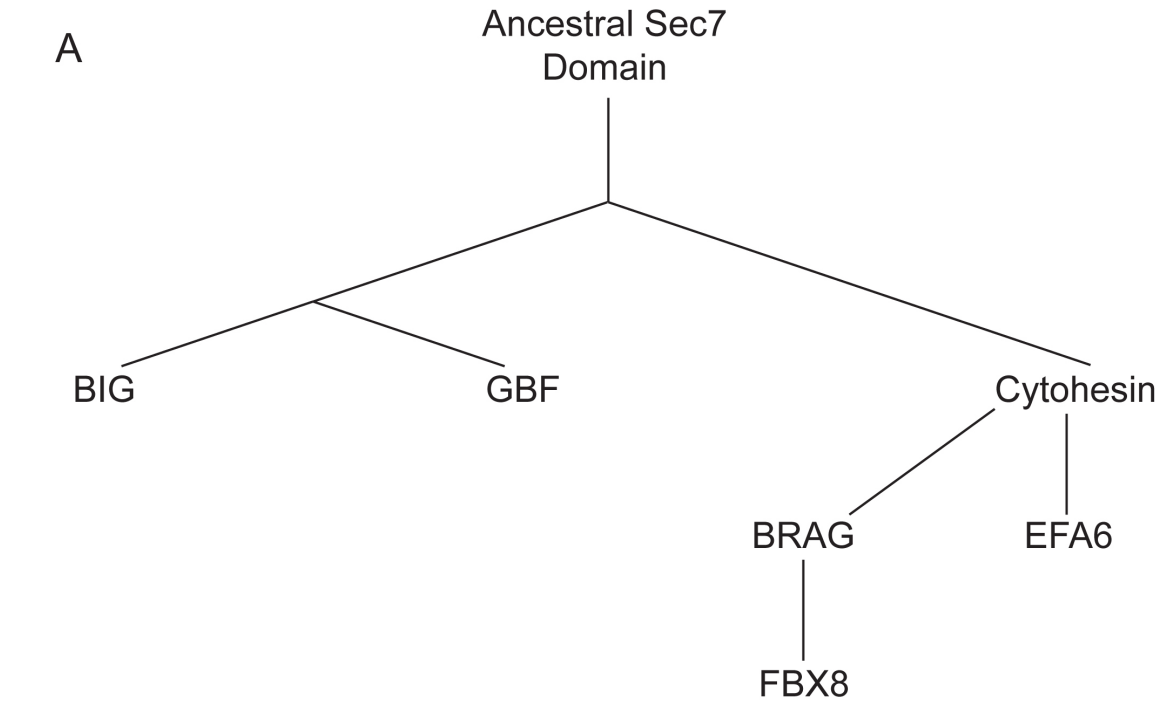

ADAP
ACAP
AGAP
ARAP
ASAP
ArfGAP2
SMAP
AGFG
GIT
Vertebrates
Class I
Arf
Class II
Arf
Metazoa
Arf I/II
Cytohesin
EFA6
BRAG
BIG

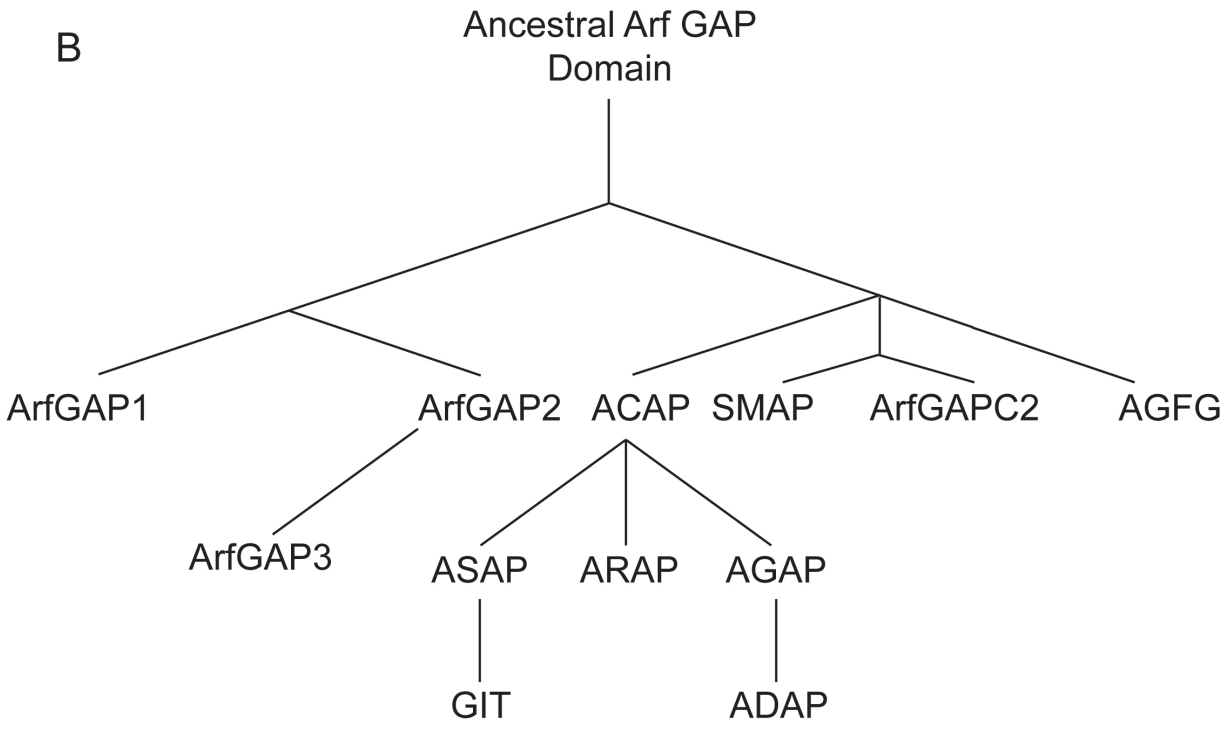
also act on other proteins, the most obvious would be the Arls; this has certainly been suggested to be the case for the *S. cerevisiae* GEF Syt1p, which is able to mediate GDP-GTP exchange on Arl1p *in vivo* and Gcs1p (*S. cerevisiae* ArfGAP1) is able to mediate GAP activity on Arl1p in the *in vitro* (Chen et al., 2010; Liu et al., 2005b). However, further analysis will be required to assess the physiological implications of these interactions, and to determine if these represent solitary cases of ArfGAPs and GEFs able to regulate non-Arf GTPases.

3.5.2 Predicting the evolutionary origins of ArfGAP and GEF subfamilies

The phylogenetic analyses presented here were unable to resolve the order in which each GAP and GEF subfamily arose, or which subfamilies gave rise to the lineage-specific forms. The reconstruction of domain structures does shed some light onto the likely possibilities. It has been previously suggested that BIG and GBF share a common ancestor and are the products of an ancient gene duplication, evidenced by the extensive conservation of the DCB, HUS, and HDS domains (Mouratou et al., 2005). Our finding that cytohesin was also present in the LECA suggests another duplication predating that giving rise to BIG and GBF. The duplication of this ancestral lineage would have given rise to one lineage that then gained the DCB, HUS, and HDS domains and gave rise to BIG and GBF (Figure 3-38). The second lineage would have instead gained a PH domain, giving rise to cytohesin. Given their conserved architectures, it is likely that cytohesin then gave rise to both BRAG and EFA6. Because of the lack of shared domains other than Sec7, the origin of

Figure 3-38. Predicted relationships between ArfGAP and ArfGEF subfamilies. Predicted relationships between subfamilies based on domain composition and BLAST analysis A) All ArfGEFs are derived from an ancestral Sec7 domain-containing protein. BIG and GBF form a clade based on reciprocal best hits and domain conservation. Domain composition and BLAST results suggest that BRAG and EFA6 are derived from cytohesin, with FBX8 arising from a gene duplication of BRAG. B) All ArfGAP proteins are derived from an ancestral ArfGAP domain-containing protein. ArfGAP1 and ArfGAP2 subfamilies preferentially retrieve each other during BLAST experiments suggesting that they may be each other's closest relatives. ACAP, SMAP, and AGFG preferentially retrieve each other over ArfGAP1, ArfGAP2, or ArfGAP3, suggesting that these three may form a separate group. Based on domain analysis, I propose that ASAP and AGAP arose via gene duplication from ACAP, with GIT and ADAP arising by gene duplication of AGAP. ArfGAPC2 retrieves SMAP as its second best hit, hinting that they may share a recent common ancestor. BLAST results indicate that ACAP is ARAP's closest homologue in terms of sequences similarity, suggesting a shared ancestor.


A
Ancestral Sec7 Domain
BIG
GBF
Cytohesin
BRAG
EFA6
FBX8


B
Ancestral Arf GAP Domain
ArfGAP1
ArfGAP2
ACAP
SMAP
ArfGAPC2
AGFG
ArfGAP3
ASAP
ARAP
AGAP
GIT
ADAP

FBX8 is more difficult to predict. However, BLAST analysis of the Sec7 domain from the human FBX8 protein retrieves BRAG as the best hit (after itself), nearly satisfying the 5-orders criterion, suggesting an origin for this later emerging subfamily (Figure 3-38).

Owing to the variability in their domain structures, the ArfGAPs are much harder to predict. However, based on the shared presence of triple ANK repeats C-terminal to the ArfGAP domains, it is worth speculating that GIT was derived from a gene duplication of either ASAP or ACAP. Similarly, it is possible that ARAP is derived from AGAP. In BLAST experiments using the human ArfGAPs to search the human genome, most ArfGAP subfamilies retrieve either ACAP or SMAP. All AGAP, ASAP, and SMAP paralogues retrieve ACAP as the top subfamily after themselves. The ACAP subfamily retrieved ASAP. AGFG retrieved SMAP. ArfGAP1 and ArfGAP2 retrieved each other, while ADAP and ARAP subfamilies retrieved a mix of AGAP, SMAP, ACAP, and ASAP. Thus, it appears that many ArfGAP subfamilies are derived from an ancestral ACAP sequence (Figure 3-38). BLAST of *A. thaliana* ArfGAPC2 sequences against its own protein database identified SMAP as the next best ArfGAP subfamily, consistent with the misclassification of multiple ArfGAPC2 sequences as SMAPs. However, these hypotheses are highly speculative, and require further testing using highly advanced phylogenetic methods, such as an in depth Scrollsaw analysis, to resolve.

3.5.3 Reconstructing the Arf regulatory system in the LECA

The above analyses allow, for the first time, a holistic view of the evolution of an entire GAP – GEF – GTPase system, and its reconstruction in the LECA. While previous analyses have pointed to the presence of a single Arf GTPase in the LECA (Berriman et al., 2005; Li et al., 2004) the data presented here indicates the presence of six ArfGAPs and three ArfGEFs, suggesting a much simpler cellular configuration of this system in the LECA, as compared to conventionally studied systems (Cox et al., 2004; Kahn et al., 2008). The finding that the GAPs greatly outnumber both the GEFs and the GTPase suggests that it is the GAPs that drive the complexity within this system. Multiple GAPs would greatly increase the potential for cross talk and integration of signals between parallel pathways, but cannot be responsible for the specific targeting of Arfs to membranes, as their association with the GTPase requires the GTP-bound form. Although we cannot say with certainty whether the GAPs or the GEFs expanded first, it is tempting to speculate that the GAPs were the first components in this system to expand, followed by the GEFs, and eventually the Arfs themselves. This scenario would indicate a period where early eukaryotes possess multiple GAPs, a single GEF, and a single GTPase. This scenario would invoke the necessity of additional upstream regulatory components in order to recruit the GEF and the Arf to the appropriate membrane. Hints to this type of regulation have already been observed; it has recently been shown that the recruitment and activation of the *S. cerevisiae* Sec7 requires interaction with Arl1, Ypt1, Ypt31, and Arf1 (McDonold and Fromme, 2014). Further analyses are required to confirm the conservation of this regulation in other systems, and to identify

upstream mechanisms of this nature for other ArfGEFs, but could nonetheless provide a basis for the targeting of a single GEF to different membranes in early eukaryotes.

It is accepted that the LECA possessed the full array of vesicular trafficking pathways necessary for transport between the organelles of the endomembrane system most of which are Arf dependent (D'Souza-Schorey and Chavrier, 2006). The LECA is proposed to have possessed a single Arf protein, as is the case for most extant eukaryotes (Li et al., 2004). The presence of a single Arf GTPase would indicate that this ancestral homologue would have been able to act at each organelle, a task that has been divided between up to six Arf proteins in mammals (Volpicelli-Daley et al., 2005). This would imply that Arf specificity is not encoded in the Arfs themselves, but rather by the Arf effectors (GAPs and GEFs). This may still be the case in extant organisms even with larger numbers of GAPs, GEFs, and GTPases. Alternatively, any organelle specificity encoded by Arfs in extant organisms, such as mammals, may be the result of lineage-specific evolution.

Identifying the site of action and the function of ArfGAPC2 *in vivo* will greatly enhance our understanding of the range of functions carried out by these Arf regulators. Similarly, further characterization of ArfGEFs in non-standard model systems will identify both novel functionality and conserved roles for ArfGEFs in membrane traffic. It is worth noting that when mapping the localization of the ancient ArfGAPs and GEFs onto the trafficking system, most major Arf-dependent transport pathway are represented (Figure 3-2) and provides a set of hypotheses for which GAPs and GEFs may interact with which coat complexes as these

functional interactions are not yet completely characterized. Testing these hypotheses *in vivo* will greatly expand our understanding of how these proteins function in the living cell.

Chapter 4: Comparative genomic and phylogenetic analysis reveals ancient complexity of the COPII coat

This chapter has been published as:

Schlacht, A., Dacks, J.B. 2015. Unexpected ancient paralogues and an evolutionary model for the COPII coat complex. *Genome Biology and Evolution* 7: 1098-1109

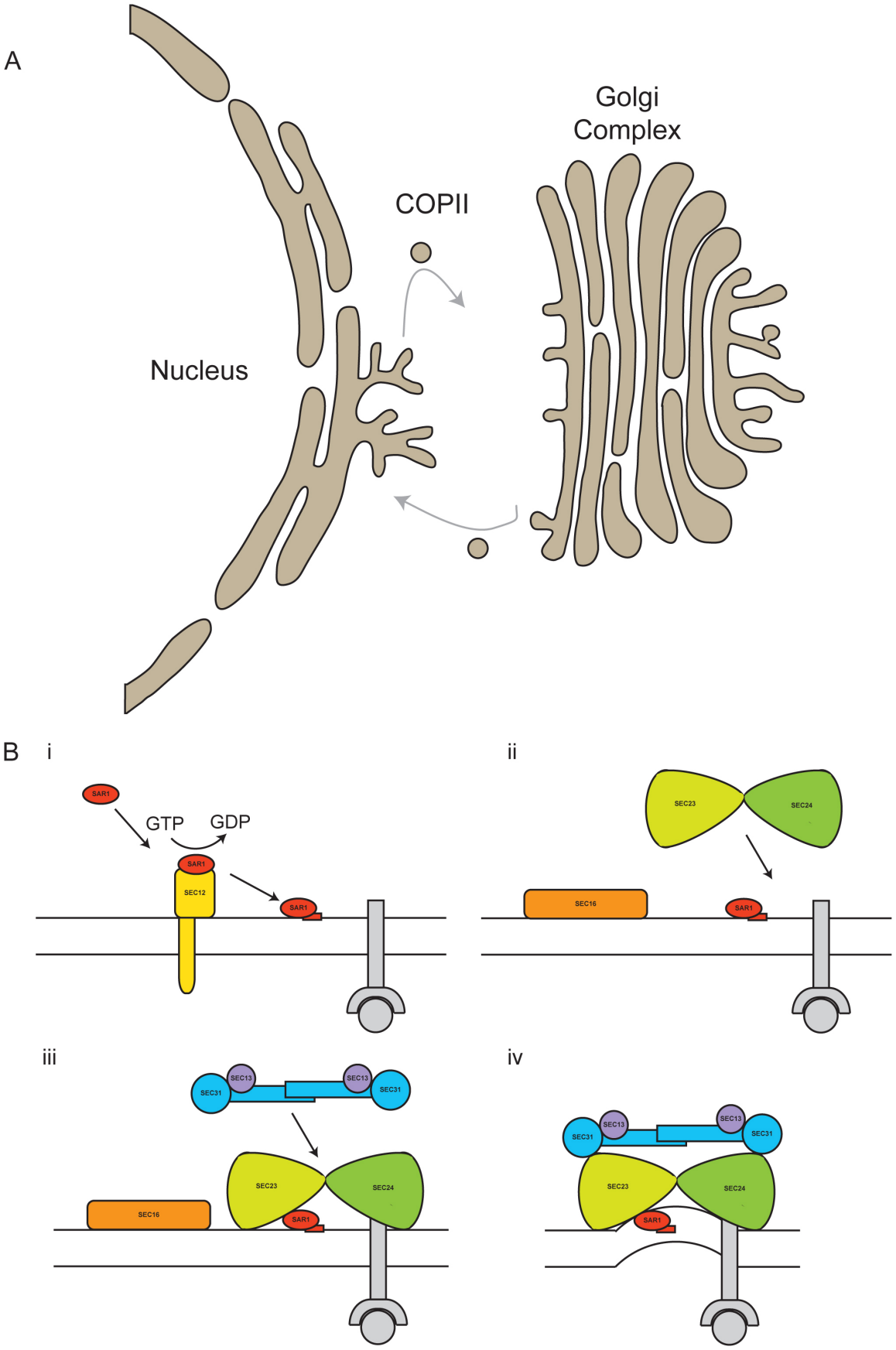
4.1 Overview

The previous chapter examined the evolution of ArfGAPs and ArfGEFs, important regulators of membrane traffic that represent an early step in the formation of transport vesicles. This analysis identified the presence of multiple GAPs and GEFs in the LECA, including a previously unreported ArfGAP subfamily, ArfGAPC2, found in diverse eukaryotes, but that has frequently been lost from some lineages, including opisthokonts. The conservation of multiple ArfGAP and ArfGEF subfamilies is suggestive of a high degree of specificity for the Arf system during coat formation. As we will see in this chapter, the next step in vesicle formation is the recruitment of cytosolic coat components to the donor membrane. While many complexes have been the subject of thorough comparative genomic and phylogenetic analysis (see chapter 5), the COPII complex is one of the only coat complexes not previously analyzed by both comparative genomic and phylogenetic methods. This chapter will describe an evolutionary analysis of the COPII complex, providing insight into the evolution of the COPII coat and its functionality in diverse eukaryotic organisms.

4.2 Assembly of the COPII complex is an important early step in membrane trafficking

The COPII complex is responsible for the exit of proteins and lipids synthesized in ER and their subsequent transport to the Golgi (Figure 4-1A; Barlowe et al., 1994). Seven interacting components are required to form a COPII vesicle. First, an ER localized GEF, Sec12, activates the small GTPase Sar1, by exchanging GDP for GTP (Barlowe and Schekman, 1993; Weissman et al., 2001; Figure 4-1Bi).

Figure 4-1. Overview of COPII coat formation. A) The COPII complex is responsible for anterograde transport from the ER to the Golgi complex. B) Overview of COPII coat formation. (i) The small GTPase Sar1 is bound by its GEF, Sec12, which catalyzes the exchange of GDP for GTP, activating Sar1 which then binds the ER membrane. (ii) Recruitment of Sar1 stimulates recruitment of the Sec23/Sec24 complex, through interaction with Sec16. Sec23 binds Sar1 and Sec24 recognizes cargo. (iii) Recruitment of Sec23/Sec24 stimulates Sec13/Sec31 complex formation. Sec31 binds the Sec23/Sec24 pre-budding complex. Together, Sec23/Sec24 and Sec13/Sec31 deform the ER membrane (iv).


A
Golgi
Complex
COPII
Nucleus
B
i
SAR1
GTP
GDP
SAR1
SEC12
SAR1
ii
SEC23
SEC24
SEC16
SAR1
iii
SEC13
SEC31
SEC13
SEC31
SEC23
SEC24
SEC16
SAR1
iv
SEC13
SEC31
SEC13
SEC31
SEC23
SEC24
SAR1

Activated Sar1 then binds the ER membrane and recruits the heterodimeric Sec23/24 adaptor complex which constitutes the Sar1-GAP and primary cargo binding subunit, respectively (Bi et al., 2002; Lee et al., 2005; Miller et al., 2003; Wendeler et al., 2007; Yoshihisa et al., 1993; Figure 4-1Bii). Recruitment of the Sec23/24 complex results in the recruitment and binding of the heterotetrameric Sec13/31 cage complex, which along with Sec23/24, are responsible for membrane deformation (Bi et al., 2002; Fromme et al., 2007; Stagg et al., 2006; Stagg et al., 2008). Figure 4-1Biii Sec16 is a multifunctional protein, implicated in the negative regulation of Sar1, controlling the timing of GTP hydrolysis (Kung et al., 2012), and acts as a scaffold, aiding the recruitment of the other COPII subunits (Espenshade et al., 1995; Gimeno et al., 1995; Gimeno et al., 1996; Shaywitz et al., 1997). Neither Sec12 nor Sec16 is included into the budding vesicle.

The COPII coat belongs to a larger family of membrane deforming complexes, including other coat complexes (COPI, AP1-5), the IFT, the SEA complex, the HOPS/CORVET tethering complex, and the NPC. As mentioned in chapter 1, these seemingly different gene families are linked through the 'protocoatomer domain architecture', a protein fold composed of a β-propeller followed by an α-solenoid that is common to all of these complexes (Devos et al., 2004; section 1.4.8). These complexes are thought be derived from an ancient membrane-deformation complex, far predating the LECA. Understanding the relationship between these complexes would, in turn, help us to understand how a highly complex membrane trafficking system evolved from an ancestor with no internal membrane compartments.

Although the COPII complex has been studied in many different organisms, including *P. pastoris* (Payne et al., 2000), *T. brucei* (Demmel et al., 2011; Sealey-Cardona et al., 2014), *A. thaliana* (De Craene et al., 2014), *H. sapiens* (Iinuma et al., 2007), and *S. cerevisiae* (Barlowe et al., 1994; Gimeno et al., 1996; Kung et al., 2012), relatively little is known about this complex outside of the latter two. In an effort to expand the understanding of nuclear pore complexes and related coat forming complexes in a variety of protistan lineages, a previous comparative genomic analysis identified components of the COPII complex in a set of diverse eukaryotic taxa (Neumann et al., 2010). Their analysis revealed that Sar1, Sec13, Sec31, Sec23, Sec24, and Sec16 are found in nearly all eukaryotes, and were therefore likely present in the LECA. This is consistent with previous large-scale analyses of the eukaryotic endomembrane machinery that found both Sar1 and Sec31 are highly conserved markers of the COPII coat across broad eukaryotic lineages (Dacks and Field, 2004). The analysis presented here extends these findings by deriving an evolutionary model describing the progression of the COPII complex from an early representative of the eukaryotic lineage to the LECA. This was done by examining two additional proteins (Sec12 and Sed4), including recently sequenced key taxa, and using in depth phylogenetic analysis to assess the evolution of each COPII component. This analysis also uncovered previously undescribed ancient paralogues of some COPII components and permitted the reconstruction of the COPII complex that was likely present in the LECA, gleaning insight into the evolution of this critical trafficking complex.

4.3 Abbreviated materials and methods

Comparative genomic analyses were carried out as described in section 2.2, using 74 genomes spanning the diversity of eukaryotes (Figure 4-2). Phylogenetic analysis was carried out as in section 2.3. The details of each phylogenetic analysis, including the number of taxa, length of masked alignment, and model parameters for each method is in Table 4-1. Phylogenetic analyses were carried out using the CIPRES web server (Miller et al., 2010). Tertiary structure predictions were carried out as described in section 2.4

Table 4-1. Parameters of phylogenetic analysis, corresponding dataset, and figure number for each COPII component.

Figure	Dataset Name	Number of taxa	Length of alignment (a.a.)	Evolutionary model used	
				Phylobayes	RAxML
4-4	Sar1_R1	120	190	LG+CAT+G	LG+CAT
4-5	Sec12_R1	51	327	LG+CAT+G+F	LG+CAT+F
4-6	Sec13_R4	87	279	LG+CAT+G	LG+CAT
4-7	Sec16_R1	72	330	LG+CAT+G+F	LG+CAT+F
4-8	Sec23_R1	123	714	LG+CAT+G+F	LG+CAT+F
4-9	Sec24_R1	223	656	LG+CAT+G+F	LG+CAT+F
4-10	Sec31_R1	90	767	LG+CAT+G+F	LG+CAT+F
4-11	Plant_Sar1_R3	108	192	LG+CAT+G+F	LG+CAT+F
4-12	Plant_Sec12	43	336	LG+CAT+G+F	LG+CAT+F
4-13	Plant_Sec13	64	299	LG+CAT+G+F	LG+CAT+F
4-14	Plant_Sec16	46	1014	LG+CAT+G+F	LG+CAT+F
4-15	Plant_Sec23_R2	129	680	LG+CAT+G+F	LG+CAT+F

4-16	Plant_Sec24	128	687	LG+CAT+G+F	LG+CAT+F
4-17	Plant_Sec31	41	940	LG+CAT+G+F	LG+CAT+F
4-18	Sec23_R4	217	671	LG+CAT+G+F	LG+CAT+F
4-19A	Sed4_R3	31	387	LG+CAT+G+F	LG+CAT+F
4-20	Opisthokonta_R3	128	583	LG+CAT+G+F	LG+CAT+F
4-21	Amoebozoa_R1	18	583	LG+CAT+G+F	LG+CAT+F
4-22	Excavata_R5	18	583	LG+CAT+G+F	LG+CAT+F
4-23	Archaeplastida_R1	56	583	LG+CAT+G+F	LG+CAT+F
4-24	SAR_R3	56	583	LG+CAT+G+F	LG+CAT+F
4-25	Ubertree_R2	30	583	LG+CAT+G+F	LG+CAT+F
4-26	Ubertree_R3	20	583	LG+CAT+G+F	LG+CAT+F
4-27	Sec24III	87	581	LG+CAT+G+F	LG+CAT+F
4-28	Ubertree_R13	40	583	LG+CAT+G+F	LG+CAT+F
4-29	Ubertree_R14	26	583	LG+CAT+G+F	LG+CAT+F
4-30	Ubertree_R15	20	583	LG+CAT+G+F	LG+CAT+F

4.4 Results

4.4.1 The COPII coat complex has sparsely and ubiquitously distributed components

Comparative genomic analysis was used to assess the distribution of each component of the COPII coat. BLAST and HMMer were used to identify homologues of each component of the coat in a broad, representative distribution of eukaryotic genomes (Figure 4-2). Consistent with the results from Neumann et al., (2010) at least one copy of Sar1, Sec23, Sec24, Sec13, and Sec31 was identified in every eukaryotic genome analyzed (Figure 4-3), providing strong evidence that these subunits were present in the LECA. The pervasiveness of these proteins in diverse eukaryotic taxa highlights the key role that Sar1, Sec23, Sec24, Sec13, and Sec31

Figure 4-2. Summary of all taxa sampled by analyses presented in this chapter and their relative phylogenetic positions. Taxa searched during comparative genomic analysis are coloured by supergroup. The secondary set of taxa in Archaeplastida (dark green) was only used to examine expansions of the COPII machinery in plants. The secondary set of Fungi (dark blue) was only used to analyze the evolution of Sed4. All other taxa were used to search for all other COPII components.

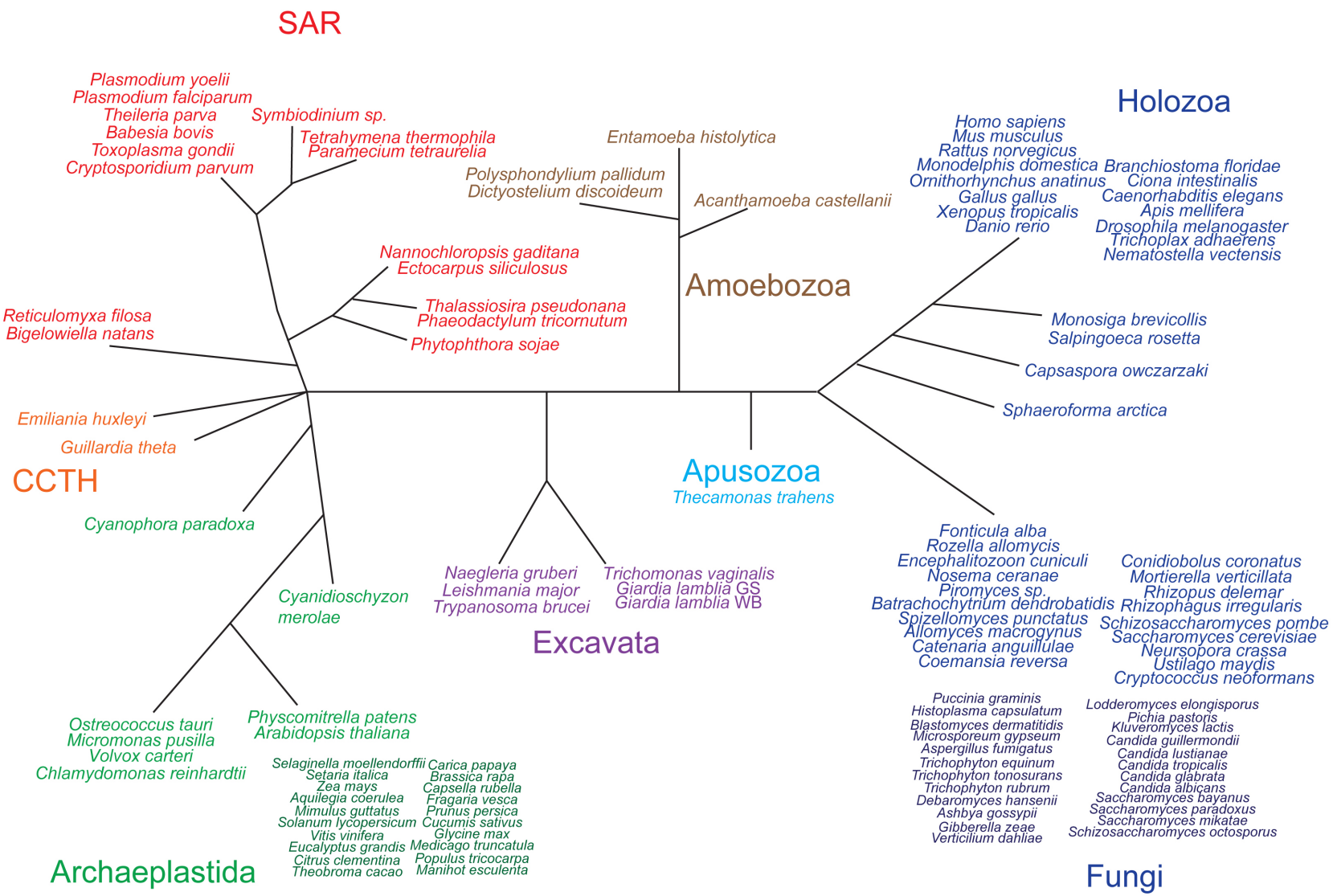
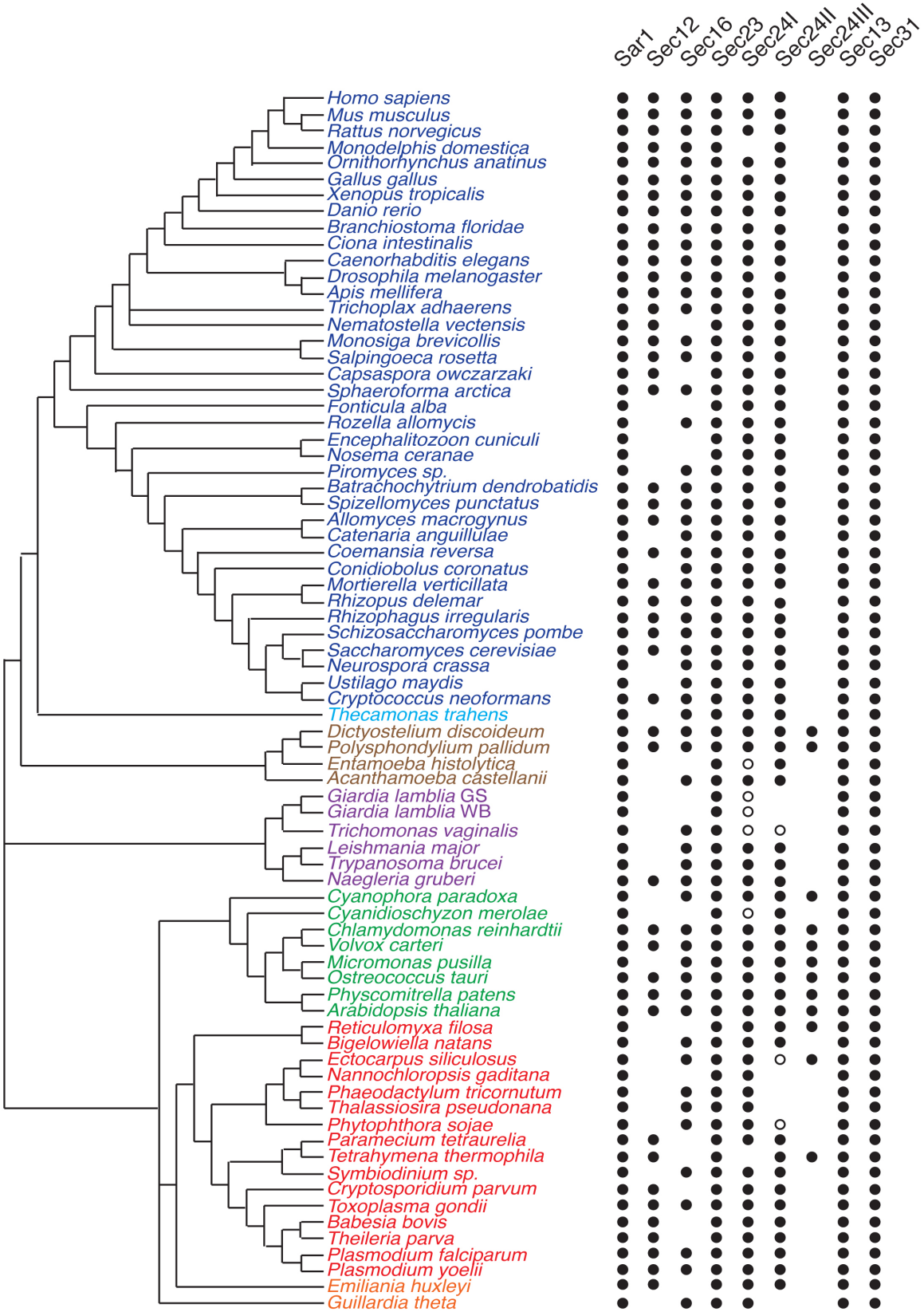

SAR
Plasmodium yoelii
Plasmodium falciparum
Theileria parva
Babesia bovis
Toxoplasma gondii
Cryptosporidium parvum
Symbiodinium sp.
Tetrahymena thermophila
Paramecium tetraurelia
Nannochloropsis gaditana
Ectocarpus siliculosus
Thalassiosira pseudonana
Phaeodactylum tricornutum
Phytophthora sojae
Reticulomyxa filosa
Bigelowiella natans
Emiliania huxleyi
Guillardia theta
CCTH
Cyanophora paradoxa
Cyanidioschyzon merolae
Ostreococcus tauri
Micromonas pusilla
Volvox carteri
Chlamydomonas reinhardtii
Physcomitrella patens
Arabidopsis thaliana
Selaginella moellendorffii
Setaria italica
Zea mays
Aquilegia coerulea
Mimulus guttatus
Solanum lycopersicum
Vitis vinifera
Eucalyptus grandis
Citrus clementina
Theobroma cacao
Carica papaya
Brassica rapa
Capsella rubella
Fragaria vesca
Prunus persica
Cucumis sativus
Glycine max
Medicago truncatula
Populus tricocarpa
Manihot esculenta
Archaeplastida
Entamoeba histolytica
Polysphondylium pallidum
Dictyostelium discoideum
Acanthamoeba castellanii
Amoebozoa
Naegleria gruberi
Leishmania major
Trypanosoma brucei
Trichomonas vaginalis
Giardia lamblia GS
Giardia lamblia WB
Excavata
Apusozoa
Thecamonas trahens
Holozoa
Homo sapiens
Mus musculus
Rattus norvegicus
Monodelphis domestica
Ornithorhynchus anatinus
Gallus gallus
Xenopus tropicalis
Danio rerio
Branchiostoma floridae
Ciona intestinalis
Caenorhabditis elegans
Apis mellifera
Drosophila melanogaster
Trichoplax adhaerens
Nematostella vectensis
Monosiga brevicollis
Salpingoeca rosetta
Capsaspora owczarzaki
Sphaeroforma arctica
Fonticula alba
Rozella allomycis
Encephalitozoon cuniculi
Nosema ceranae
Piromyces sp.
Batrachochytrium dehdrobatidis
Spizellomyces punctatus
Allomyces macrogynus
Catenaria anguillulae
Coemansia reversa
Conidiobolus coronatus
Mortierella verticillata
Rhizopus delemar
Rhizophagus irregularis
Schizosaccharomyces pombe
Saccharomyces cerevisiae
Neursopora crassa
Ustilago maydis
Cryptococcus neoformans
Puccinia graminis
Histoplasma capsulatum
Blastomyces dermatitidis
Microsporeum gypseum
Aspergillus fumigatus
Trichophyton equinum
Trichophyton tonosurans
Trichophyton rubrum
Debaromyces hansenii
Ashbya gossypii
Gibberella zeae
Verticilium dahliae
Lodderomyces elongisporus
Pichia pastoris
Kluveromyces lactis
Candida guillermondii
Candida lustianae
Candida tropicalis
Candida glabrata
Candida albicans
Saccharomyces bayanus
Saccharomyces paradoxus
Saccharomyces mikatae
Schizosaccharomyces octosporus
Fungi

Figure 4-3. Comparative genomic analysis reveals presence of COPII coat components across the diversity of eukaryotes. At least one orthologue of each Sar1, Sec23, Sec24, Sec13, and Sec31 has been identified in all taxa sampled, whereas Sec12 and Sec16 are missing from multiple eukaryotic taxa, but are found in every eukaryotic supergroup. These distributions indicate that all seven components were present in the LECA. Black dots indicate the presence of at least one orthologue of the indicated protein (column) in the corresponding organism (row), open dots represent additional Sec24 sequences that did not fall into any clade during phylogenetic analysis and are classified based on best BLAST hit. Empty space indicates that no orthologue was identified. Orthologous sequences were identified using BLAST and HMMer.


Sar1
Sec12
Sec16
Sec23
Sec24I
Sec24II
Sec24III
Sec13
Sec31
Homo sapiens
Mus musculus
Rattus norvegicus
Monodelphis domestica
Ornithorhynchus anatinus
Gallus gallus
Xenopus tropicalis
Danio rerio
Branchiostoma floridae
Ciona intestinalis
Caenorhabditis elegans
Drosophila melanogaster
Apis mellifera
Trichoplax adhaerens
Nematostella vectensis
Monosiga brevicollis
Salpingoeca rosetta
Capsaspora owczarzaki
Sphaeroforma arctica
Fonticula alba
Rozella allomycis
Encephalitozoon cuniculi
Nosema ceranae
Piromyces sp.
Batrachochytrium dendrobatidis
Spizellomyces punctatus
Allomyces macrogynus
Catenaria anguillulae
Coemansia reversa
Conidiobolus coronatus
Mortierella verticillata
Rhizopus delemar
Rhizophagus irregularis
Schizosaccharomyces pombe
Saccharomyces cerevisiae
Neurospora crassa
Ustilago maydis
Cryptococcus neoformans
Thecamonas trahens
Dictyostelium discoideum
Polysphondylium pallidum
Entamoeba histolytica
Acanthamoeba castellanii
Giardia lamblia GS
Giardia lamblia WB
Trichomonas vaginalis
Leishmania major
Trypanosoma brucei
Naegleria gruberi
Cyanophora paradoxa
Cyanidioschyzon merolae
Chlamydomonas reinhardtii
Volvox carteri
Micromonas pusilla
Ostreococcus tauri
Physcomitrella patens
Arabidopsis thaliana
Reticulomyxa filosa
Bigelowiella natans
Ectocarpus siliculosus
Nannochloropsis gaditana
Phaeodactylum tricornutum
Thalassiosira pseudonana
Phytophthora sojae
Paramecium tetraurelia
Tetrahymena thermophila
Symbiodinium sp.
Cryptosporidium parvum
Toxoplasma gondii
Babesia bovis
Theileria parva
Plasmodium falciparum
Plasmodium yoelii
Emiliania huxleyi
Guillardia theta

play in forming the COPII coat as seen in *in vitro* analyses, which have identified these five components as necessary and sufficient to bud vesicles from synthetic liposomes (Barlowe et al., 1994; Salama et al., 1993).

In contrast to the above machinery, Sec12 and Sec16 were unidentifiable in multiple taxa. Absences were not limited to one particular group of organisms, but were distributed across the six supergroups (Figure 4-3). Sec12 was missing more frequently than Sec16, and in only eight instances are they both missing from the same organism. Of these, four (*i.e., Encephalitozoon cuniculi, Nosema ceranae, E. histolytica,* and *G. lamblia*) are parasites and are known for high levels of sequence divergence or cellular reduction, possibly accounting for our inability to detect these proteins. *Cyanidioschyzon merolae* is an extremophile with a minimal membrane trafficking system (Matsuzaki et al., 2004) and *Nannochloropsis gaditana* is a Eustigmatophycean microalga with a reduced cellular system (Lubián, 1982; Radakovits et al., 2012). These reduced cellular configurations likely resulted in a stripping down and loss of non-essential cellular machinery. By contrast, *Fonticula alba* and *Reticulomyxa filosa* are both free-living heterotrophs; sequence divergence is the most likely explanation for the absence from these two organisms, and may also be the case for the other putative absences. In Sec16, the central conserved domain is the only region that is strongly conserved between taxa; therefore, sequence divergence in the flanking regions drastically increases the likelihood of false-negatives. This is also likely the case for Sec12; low sequence conservation and the presence of multiple WD40 repeats makes it difficult to distinguish from other WD40 repeat containing proteins. This was especially apparent when trying to

identify the *S. cerevisiae* Sec12 using the *H. sapiens* sequence; multiple rounds of psi-BLAST were required to show that they are indeed homologues, as BLASTp did not provide enough sensitivity to do so.

4.4.2 Lineage-specific expansions and an ancient Sec23 duplication

Phylogenetic analyses were carried out to determine the number of paralogues of each COPII component present in the LECA, and to find expansions and reductions in various eukaryotic lineages. Sequences obtained from the comparative genomic analyses were used to construct phylogenetic datasets. Our analyses indicated that for six of the seven components analyzed, only one paralogue was present in the LECA (Figures 4-4 – 4-10). These trees are characterized by one clade per supergroup, along with weak backbone support. Five subunits (Sar1, Sec23, Sec24, Sec31, and Sec16) have undergone expansions in vertebrates, correlating with increasing organismal complexity, possibly the result of selection for additional paralogues permitting tissue specificity or differential regulation in these organisms. All seven subunits have undergone expansions in *A. thaliana* and *P. patens*. To determine whether these additional paralogues are lineage-specific, or if they represent gene duplications that occurred earlier in the plant lineage, additional archaeplastid genomes were sampled and individual protein trees were re-run using only the archaeplastid taxa (Figures 4-11 – 4-17). Similar to other protein families, clear expansions have occurred in individual species and in higher archaeplastid orders, roughly correlating with the evolution of multicellularity in plants (Rutherford and Moore, 2002; Sanderfoot, 2007).

Figure 4-4. Phylogenetic analysis of Sar1 identifies a single paralogue in the LECA. Phylogenetic analysis was carried out to determine the number of Sar1 paralogues present in the LECA and to identify lineage-specific expansions. Although many lineage-specific expansions have occurred, such as in multicellular plants and vertebrates (upper and lower grey boxes, respectively), the majority of taxa possess a single Sar1 paralogue. Additionally, only one clade of each supergroup is visible, pointing to a single Sar1 protein in the LECA. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (Phylobayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95, closed light circles ≥ 0.95/75, open circles ≥ 0.8/50.

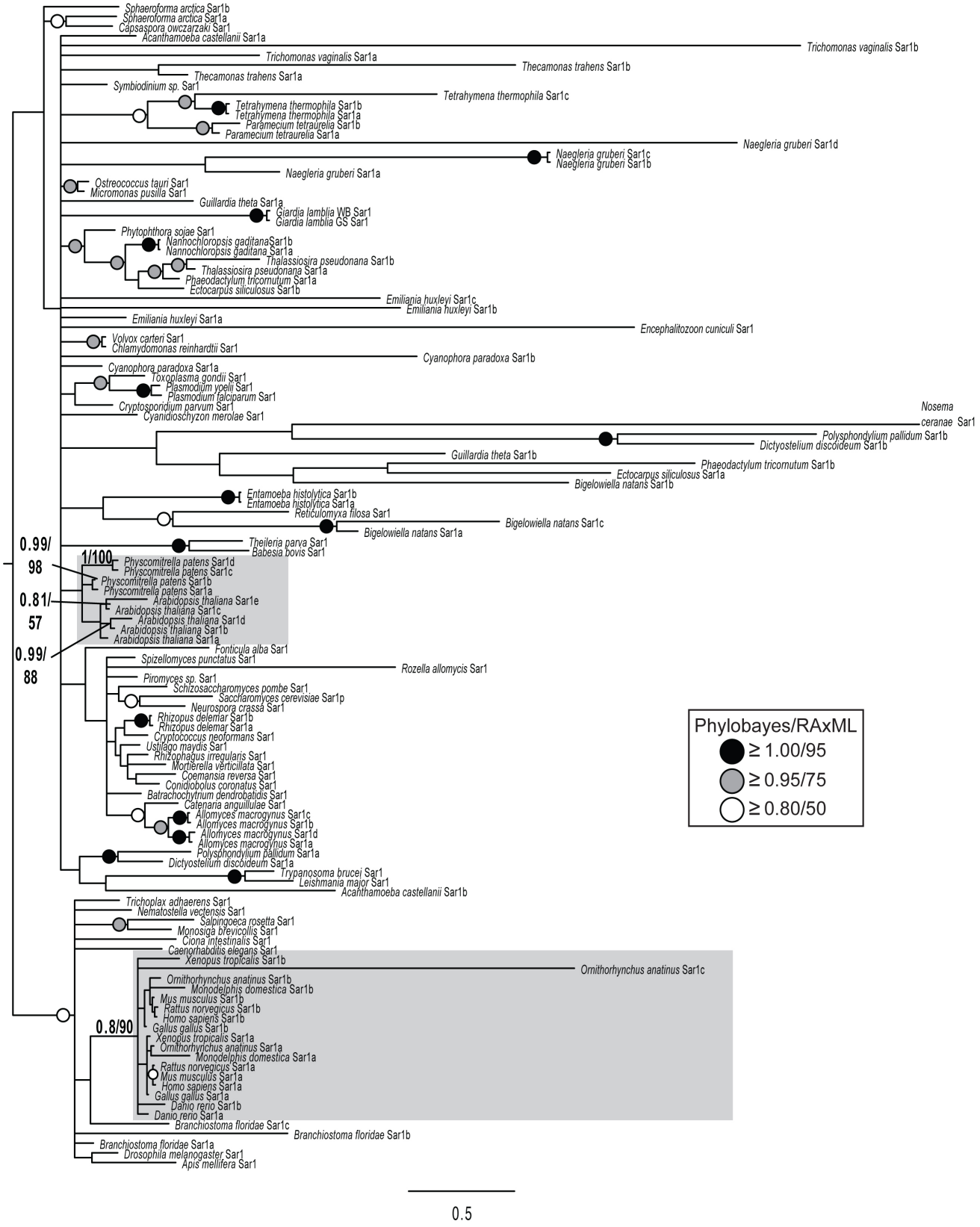

Sphaeroforma arctica Sar1b
Sphaeroforma arctica Sar1a
Capsaspora owczarzaki Sar1
Acanthamoeba castellanii Sar1a
Trichomonas vaginalis Sar1b
Trichomonas vaginalis Sar1a
Thecamonas trahens Sar1b
Thecamonas trahens Sar1a
Symbiodinium sp. Sar1
Tetrahymena thermophila Sar1c
Tetrahymena thermophila Sar1b
Tetrahymena thermophila Sar1a
Paramecium tetraurelia Sar1b
Paramecium tetraurelia Sar1a
Naegleria gruberi Sar1d
Naegleria gruberi Sar1c
Naegleria gruberi Sar1b
Naegleria gruberi Sar1a
Ostreococcus tauri Sar1
Micromonas pusilla Sar1
Guillardia theta Sar1a
Giardia lamblia WB Sar1
Giardia lamblia GS Sar1
Phytophthora sojae Sar1
Nannochloropsis gaditana Sar1b
Nannochloropsis gaditana Sar1a
Thalassiosira pseudonana Sar1b
Thalassiosira pseudonana Sar1a
Phaeodactylum tricornutum Sar1a
Ectocarpus siliculosus Sar1b
Emiliania huxleyi Sar1c
Emiliania huxleyi Sar1b
Emiliania huxleyi Sar1a
Encephalitozoon cuniculi Sar1
Volvox carteri Sar1
Chlamydomonas reinhardtii Sar1
Cyanophora paradoxa Sar1b
Cyanophora paradoxa Sar1a
Toxoplasma gondii Sar1
Plasmodium yoelii Sar1
Plasmodium falciparum Sar1
Cryptosporidium parvum Sar1
Cyanidioschyzon merolae Sar1
Nosema ceranae Sar1
Polysphondylium pallidum Sar1b
Dictyostelium discoideum Sar1b
Guillardia theta Sar1b
Phaeodactylum tricornutum Sar1b
Ectocarpus siliculosus Sar1a
Bigelowiella natans Sar1b
Entamoeba histolytica Sar1b
Entamoeba histolytica Sar1a
Reticulomyxa filosa Sar1
Bigelowiella natans Sar1c
Bigelowiella natans Sar1a
Theileria parva Sar1
Babesia bovis Sar1
0.99/98
1/100
Physcomitrella patens Sar1d
Physcomitrella patens Sar1c
Physcomitrella patens Sar1b
Physcomitrella patens Sar1a
0.81/57
Arabidopsis thaliana Sar1e
Arabidopsis thaliana Sar1c
Arabidopsis thaliana Sar1d
Arabidopsis thaliana Sar1b
Arabidopsis thaliana Sar1a
0.99/88
Fonticula alba Sar1
Spizellomyces punctatus Sar1
Rozella allomycis Sar1
Piromyces sp. Sar1
Schizosaccharomyces pombe Sar1
Saccharomyces cerevisiae Sar1p
Neurospora crassa Sar1
Rhizopus delemar Sar1b
Rhizopus delemar Sar1a
Cryptococcus neoformans Sar1
Ustilago maydis Sar1
Rhizophagus irregularis Sar1
Mortierella verticillata Sar1
Coemansia reversa Sar1
Conidiobolus coronatus Sar1
Batrachochytrium dendrobatidis Sar1
Catenaria anguillulae Sar1
Allomyces macrogynus Sar1c
Allomyces macrogynus Sar1b
Allomyces macrogynus Sar1d
Allomyces macrogynus Sar1a
Polysphondylium pallidum Sar1a
Dictyostelium discoideum Sar1a
Trypanosoma brucei Sar1
Leishmania major Sar1
Acanthamoeba castellanii Sar1b
Trichoplax adhaerens Sar1
Nematostella vectensis Sar1
Salpingoeca rosetta Sar1
Monosiga brevicollis Sar1
Ciona intestinalis Sar1
Caenorhabditis elegans Sar1
Xenopus tropicalis Sar1b
Ornithorhynchus anatinus Sar1c
Ornithorhynchus anatinus Sar1b
Monodelphis domestica Sar1b
Mus musculus Sar1b
Rattus norvegicus Sar1b
Homo sapiens Sar1b
Gallus gallus Sar1b
0.8/90
Xenopus tropicalis Sar1a
Ornithorhynchus anatinus Sar1a
Monodelphis domestica Sar1a
Rattus norvegicus Sar1a
Mus musculus Sar1a
Homo sapiens Sar1a
Gallus gallus Sar1a
Dahio rerio Sar1b
Danio rerio Sar1a
Branchiostoma floridae Sar1c
Branchiostoma floridae Sar1b
Branchiostoma floridae Sar1a
Drosophila melanogaster Sar1
Apis mellifera Sar1
Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50
0.5

Figure 4-5. Phylogenetic analysis of Sec12 sequences identifies a single paralogue in the LECA. Phylogenetic analysis was carried out to determine the number of Sec12 paralogues present in the LECA. By and large, a single clade for each supergroup was reconstructed. Additionally, the majority of taxa possess only a single Sec12 sequence indicating that only one Sec12 sequence was present in the LECA. Lineage-specific expansions have also occurred in multicellular plants and *A. macrogynus* (upper and lower grey boxes, respectively). The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (Phylobayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95, closed light circles ≥ 0.95/75, open circles ≥ 0.8/50.

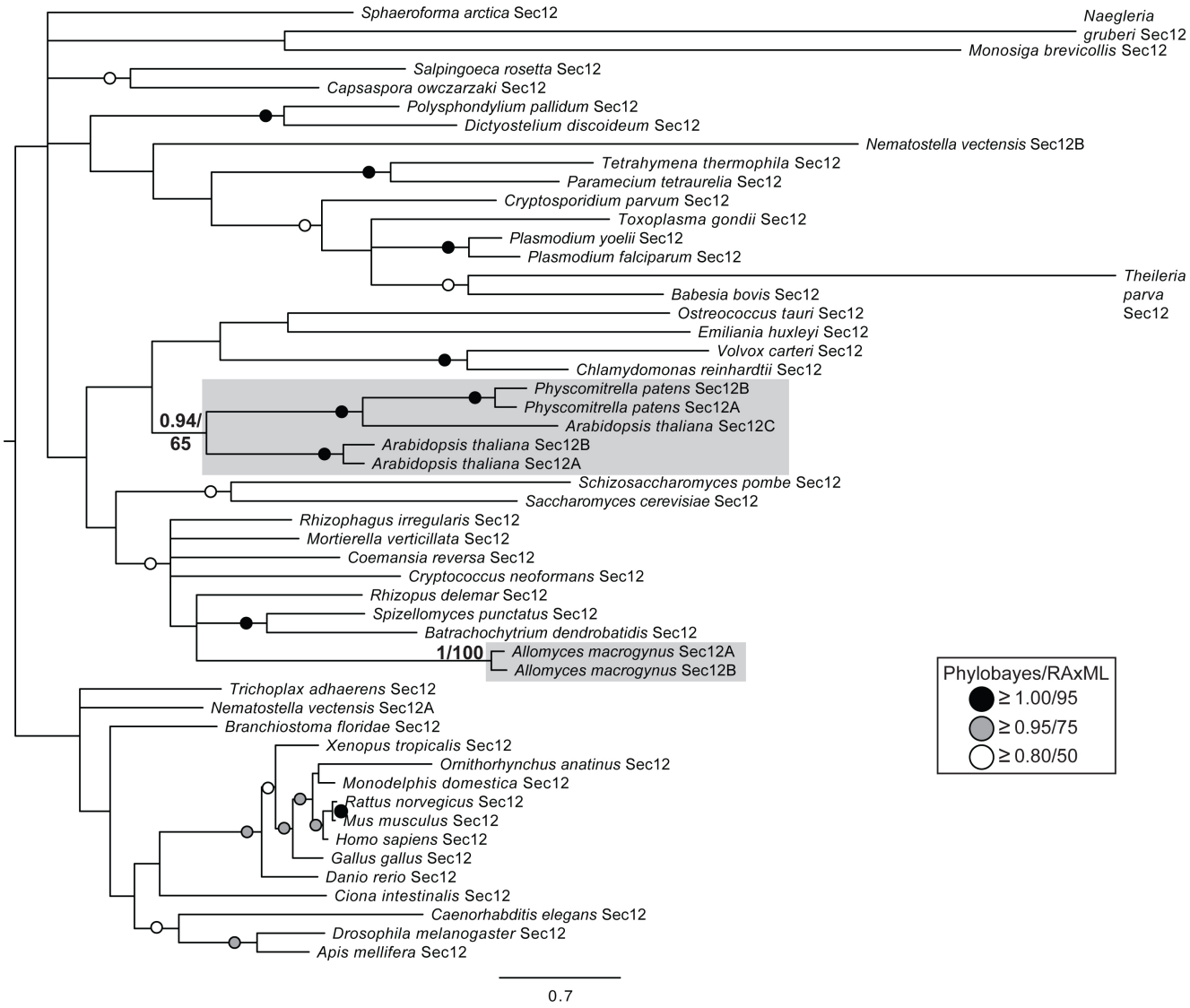

Sphaeroforma arctica Sec12
Naegleria gruberi Sec12
Monosiga brevicollis Sec12
Salpingoeca rosetta Sec12
Capsaspora owczarzaki Sec12
Polysphondylium pallidum Sec12
Dictyostelium discoideum Sec12
Nematostella vectensis Sec12B
Tetrahymena thermophila Sec12
Paramecium tetraurelia Sec12
Cryptosporidium parvum Sec12
Toxoplasma gondii Sec12
Plasmodium yoelii Sec12
Plasmodium falciparum Sec12
Theileria parva Sec12
Babesia bovis Sec12
Ostreococcus tauri Sec12
Emiliania huxleyi Sec12
Volvox carteri Sec12
Chlamydomonas reinhardtii Sec12
Physcomitrella patens Sec12B
Physcomitrella patens Sec12A
Arabidopsis thaliana Sec12C
0.94/
65
Arabidopsis thaliana Sec12B
Arabidopsis thaliana Sec12A
Schizosaccharomyces pombe Sec12
Saccharomyces cerevisiae Sec12
Rhizophagus irregularis Sec12
Mortierella verticillata Sec12
Coemansia reversa Sec12
Cryptococcus neoformans Sec12
Rhizopus delemar Sec12
Spizellomyces punctatus Sec12
Batrachochytrium dendrobatidis Sec12
1/100
Allomyces macrogynus Sec12A
Allomyces macrogynus Sec12B
Trichoplax adhaerens Sec12
Nematostella vectensis Sec12A
Branchiostoma floridae Sec12
Xenopus tropicalis Sec12
Ornithorhynchus anatinus Sec12
Monodelphis domestica Sec12
Rattus norvegicus Sec12
Mus musculus Sec12
Homo sapiens Sec12
Gallus gallus Sec12
Danio rerio Sec12
Ciona intestinalis Sec12
Caenorhabditis elegans Sec12
Drosophila melanogaster Sec12
Apis mellifera Sec12
0.7
Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50

Figure 4-6. Phylogenetic analysis of Sec13 sequences identifies a single paralogue in the LECA. To determine the number of Sec13 paralogues present in the LECA, phylogenetic analysis was carried out. The majority of taxa possess one Sec13 paralogue indicating that only one Sec13 sequence was present in the LECA. Multiple instances of lineage-specific expansion are apparent, most notably in multicellular plants (grey box). By contrast, vertebrates have not undergone any expansion. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (Phylobayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95, closed light circles ≥ 0.95/75, open circles ≥ 0.8/50.


Branchiostoma floridae Sec13
Ciona intestinalis Sec13
Drosophila melanogaster Sec13
Apis mellifera Sec13
Trichoplax adhaerens Sec13
Nematostella vectensis Sec13
Thecamonas trahens Sec13
Tetrahymena thermophila Sec13
Salpingoeca rosetta Sec13
Sphaeroforma arctica Sec13A
Reticulomyxa filosa Sec13
Paramecium tetraurelia Sec13C
Paramecium tetraurelia Sec13B
Paramecium tetraurelia Sec13A
Phytophthora sojae Sec13
Monosiga brevicollis Sec13
Guillardia theta Sec13
Phaeodactylum tricornutum Sec13
Ectocarpus siliculosus Sec13
Emiliania huxleyi Sec13
Rozella allomycis Sec13
Polysphondylium pallidum Sec13
Dictyostelium discoideum Sec13
Ostreococcus tauri Sec13
Micromonas pusila Sec13
Volvox carteri Sec13
Chlamydomonas reinhardtii Sec13
Toxoplasma gondii Sec13
Cryptosporidium parvum Sec13
Naegleria gruberi Sec13
Cyanophora paradoxa Sec13
Capsaspora owczarzaki Sec13
Cyanidioschyzon merolae Sec13
Caenorhabditis elegans Sec13
Bigelowiella natans Sec13
Physcomitrella patens Sec13D
Physcomitrella patens Sec13C
Physcomitrella patens Sec13B
Physcomitrella patens Sec13A
0.94/
76
Arabidopsis thaliana Sec13B
Arabidopsis thaliana Sec13A
Piromyces sp. Sec13
Rhizophagus irregularis Sec13
Rhizopus delemar Sec13
Schizosaccharomyces pombe Sec13
Saccharomyces cerevisiae Sec13
Neurospora crassa Sec13
Mortierella verticillata Sec13
Coemansia reversa Sec13
Ustilago maydis Sec13
Cryptococcus neoformans Sec13
Conidiobolus coronatus Sec13
Batrachochytrium dendrobatidis Sec13
Spizellomyces punctatus Sec13
Catenaria anguillulae Sec13
Allomyces macrogynus Sec13A
Allomyces macrogynus Sec13B
Xenopus tropicalis Sec13
Ornithorhynchus anatinus Sec13
Monodelphis domestica Sec13
Rattus norvegicus Sec13
Mus musculus Sec13
Homo sapiens Sec13
Gallus gallus Sec13
Danio rerio Sec13
Sphaeroforma arctica Sec13B
Trichomonas vaginalis Sec13B
Trichomonas vaginalis Sec13A
Nannochloropsis gaditana Sec13
Paramecium tetraurelia Sec13D
Fonticula alba Sec13
Giardia lamblia WB Sec13
Giardia lamblia GS Sec13
Nosema ceranae Sec13
Encephalitozoon cuniculi Sec13
Trypanosoma brucei Sec13
Leishmania major Sec13A
Symbiodinium sp. Sec13d
Symbiodinium sp. Sec13c
Symbiodinium sp. Sec13b
Plasmodium yoelii Sec13
Plasmodium falciparum Sec13
Theileria parva 13
Babesia bovis Sec13B
Babesia bovis Sec13A
Leishmania major Sec13B
Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50
0.9

Figure 4-7. Phylogenetic analysis points to a single Sec16 paralogue in the LECA. Phylogenetic analysis was carried out to determine the number of Sec16 paralogues present in the LECA. The majority of taxa sampled possess a single Sec16 paralogue, indicating that a single Sec16 sequence was present in the LECA. Lineage specific expansions have occurred such as in metazoans and Archaeplastida (lower and upper grey boxes, respectively), as well as *G. theta, T. pseudonana,* and *C. neoformans*. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (Phylobayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95, closed light circles ≥ 0.95/75, open circles ≥ 0.8/50.

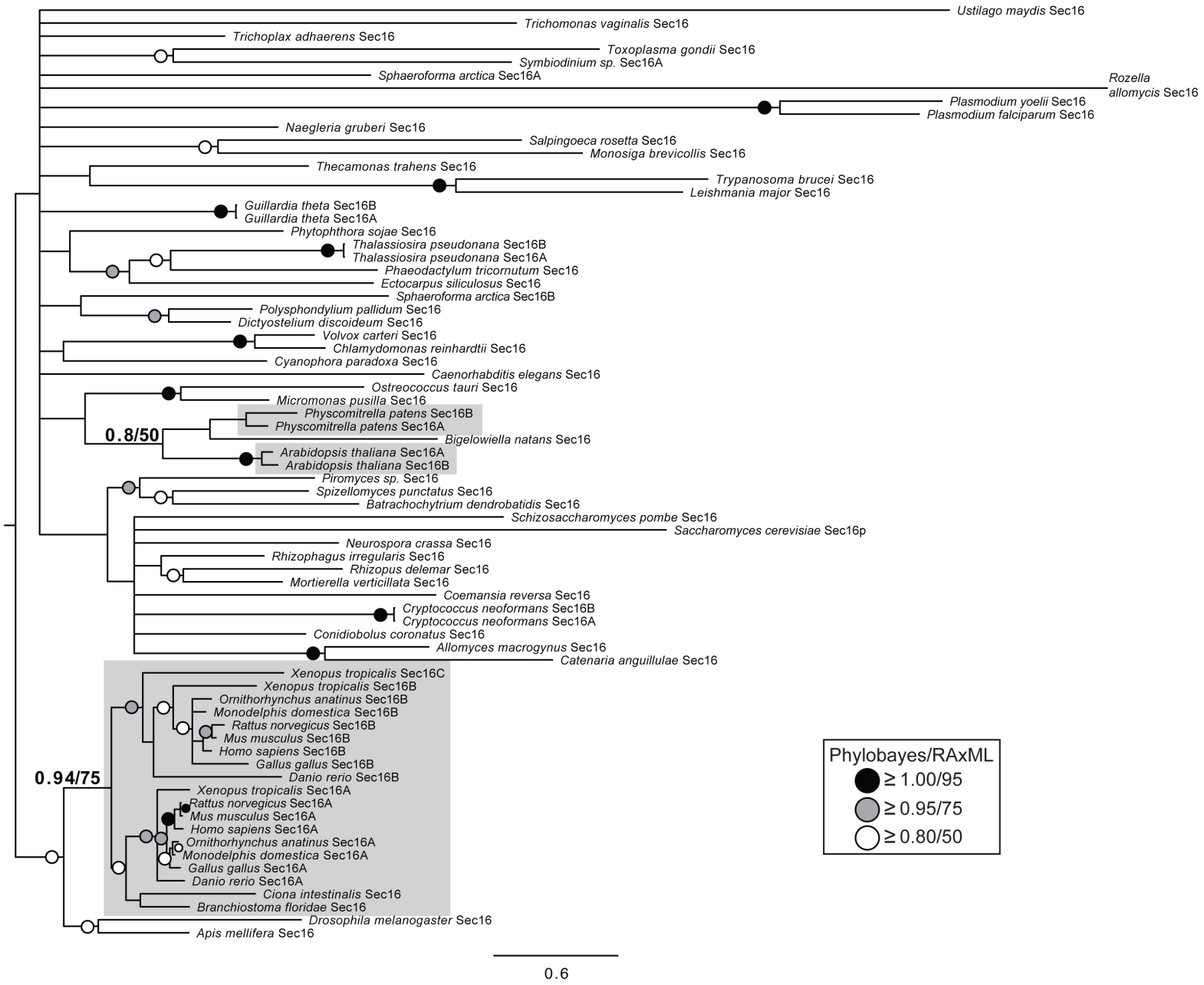

Ustilago maydis Sec16
Trichomonas vaginalis Sec16
Trichoplax adhaerens Sec16
Toxoplasma gondii Sec16
Symbiodinium sp. Sec16A
Sphaeroforma arctica Sec16A
Rozella allomycis Sec16
Plasmodium yoelii Sec16
Plasmodium falciparum Sec16
Naegleria gruberi Sec16
Salpingoeca rosetta Sec16
Monosiga brevicollis Sec16
Thecamonas trahens Sec16
Trypanosoma brucei Sec16
Leishmania major Sec16
Guillardia theta Sec16B
Guillardia theta Sec16A
Phytophthora sojae Sec16
Thalassiosira pseudonana Sec16B
Thalassiosira pseudonana Sec16A
Phaeodactylum tricornutum Sec16
Ectocarpus siliculosus Sec16
Sphaeroforma arctica Sec16B
Polysphondylium pallidum Sec16
Dictyostelium discoideum Sec16
Volvox carteri Sec16
Chlamydomonas reinhardtii Sec16
Cyanophora paradoxa Sec16
Caenorhabditis elegans Sec16
Ostreococcus tauri Sec16
Micromonas pusilla Sec16
Physcomitrella patens Sec16B
Physcomitrella patens Sec16A
0.8/50
Bigelowiella natans Sec16
Arabidopsis thaliana Sec16A
Arabidopsis thaliana Sec16B
Piromyces sp. Sec16
Spizellomyces punctatus Sec16
Batrachochytrium dendrobatidis Sec16
Schizosaccharomyces pombe Sec16
Saccharomyces cerevisiae Sec16p
Neurospora crassa Sec16
Rhizophagus irregularis Sec16
Rhizopus delemar Sec16
Mortierella verticillata Sec16
Coemansia reversa Sec16
Cryptococcus neoformans Sec16B
Cryptococcus neoformans Sec16A
Conidiobolus coronatus Sec16
Allomyces macrogynus Sec16
Catenaria anguillulae Sec16
Xenopus tropicalis Sec16C
Xenopus tropicalis Sec16B
Ornithorhynchus anatinus Sec16B
Monodelphis domestica Sec16B
Rattus norvegicus Sec16B
Mus musculus Sec16B
Homo sapiens Sec16B
Gallus gallus Sec16B
Danio rerio Sec16B
0.94/75
Xenopus tropicalis Sec16A
Rattus norvegicus Sec16A
Mus musculus Sec16A
Homo sapiens Sec16A
Ornithorhynchus anatinus Sec16A
Monodelphis domestica Sec16A
Gallus gallus Sec16A
Danio rerio Sec16A
Ciona intestinalis Sec16
Branchiostoma floridae Sec16
Drosophila melanogaster Sec16
Apis mellifera Sec16
0.6
Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50

Figure 4-8. Phylogenetic analysis identifies a single Sec23 sequence in the LECA. Phylogenetic analysis was carried out to determine the number of Sec23 paralogues present in the LECA. By and large, a single clade for each supergroup appears to have been reconstructed. The exception is the Archaeplastida, where two *A. thaliana, P. patens,* and *M. pusilla* containing clades are present (upper and middle grey boxes). Nonetheless, the data suggest the presence of a single Sec23 paralogue in the LECA. Major expansions of Sec23 can be seen in vertebrates and in multicellular plants (grey boxes). Smaller, species-specific expansions are also visible in other lineages. Although likely an artefact, a sequence from *G. lamblia* also grouped with the *A. thaliana* clade (upper grey box). The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (Phylobayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95, closed light circles ≥ 0.95/75, open circles ≥ 0.8/50.

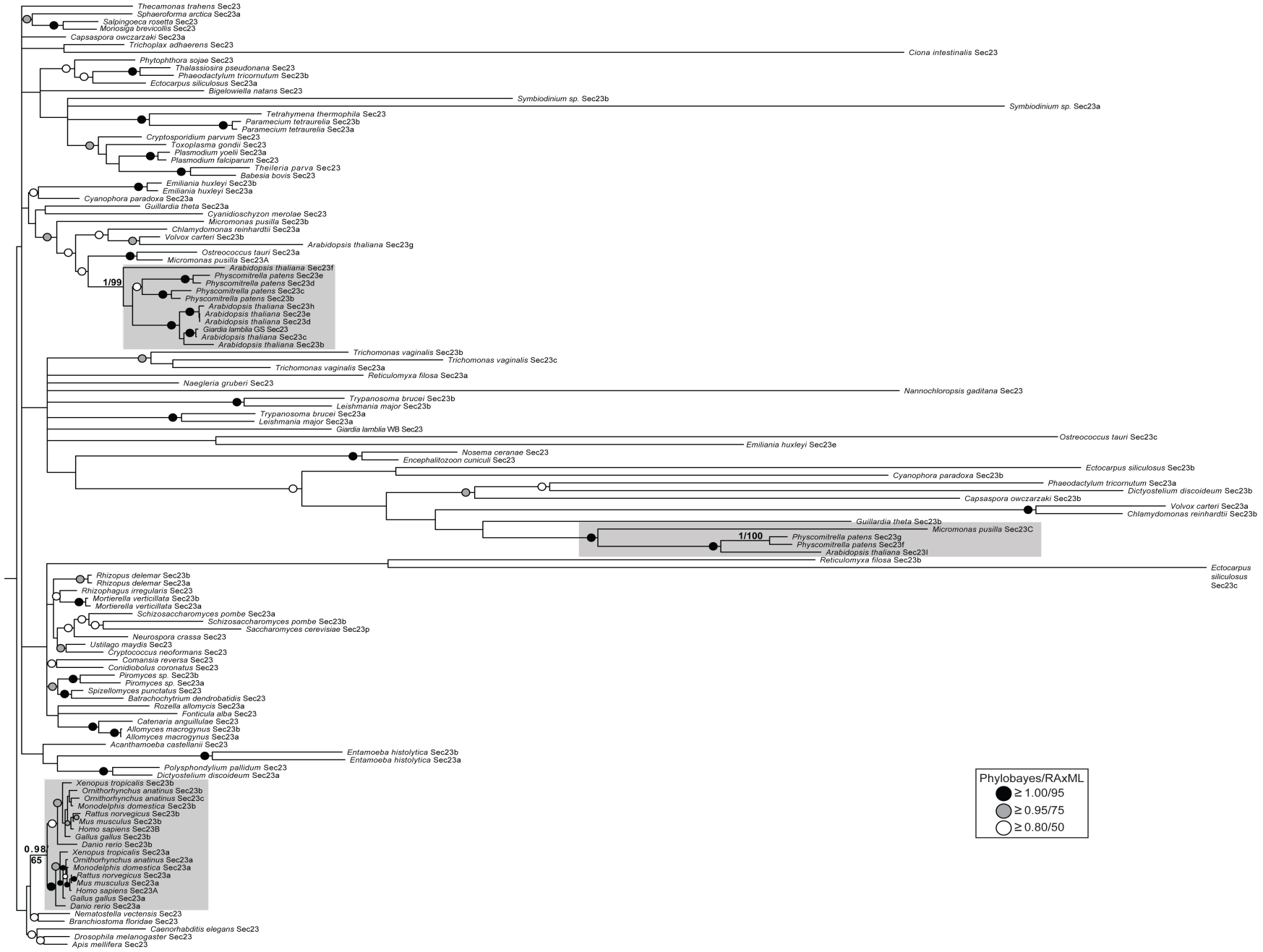

Thecamonas trahens Sec23
Sphaeroforma arctica Sec23a
Salpingoeca rosetta Sec23
Monosiga brevicollis Sec23
Capsaspora owczarzaki Sec23a
Trichoplax adhaerens Sec23
Ciona intestinalis Sec23
Phytophthora sojae Sec23
Thalassiosira pseudonana Sec23
Phaeodactylum tricornutum Sec23b
Ectocarpus siliculosus Sec23a
Bigelowiella natans Sec23
Symbiodinium sp. Sec23b
Symbiodinium sp. Sec23a
Tetrahymena thermophila Sec23
Paramecium tetraurelia Sec23b
Paramecium tetraurelia Sec23a
Cryptosporidium parvum Sec23
Toxoplasma gondii Sec23
Plasmodium yoelii Sec23a
Plasmodium falciparum Sec23
Theileria parva Sec23
Babesia bovis Sec23
Emiliania huxleyi Sec23b
Emiliania huxleyi Sec23a
Cyanophora paradoxa Sec23a
Guillardia theta Sec23a
Cyanidioschyzon merolae Sec23
Micromonas pusilla Sec23b
Chlamydomonas reinhardtii Sec23a
Volvox carteri Sec23b
Arabidopsis thaliana Sec23g
Ostreococcus tauri Sec23a
Micromonas pusilla Sec23A
Arabidopsis thaliana Sec23f
1/99
Physcomitrella patens Sec23e
Physcomitrella patens Sec23d
Physcomitrella patens Sec23c
Physcomitrella patens Sec23b
Arabidopsis thaliana Sec23h
Arabidopsis thaliana Sec23e
Arabidopsis thaliana Sec23d
Giardia lamblia GS Sec23
Arabidopsis thaliana Sec23c
Arabidopsis thaliana Sec23b
Trichomonas vaginalis Sec23b
Trichomonas vaginalis Sec23c
Trichomonas vaginalis Sec23a
Reticulomyxa filosa Sec23a
Naegleria gruberi Sec23
Nannochloropsis gaditana Sec23
Trypanosoma brucei Sec23b
Leishmania major Sec23b
Trypanosoma brucei Sec23a
Leishmania major Sec23a
Giardia lamblia WB Sec23
Ostreococcus tauri Sec23c
Emiliania huxleyi Sec23e
Nosema ceranae Sec23
Encephalitozoon cuniculi Sec23
Ectocarpus siliculosus Sec23b
Cyanophora paradoxa Sec23b
Phaeodactylum tricornutum Sec23a
Dictyostelium discoideum Sec23b
Capsaspora owczarzaki Sec23b
Volvox carteri Sec23a
Chlamydomonas reinhardtii Sec23b
Guillardia theta Sec23b
Micromonas pusilla Sec23C
1/100
Physcomitrella patens Sec23g
Physcomitrella patens Sec23f
Arabidopsis thaliana Sec23l
Reticulomyxa filosa Sec23b
Ectocarpus siliculosus Sec23c
Rhizopus delemar Sec23b
Rhizopus delemar Sec23a
Rhizophagus irregularis Sec23
Mortierella verticillata Sec23b
Mortierella verticillata Sec23a
Schizosaccharomyces pombe Sec23a
Schizosaccharomyces pombe Sec23b
Saccharomyces cerevisiae Sec23p
Neurospora crassa Sec23
Ustilago maydis Sec23
Cryptococcus neoformans Sec23
Comansia reversa Sec23
Conidiobolus coronatus Sec23
Piromyces sp. Sec23b
Piromyces sp. Sec23a
Spizellomyces punctatus Sec23
Batrachochytrium dendrobatidis Sec23
Rozella allomycis Sec23a
Fonticula alba Sec23
Catenaria anguillulae Sec23
Allomyces macrogynus Sec23b
Allomyces macrogynus Sec23a
Acanthamoeba castellanii Sec23
Entamoeba histolytica Sec23b
Entamoeba histolytica Sec23a
Polysphondylium pallidum Sec23
Dictyostelium discoideum Sec23a
Xenopus tropicalis Sec23b
Ornithorhynchus anatinus Sec23b
Ornithorhynchus anatinus Sec23c
Monodelphis domestica Sec23b
Rattus norvegicus Sec23b
Mus musculus Sec23b
Homo sapiens Sec23B
Gallus gallus Sec23b
Danio rerio Sec23b
0.98
65
Xenopus tropicalis Sec23a
Ornithorhynchus anatinus Sec23a
Monodelphis domestica Sec23a
Rattus norvegicus Sec23a
Mus musculus Sec23a
Homo sapiens Sec23A
Gallus gallus Sec23a
Danio rerio Sec23a
Nematostella vectensis Sec23
Branchiostoma floridae Sec23
Caenorhabditis elegans Sec23
Drosophila melanogaster Sec23
Apis mellifera Sec23
Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50
0.7

Figure 4-9. Phylogenetic analysis of Sec24 sequences suggests the presence of multiple paralogues in the LECA. To determine the number of ancient Sec24 paralogues, phylogenetic analysis was carried out. Multiple clades of some major eukaryotic lineages (Fungi, Holozoa, Archaeplastida, and SAR/CCTH) were reconstructed suggesting that multiple paralogues of Sec24 may have been present in the LECA, although backbone nodes are largely unresolved. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (Phylobayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95, closed light circles ≥ 0.95/75, open circles ≥ 0.8/50. For enlarged clades see Appendix.

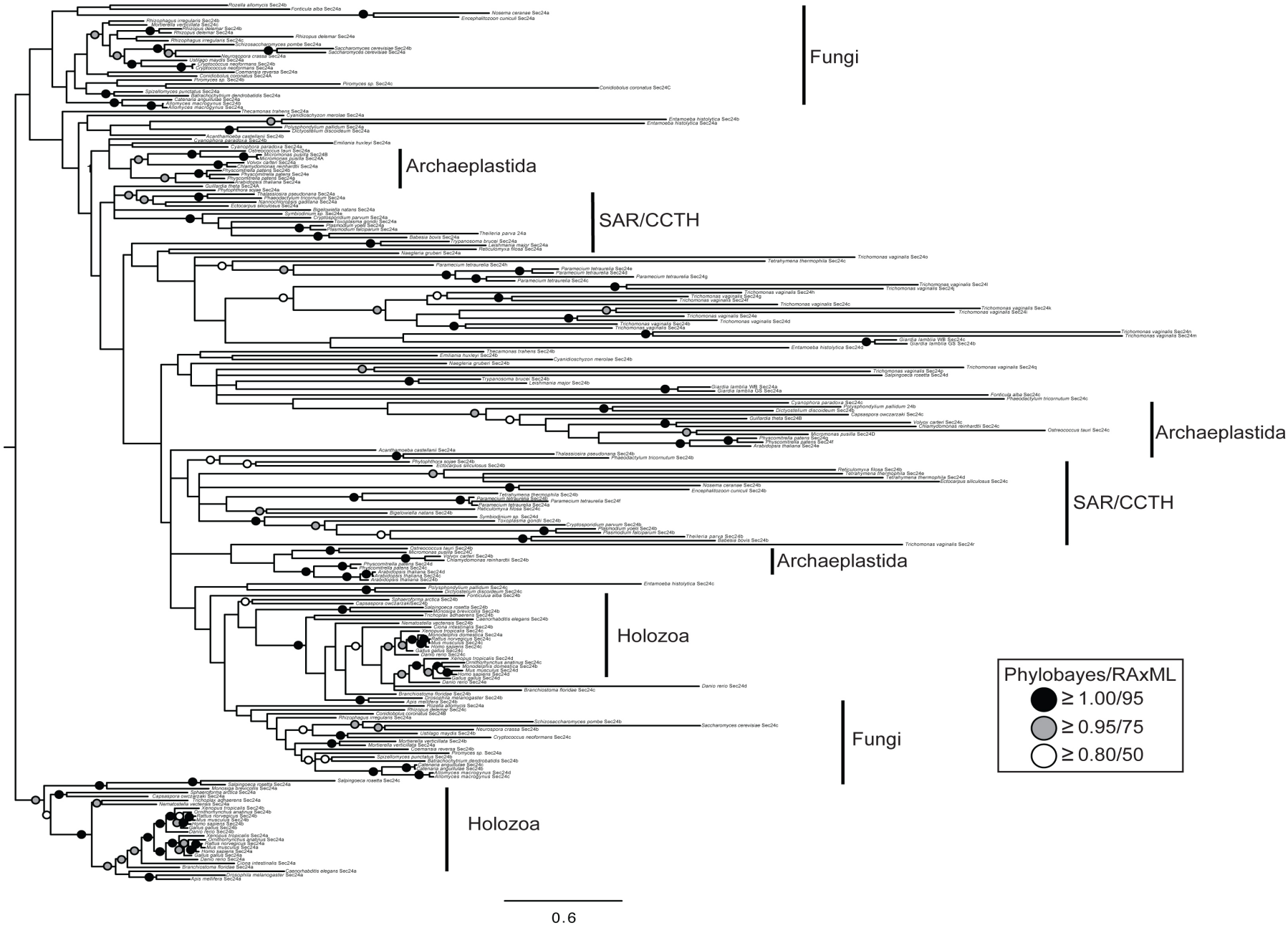

Fungi
Archaeplastida
SAR/CCTH
Archaeplastida
SAR/CCTH
Archaeplastida
Holozoa
Fungi
Holozoa
Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50
0.6

Figure 4-10. Phylogenetic analysis indicates that a single Sec31 paralogue was present in the LECA. Phylogenetic analysis was carried out to determine the number of Sec31 paralogues present in the LECA. Generally, a single clade is reconstructed for most major eukaryotic lineages, suggesting that only one Sec31 sequence was present in the LECA. Expansion has occurred in many eukaryotic lineages independently, notably in vertebrates and in multicellular plants (grey boxes). The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (Phylobayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95, closed light circles ≥ 0.95/75, open circles ≥ 0.8/50.

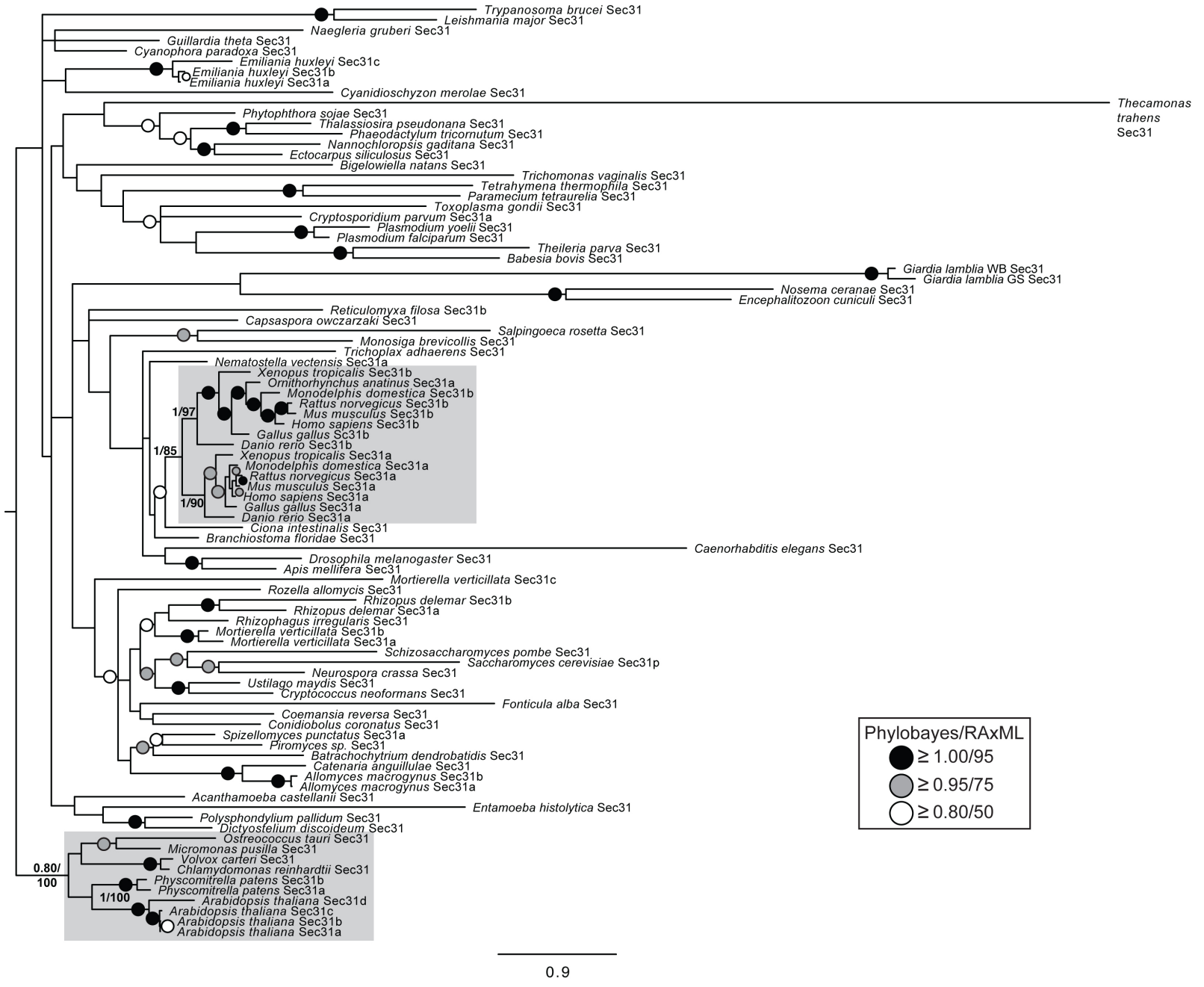

Trypanosoma brucei Sec31
Leishmania major Sec31
Naegleria gruberi Sec31
Guillardia theta Sec31
Cyanophora paradoxa Sec31
Emiliania huxleyi Sec31c
Emiliania huxleyi Sec31b
Emiliania huxleyi Sec31a
Cyanidioschyzon merolae Sec31
Thecamonas trahens Sec31
Phytophthora sojae Sec31
Thalassiosira pseudonana Sec31
Phaeodactylum tricornutum Sec31
Nannochloropsis gaditana Sec31
Ectocarpus siliculosus Sec31
Bigelowiella natans Sec31
Trichomonas vaginalis Sec31
Tetrahymena thermophila Sec31
Paramecium tetraurelia Sec31
Toxoplasma gondii Sec31
Cryptosporidium parvum Sec31a
Plasmodium yoelii Sec31
Plasmodium falciparum Sec31
Theileria parva Sec31
Babesia bovis Sec31
Giardia lamblia WB Sec31
Giardia lamblia GS Sec31
Nosema ceranae Sec31
Encephalitozoon cuniculi Sec31
Reticulomyxa filosa Sec31b
Capsaspora owczarzaki Sec31
Salpingoeca rosetta Sec31
Monosiga brevicollis Sec31
Trichoplax adhaerens Sec31
Nematostella vectensis Sec31a
Xenopus tropicalis Sec31b
Ornithorhynchus anatinus Sec31a
Monodelphis domestica Sec31b
Rattus norvegicus Sec31b
Mus musculus Sec31b
Homo sapiens Sec31b
Gallus gallus Sc31b
Danio rerio Sec31b
Xenopus tropicalis Sec31a
Monodelphis domestica Sec31a
Rattus norvegicus Sec31a
Mus musculus Sec31a
Homo sapiens Sec31a
Gallus gallus Sec31a
Danio rerio Sec31a
Ciona intestinalis Sec31
Branchiostoma floridae Sec31
Caenorhabditis elegans Sec31
Drosophila melanogaster Sec31
Apis mellifera Sec31
Mortierella verticillata Sec31c
Rozella allomycis Sec31
Rhizopus delemar Sec31b
Rhizopus delemar Sec31a
Rhizophagus irregularis Sec31
Mortierella verticillata Sec31b
Mortierella verticillata Sec31a
Schizosaccharomyces pombe Sec31
Saccharomyces cerevisiae Sec31p
Neurospora crassa Sec31
Ustilago maydis Sec31
Cryptococcus neoformans Sec31
Fonticula alba Sec31
Coemansia reversa Sec31
Conidiobolus coronatus Sec31
Spizellomyces punctatus Sec31a
Piromyces sp. Sec31
Batrachochytrium dendrobatidis Sec31
Catenaria anguillulae Sec31
Allomyces macrogynus Sec31b
Allomyces macrogynus Sec31a
Acanthamoeba castellanii Sec31
Entamoeba histolytica Sec31
Polysphondylium pallidum Sec31
Dictyostelium discoideum Sec31
Ostreococcus tauri Sec31
Micromonas pusilla Sec31
Volvox carteri Sec31
Chlamydomonas reinhardtii Sec31
Physcomitrella patens Sec31b
Physcomitrella patens Sec31a
Arabidopsis thaliana Sec31d
Arabidopsis thaliana Sec31c
Arabidopsis thaliana Sec31b
Arabidopsis thaliana Sec31a
1/97
1/85
1/90
0.80/
100
1/100
Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50
0.9

Figure 4-11. Phylogenetic analysis identifies multiple expansions of Sar1 in Archaeplastida. To determine whether the expansions of Sar1 observed in *P. patens* and *A. thaliana* (Figure 4-4) were limited to those two taxa or whether expansions have also occurred in other plant lineages, an expanded phylogenetic analysis of archaeplastid Sar1 sequences was carried out. Multiple lineage-specific expansions can be observed in the Grasses (*Zea mays, Oryzia sativa, Setaria italica*), Papilionoidae (*Medicago truncatula, Glycine max*), Rosales-Cucurbitales (*Prunus persica, Fragaria vesca, Cucumis sativus*), and Brassicaceae (*A. thaliana, Capsella rubella,* and *Brassica rapa)* (grey boxes). Species-specific expansions are also prevalent, indicating that multiple expansions of Sar1 have occurred during archaeplastid evolution. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (Phylobayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95, closed light circles ≥ 0.95/75, open circles ≥ 0.8/50.

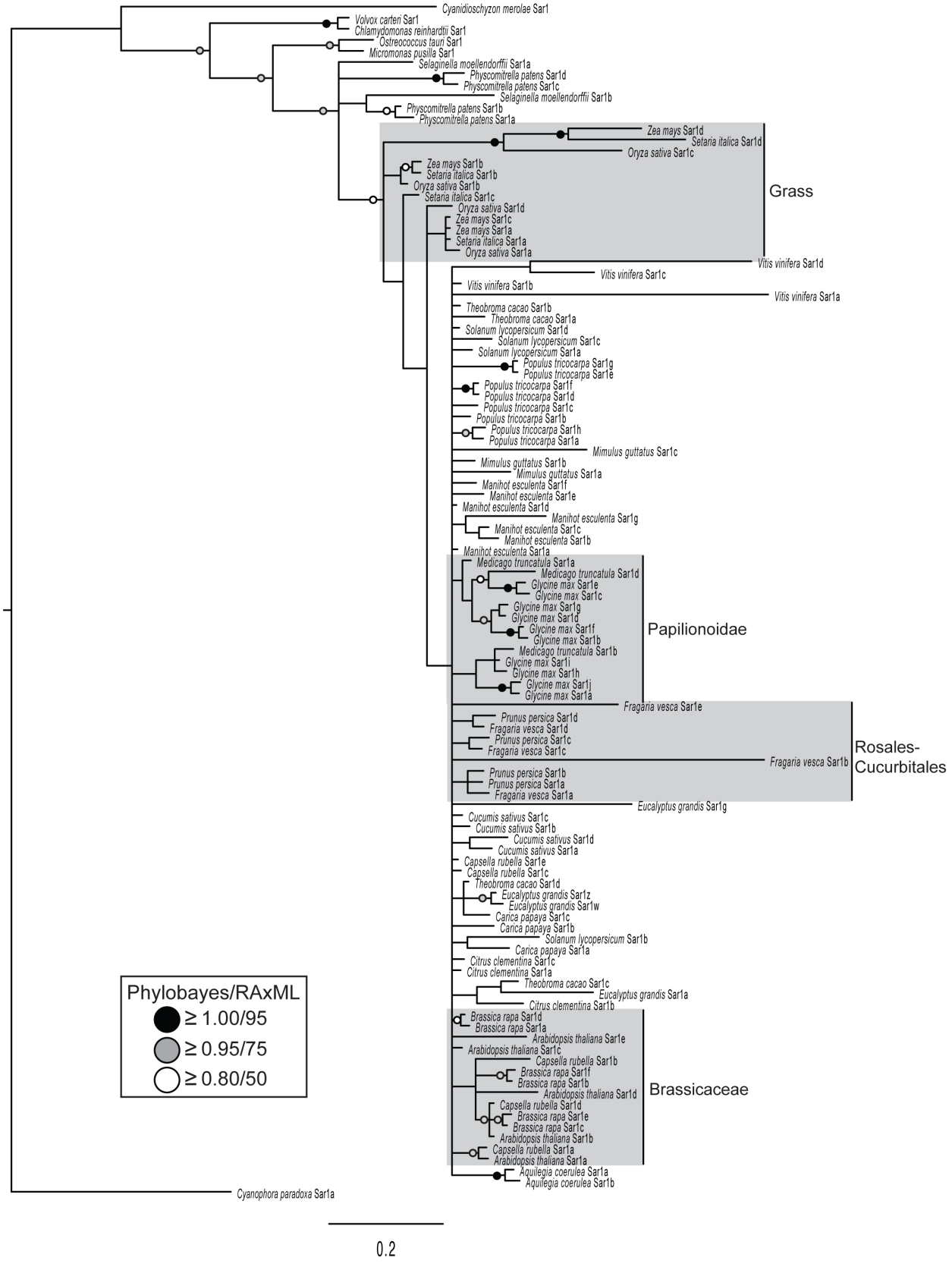

Cyanidioschyzon merolae Sar1
Volvox carteri Sar1
Chlamydomonas reinhardtii Sar1
Ostreococcus tauri Sar1
Micromonas pusilla Sar1
Selaginella moellendorffii Sar1a
Physcomitrella patens Sar1d
Physcomitrella patens Sar1c
Selaginella moellendorffii Sar1b
Physcomitrella patens Sar1b
Physcomitrella patens Sar1a
Zea mays Sar1d
Setaria italica Sar1d
Oryza sativa Sar1c
Zea mays Sar1b
Setaria italica Sar1b
Oryza sativa Sar1b
Setaria italica Sar1c
Oryza sativa Sar1d
Zea mays Sar1c
Zea mays Sar1a
Setaria italica Sar1a
Oryza sativa Sar1a
Grass
Vitis vinifera Sar1d
Vitis vinifera Sar1c
Vitis vinifera Sar1b
Vitis vinifera Sar1a
Theobroma cacao Sar1b
Theobroma cacao Sar1a
Solanum lycopersicum Sar1d
Solanum lycopersicum Sar1c
Solanum lycopersicum Sar1a
Populus tricocarpa Sar1g
Populus tricocarpa Sar1e
Populus tricocarpa Sar1f
Populus tricocarpa Sar1d
Populus tricocarpa Sar1c
Populus tricocarpa Sar1b
Populus tricocarpa Sar1h
Populus tricocarpa Sar1a
Mimulus guttatus Sar1c
Mimulus guttatus Sar1b
Mimulus guttatus Sar1a
Manihot esculenta Sar1f
Manihot esculenta Sar1e
Manihot esculenta Sar1d
Manihot esculenta Sar1g
Manihot esculenta Sar1c
Manihot esculenta Sar1b
Manihot esculenta Sar1a
Medicago truncatula Sar1a
Medicago truncatula Sar1d
Glycine max Sar1e
Glycine max Sar1c
Glycine max Sar1g
Glycine max Sar1d
Glycine max Sar1f
Glycine max Sar1b
Medicago truncatula Sar1b
Glycine max Sar1i
Glycine max Sar1h
Glycine max Sar1j
Glycine max Sar1a
Papilionoidae
Fragaria vesca Sar1e
Prunus persica Sar1d
Fragaria vesca Sar1d
Prunus persica Sar1c
Fragaria vesca Sar1c
Fragaria vesca Sar1b
Prunus persica Sar1b
Prunus persica Sar1a
Fragaria vesca Sar1a
Rosales-
Cucurbitales
Eucalyptus grandis Sar1g
Cucumis sativus Sar1c
Cucumis sativus Sar1b
Cucumis sativus Sar1d
Cucumis sativus Sar1a
Capsella rubella Sar1e
Capsella rubella Sar1c
Theobroma cacao Sar1d
Eucalyptus grandis Sar1z
Eucalyptus grandis Sar1w
Carica papaya Sar1c
Carica papaya Sar1b
Solanum lycopersicum Sar1b
Carica papaya Sar1a
Citrus clementina Sar1c
Citrus clementina Sar1a
Theobroma cacao Sar1c
Eucalyptus grandis Sar1a
Citrus clementina Sar1b
Brassica rapa Sar1d
Brassica rapa Sar1a
Arabidopsis thaliana Sar1e
Arabidopsis thaliana Sar1c
Capsella rubella Sar1b
Brassica rapa Sar1f
Brassica rapa Sar1b
Arabidopsis thaliana Sar1d
Capsella rubella Sar1d
Brassica rapa Sar1e
Brassica rapa Sar1c
Arabidopsis thaliana Sar1b
Capsella rubella Sar1a
Arabidopsis thaliana Sar1a
Brassicaceae
Aquilegia coerulea Sar1a
Aquilegia coerulea Sar1b
Cyanophora paradoxa Sar1a
Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50
0.2

Figure 4-12. Phylogenetic analysis identifies an early duplication of Sec12 in embryophytes. To determine the relative timing of the gene duplication giving rise to the two archaeplastid Sec12 paralogues observed in Figure 4-5, an expanded phylogenetic analysis of archaeplastid Sec12 sequences was carried out. The analysis identified a gene duplication early in embryophytes (grey box), followed by multiple instances of species-specific expansion. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (Phylobayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95, closed light circles ≥ 0.95/75, open circles ≥ 0.8/50.

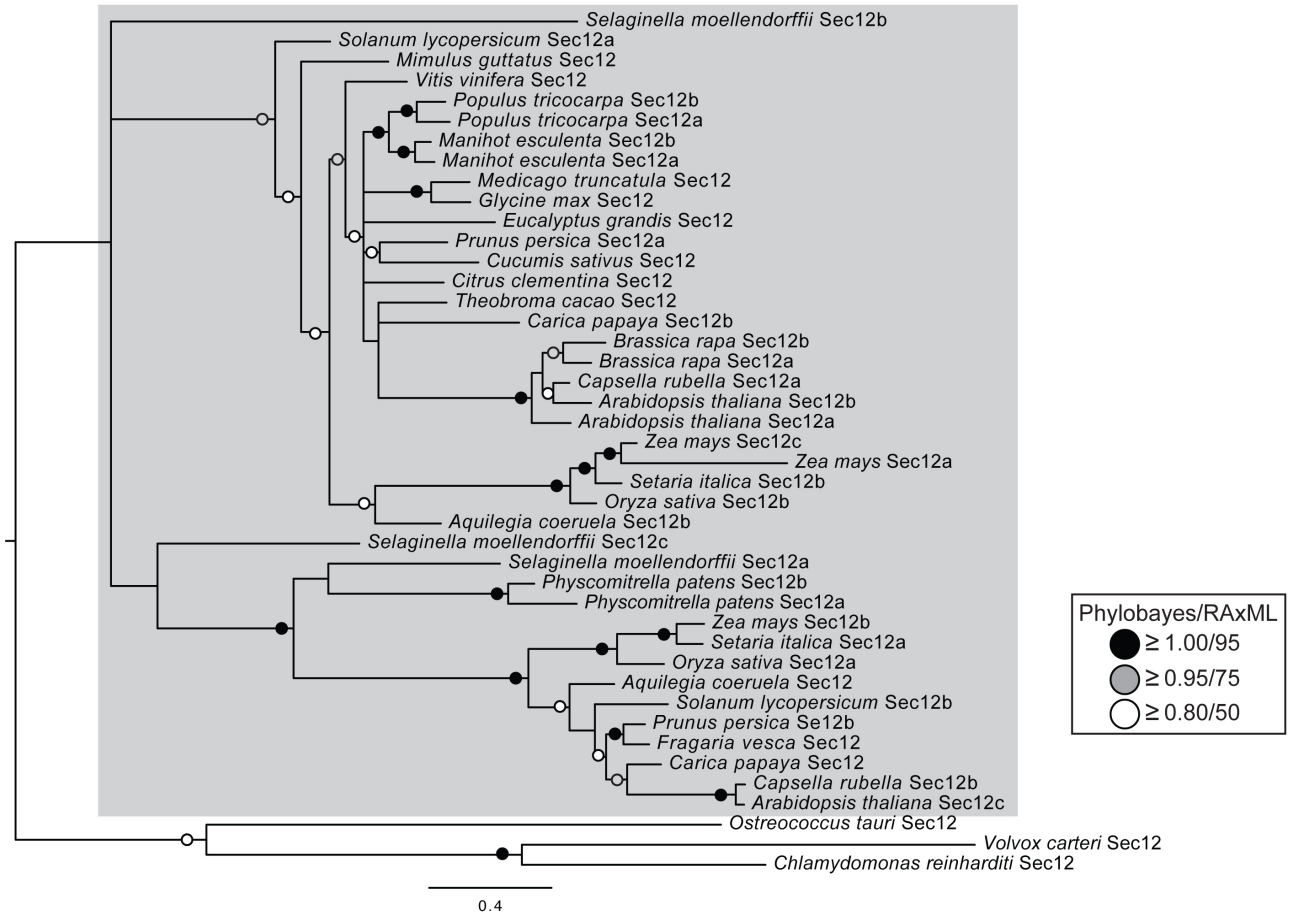

Selaginella moellendorffii Sec12b
Solanum lycopersicum Sec12a
Mimulus guttatus Sec12
Vitis vinifera Sec12
Populus tricocarpa Sec12b
Populus tricocarpa Sec12a
Manihot esculenta Sec12b
Manihot esculenta Sec12a
Medicago truncatula Sec12
Glycine max Sec12
Eucalyptus grandis Sec12
Prunus persica Sec12a
Cucumis sativus Sec12
Citrus clementina Sec12
Theobroma cacao Sec12
Carica papaya Sec12b
Brassica rapa Sec12b
Brassica rapa Sec12a
Capsella rubella Sec12a
Arabidopsis thaliana Sec12b
Arabidopsis thaliana Sec12a
Zea mays Sec12c
Zea mays Sec12a
Setaria italica Sec12b
Oryza sativa Sec12b
Aquilegia coeruela Sec12b
Selaginella moellendorffii Sec12c
Selaginella moellendorffii Sec12a
Physcomitrella patens Sec12b
Physcomitrella patens Sec12a
Zea mays Sec12b
Setaria italica Sec12a
Oryza sativa Sec12a
Aquilegia coeruela Sec12
Solanum lycopersicum Sec12b
Prunus persica Se12b
Fragaria vesca Sec12
Carica papaya Sec12
Capsella rubella Sec12b
Arabidopsis thaliana Sec12c
Ostreococcus tauri Sec12
Volvox carteri Sec12
Chlamydomonas reinharditi Sec12
0.4
Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50

Figure 4-13. Phylogenetic analysis identifies multiple expansions of Sec13 in embryophytes. To determine whether the expansions of Sec13 in *A. thaliana* and *P. patens* (Figure 4-6) are found broadly across the Archaeplastida, phylogenetic analysis using an expanded archaeplastid dataset was carried out. Two instances of lineage-specific expansion were reconstructed; gene duplications in the Brassicaceae (*A. thaliana, C. rubella,* and *B. rapa*, lower grey box) and in the Grass lineage (*Z. mays, O. sativa,* and *S. italica*. upper grey box) have occurred independently. Multiple instances of genome-specific expansion have also occurred (*P. patens, F. vesca, G. max*). The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (Phylobayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95, closed light circles ≥ 0.95/75, open circles ≥ 0.8/50.

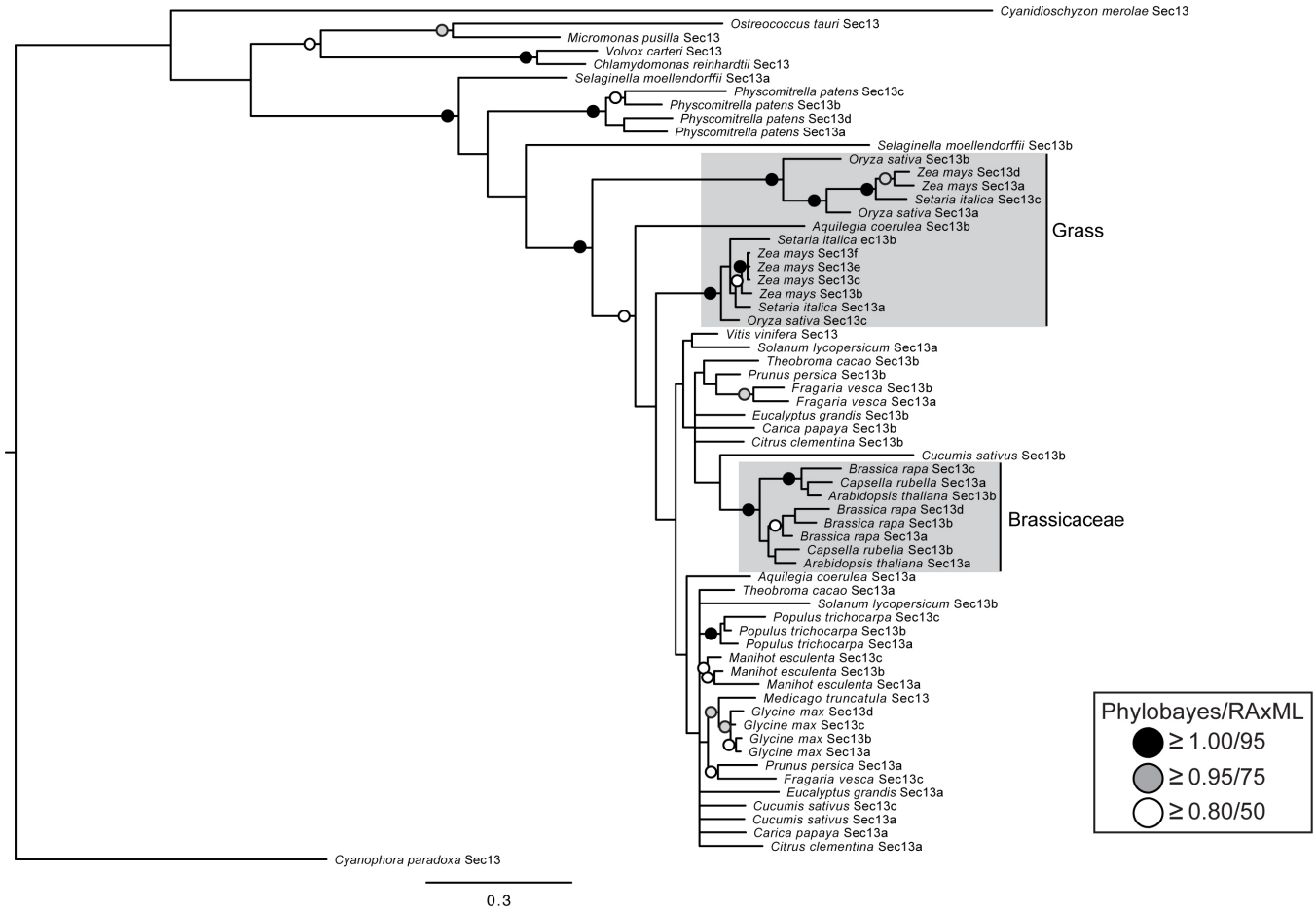

Cyanidioschyzon merolae Sec13
Ostreococcus tauri Sec13
Micromonas pusilla Sec13
Volvox carteri Sec13
Chlamydomonas reinhardtii Sec13
Selaginella moellendorffii Sec13a
Physcomitrella patens Sec13c
Physcomitrella patens Sec13b
Physcomitrella patens Sec13d
Physcomitrella patens Sec13a
Selaginella moellendorffii Sec13b
Oryza sativa Sec13b
Zea mays Sec13d
Zea mays Sec13a
Setaria italica Sec13c
Oryza sativa Sec13a
Aquilegia coerulea Sec13b
Setaria italica ec13b
Zea mays Sec13f
Zea mays Sec13e
Zea mays Sec13c
Zea mays Sec13b
Setaria italica Sec13a
Oryza sativa Sec13c
Grass
Vitis vinifera Sec13
Solanum lycopersicum Sec13a
Theobroma cacao Sec13b
Prunus persica Sec13b
Fragaria vesca Sec13b
Fragaria vesca Sec13a
Eucalyptus grandis Sec13b
Carica papaya Sec13b
Citrus clementina Sec13b
Cucumis sativus Sec13b
Brassica rapa Sec13c
Capsella rubella Sec13a
Arabidopsis thaliana Sec13b
Brassica rapa Sec13d
Brassica rapa Sec13b
Brassica rapa Sec13a
Capsella rubella Sec13b
Arabidopsis thaliana Sec13a
Brassicaceae
Aquilegia coerulea Sec13a
Theobroma cacao Sec13a
Solanum lycopersicum Sec13b
Populus trichocarpa Sec13c
Populus trichocarpa Sec13b
Populus trichocarpa Sec13a
Manihot esculenta Sec13c
Manihot esculenta Sec13b
Manihot esculenta Sec13a
Medicago truncatula Sec13
Glycine max Sec13d
Glycine max Sec13c
Glycine max Sec13b
Glycine max Sec13a
Prunus persica Sec13a
Fragaria vesca Sec13c
Eucalyptus grandis Sec13a
Cucumis sativus Sec13c
Cucumis sativus Sec13a
Carica papaya Sec13a
Citrus clementina Sec13a
Cyanophora paradoxa Sec13
0.3
Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50

Figure 4-14. Phylogenetic analysis identifies multiple expansions of Sec16 in embryophytes. To determine whether the expansions of Sec16 observed in *A. thaliana* and *P. patens* is limited to these two taxa or found broadly across Archaeplastida, an expanded phylogenetic analysis of plant Sec16 sequences was carried out. Two instances of lineage-specific expansion have occurred, one in the Brassicaceae (*A. thaliana, C. rubella,* and *B. rapa*, lower grey box), and in the Grass lineage (*Z. mays, O. sativa,* and *S. italica*, upper grey box). Multiple instances of genome-specific expansion have also occurred (*e.g., G. max* and *Solanum lycopersicum*). All expansions are contained within the embryophyte clade. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (Phylobayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95, closed light circles ≥ 0.95/75, open circles ≥ 0.8/50.

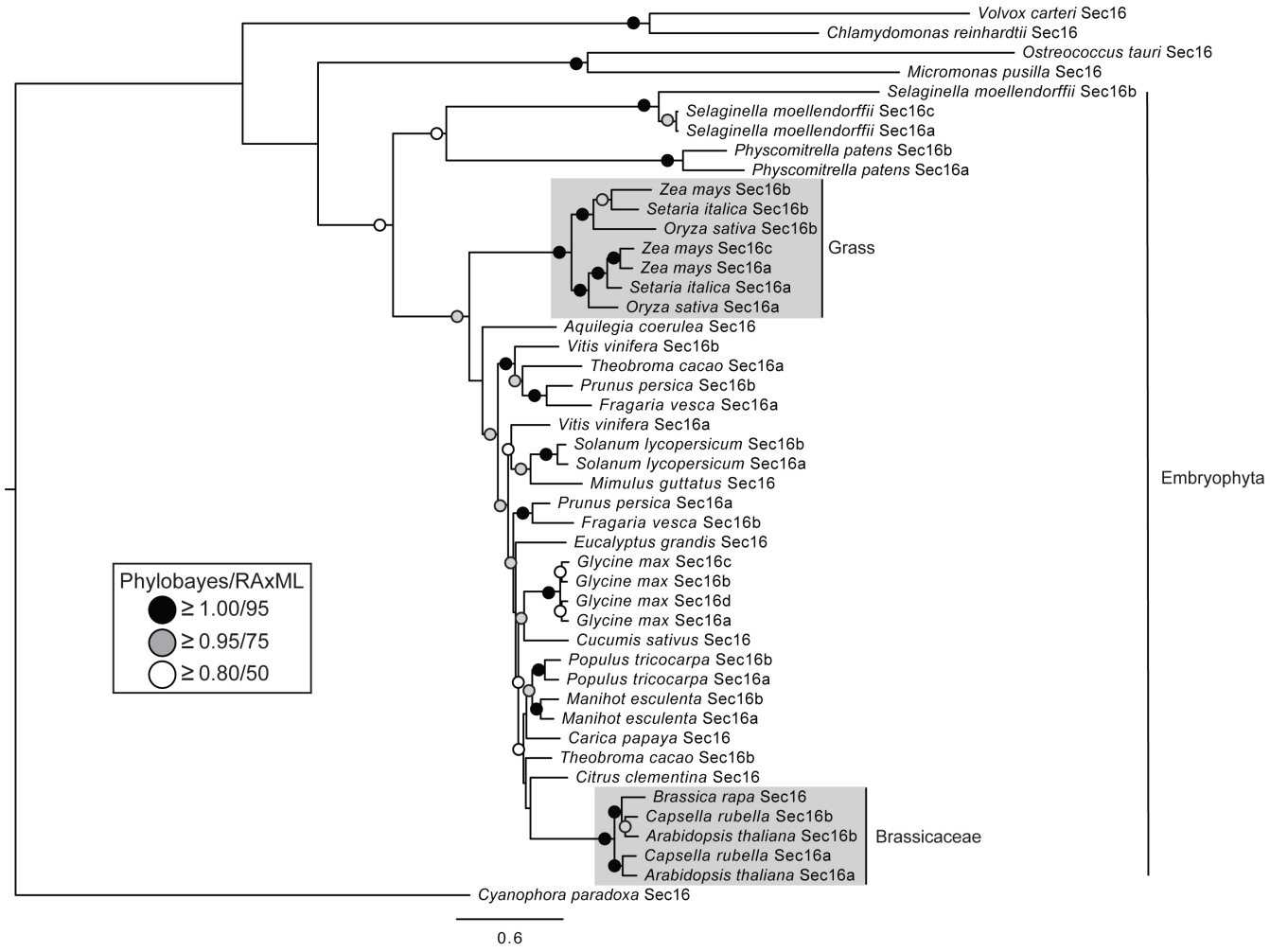

Volvox carteri Sec16
Chlamydomonas reinhardtii Sec16
Ostreococcus tauri Sec16
Micromonas pusilla Sec16
Selaginella moellendorffii Sec16b
Selaginella moellendorffii Sec16c
Selaginella moellendorffii Sec16a
Physcomitrella patens Sec16b
Physcomitrella patens Sec16a
Zea mays Sec16b
Setaria italica Sec16b
Oryza sativa Sec16b
Zea mays Sec16c
Zea mays Sec16a
Setaria italica Sec16a
Oryza sativa Sec16a
Grass
Aquilegia coerulea Sec16
Vitis vinifera Sec16b
Theobroma cacao Sec16a
Prunus persica Sec16b
Fragaria vesca Sec16a
Vitis vinifera Sec16a
Solanum lycopersicum Sec16b
Solanum lycopersicum Sec16a
Mimulus guttatus Sec16
Prunus persica Sec16a
Fragaria vesca Sec16b
Eucalyptus grandis Sec16
Glycine max Sec16c
Glycine max Sec16b
Glycine max Sec16d
Glycine max Sec16a
Cucumis sativus Sec16
Populus tricocarpa Sec16b
Populus tricocarpa Sec16a
Manihot esculenta Sec16b
Manihot esculenta Sec16a
Carica papaya Sec16
Theobroma cacao Sec16b
Citrus clementina Sec16
Brassica rapa Sec16
Capsella rubella Sec16b
Arabidopsis thaliana Sec16b
Capsella rubella Sec16a
Arabidopsis thaliana Sec16a
Brassicaceae
Embryophyta
Cyanophora paradoxa Sec16
Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50
0.6

Figure 4-15. Phylogenetic analysis of archaeplastid Sec23 sequences identifies three major clades of Sec23. To determine whether the archaeplastid Sec23 clades observed in Figure 4-8 are the result of an ancient gene duplication or phylogenetic artefact, an expanded analysis of plant Sec23 sequences was carried out. Three clades of Sec23 were reconstructed (grey boxes). Only two paralogues were found in the glaucophyta *Cyanophora paradoxa* (middle and lower grey boxes), suggesting that the ancestral archaeplastid possessed two Sec23 paralogues. A second duplication then occurred in the ancestor of Viridiplantae (upper grey box). Duplication in spermatophytes has also occurred in one of the clades (vertical lines, middle grey box). Multiple instances of species-specific expansions have also occurred in each of the three clades. Together, this suggests that the archaeplastid ancestor likely possessed two Sec23 paralogues. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (Phylobayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95, closed light circles ≥ 0.95/75, open circles ≥ 0.8/50.

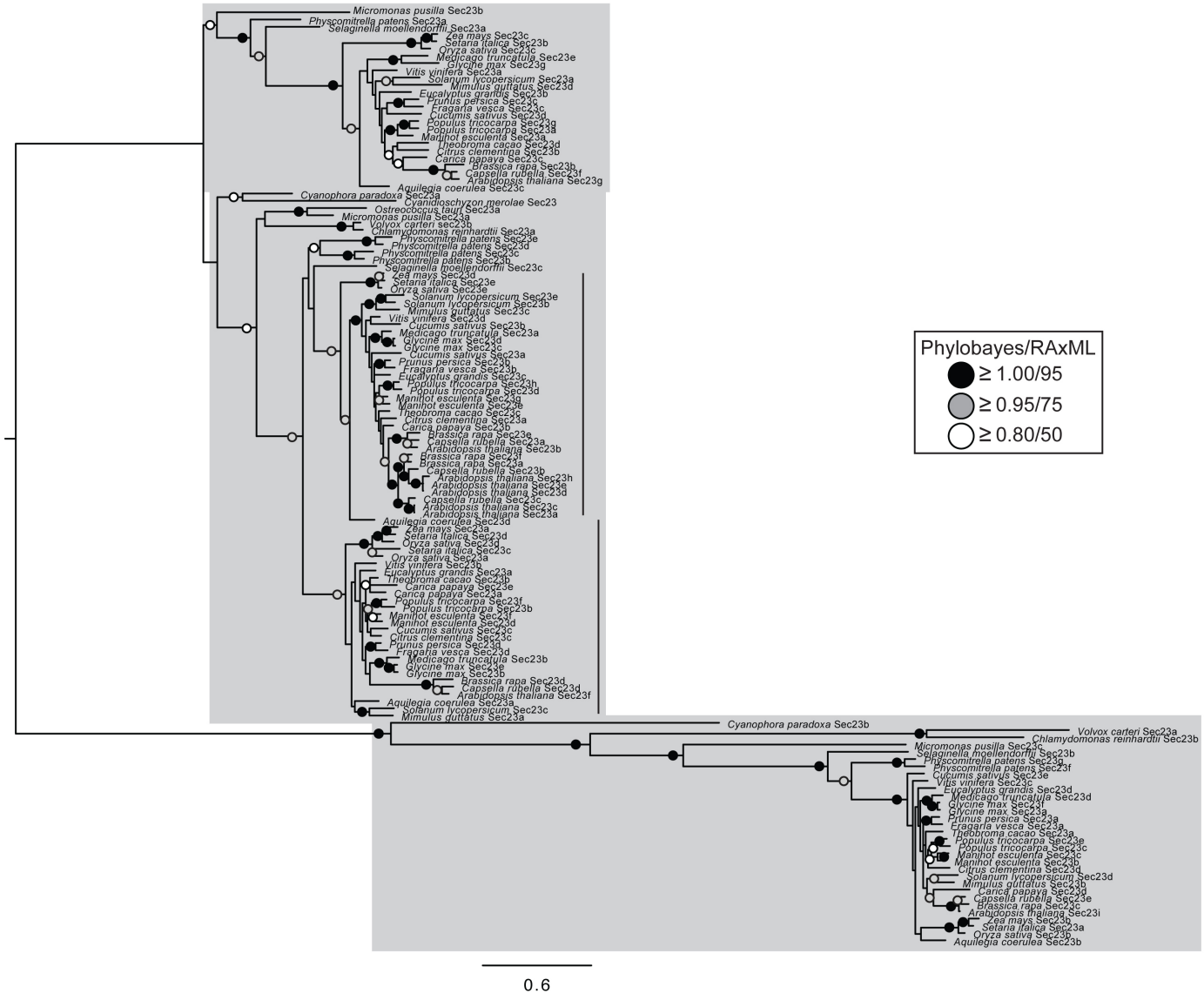

Micromonas pusilla Sec23b
Physcomitrella patens Sec23a
Selaginella moellendorffii Sec23a
Zea mays Sec23c
Setaria italica Sec23b
Oryza sativa Sec23c
Medicago truncatula Sec23e
Glycine max Sec23g
Vitis vinifera Sec23a
Solanum lycopersicum Sec23a
Mimulus guttatus Sec23d
Eucalyptus grandis Sec23b
Prunus persica Sec23c
Fragaria vesca Sec23c
Cucumis sativus Sec23d
Populus tricocarpa Sec23g
Populus tricocarpa Sec23a
Manihot esculenta Sec23a
Theobroma cacao Sec23d
Citrus clementina Sec23b
Carica papaya Sec23c
Brassica rapa Sec23h
Capsella rubella Sec23f
Arabidopsis thaliana Sec23g
Aquilegia coerulea Sec23c
Cyanophora paradoxa Sec23a
Cyanidioschyzon merolae Sec23
Ostreococcus tauri Sec23a
Micromonas pusilla Sec23a
Volvox carteri sec23b
Chlamydomonas reinhardtii Sec23a
Physcomitrella patens Sec23e
Physcomitrella patens Sec23d
Physcomitrella patens Sec23c
Physcomitrella patens Sec23b
Selaginella moellendorffii Sec23c
Zea mays Sec23d
Setaria italica Sec23e
Oryza sativa Sec23e
Solanum lycopersicum Sec23e
Solanum lycopersicum Sec23b
Mimulus guttatus Sec23c
Vitis vinifera Sec23d
Cucumis sativus Sec23b
Medicago truncatula Sec23a
Glycine max Sec23d
Glycine max Sec23c
Cucumis sativus Sec23a
Prunus persica Sec23b
Fragaria vesca Sec23b
Eucalyptus grandis Sec23c
Populus tricocarpa Sec23h
Populus tricocarpa Sec23d
Manihot esculenta Sec23g
Manihot esculenta Sec23e
Theobroma cacao Sec23c
Citrus clementina Sec23a
Carica papaya Sec23b
Brassica rapa Sec23e
Capsella rubella Sec23a
Arabidopsis thaliana Sec23b
Brassica rapa Sec23f
Brassica rapa Sec23a
Capsella rubella Sec23b
Arabidopsis thaliana Sec23h
Arabidopsis thaliana Sec23e
Arabidopsis thaliana Sec23d
Capsella rubella Sec23c
Arabidopsis thaliana Sec23a
Aquilegia coerulea Sec23d
Zea mays Sec23a
Setaria italica Sec23d
Oryza sativa Sec23g
Setaria italica Sec23c
Oryza sativa Sec23a
Vitis vinifera Sec23b
Eucalyptus grandis Sec23a
Theobroma cacao Sec23b
Carica papaya Sec23e
Carica papaya Sec23a
Populus tricocarpa Sec23f
Populus tricocarpa Sec23b
Manihot esculenta Sec23f
Manihot esculenta Sec23d
Cucumis sativus Sec23c
Citrus clementina Sec23c
Prunus persica Sec23d
Fragaria vesca Sec23d
Medicago truncatula Sec23b
Glycine max Sec23e
Glycine max Sec23b
Brassica rapa Sec23d
Capsella rubella Sec23d
Arabidopsis thaliana Sec23f
Aquilegia coerulea Sec23a
Solanum lycopersicum Sec23c
Mimulus guttatus Sec23a
Cyanophora paradoxa Sec23b
Volvox carteri Sec23a
Chlamydomonas reinhardtii Sec23b
Micromonas pusilla Sec23c
Selaginella moellendorffii Sec23b
Physcomitrella patens Sec23g
Physcomitrella patens Sec23f
Cucumis sativus Sec23e
Vitis vinifera Sec23c
Eucalyptus grandis Sec23d
Medicago truncatula Sec23d
Glycine max Sec23f
Glycine max Sec23a
Prunus persica Sec23a
Fragaria vesca Sec23a
Theobroma cacao Sec23a
Populus tricocarpa Sec23e
Populus tricocarpa Sec23c
Manihot esculenta Sec23c
Manihot esculenta Sec23b
Citrus clementina Sec23d
Solanum lycopersicum Sec23d
Mimulus guttatus Sec23b
Carica papaya Sec23d
Capsella rubella Sec23e
Brassica rapa Sec23c
Arabidopsis thaliana Sec23i
Zea mays Sec23b
Setaria italica Sec23a
Oryza sativa Sec23b
Aquilegia coerulea Sec23b
Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50
0.6

Figure 4-16. Phylogenetic analysis of archaeplastid Sec24 sequences identifies three paralogues in an early archaeplastid ancestor. In order to gain a better understanding of the relationship between the three Archaeplastid clades observed in Figure 4-9, an expanded analysis of plant Sec24 sequences was carried out. Three Sec24 clades were reconstructed, indicating that multiple gene duplications occurred in an early archaeplastid ancestor. These duplications likely occurred after the divergence of glaucophytes, because both *C. paradoxa* sequences group to the exclusion of the other Sec24 sequences. Two instances of lineage-specific expansion have occurred in addition to multiple occurrences of species-specific expansion within these larger clades (grey boxes). The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (Phylobayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95, closed light circles ≥ 0.95/75, open circles ≥ 0.8/50.

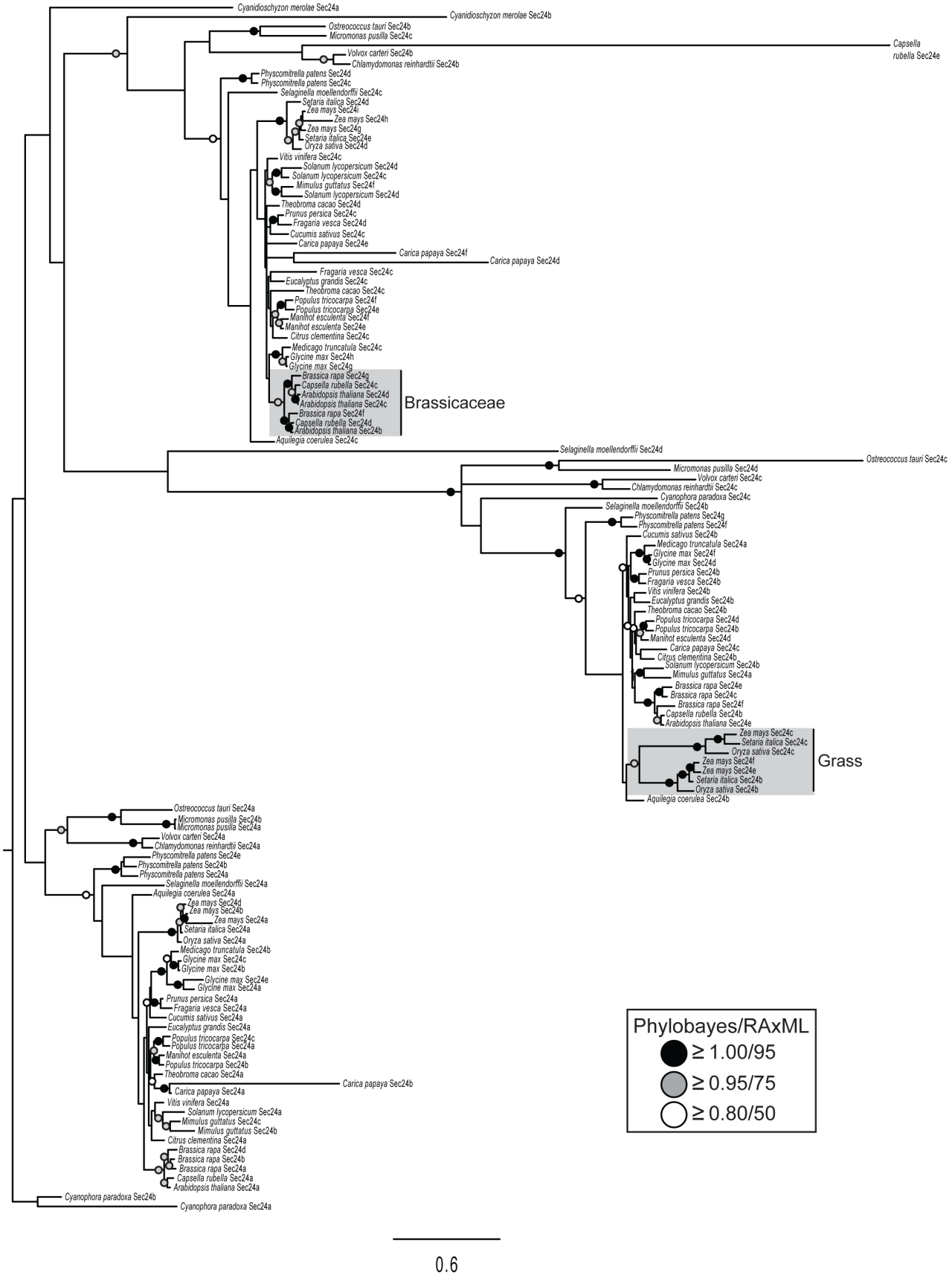

Cyanidioschyzon merolae Sec24a
Cyanidioschyzon merolae Sec24b
Ostreococcus tauri Sec24b
Micromonas pusilla Sec24c
Capsella rubella Sec24e
Volvox carteri Sec24b
Chlamydomonas reinhardtii Sec24b
Physcomitrella patens Sec24d
Physcomitrella patens Sec24c
Selaginella moellendorffii Sec24c
Setaria italica Sec24d
Zea mays Sec24i
Zea mays Sec24h
Zea mays Sec24g
Setaria italica Sec24e
Oryza sativa Sec24d
Vitis vinifera Sec24c
Solanum lycopersicum Sec24d
Solanum lycopersicum Sec24c
Mimulus guttatus Sec24f
Solanum lycopersicum Sec24d
Theobroma cacao Sec24d
Prunus persica Sec24c
Fragaria vesca Sec24d
Cucumis sativus Sec24c
Carica papaya Sec24e
Carica papaya Sec24f
Carica papaya Sec24d
Fragaria vesca Sec24c
Eucalyptus grandis Sec24c
Theobroma cacao Sec24c
Populus tricocarpa Sec24f
Populus tricocarpa Sec24e
Manihot esculenta Sec24f
Manihot esculenta Sec24e
Citrus clementina Sec24c
Medicago truncatula Sec24c
Glycine max Sec24h
Glycine max Sec24g
Brassica rapa Sec24g
Capsella rubella Sec24c
Arabidopsis thaliana Sec24d
Arabidopsis thaliana Sec24c
Brassica rapa Sec24f
Capsella rubella Sec24d
Arabidopsis thaliana Sec24b
Brassicaceae
Aquilegia coerulea Sec24c
Selaginella moellendorffii Sec24d
Ostreococcus tauri Sec24c
Micromonas pusilla Sec24d
Volvox carteri Sec24c
Chlamydomonas reinhardtii Sec24c
Cyanophora paradoxa Sec24c
Selaginella moellendorffii Sec24b
Physcomitrella patens Sec24g
Physcomitrella patens Sec24f
Cucumis sativus Sec24b
Medicago truncatula Sec24a
Glycine max Sec24f
Glycine max Sec24d
Prunus persica Sec24b
Fragaria vesca Sec24b
Vitis vinifera Sec24b
Eucalyptus grandis Sec24b
Theobroma cacao Sec24b
Populus tricocarpa Sec24d
Populus tricocarpa Sec24b
Manihot esculenta Sec24d
Carica papaya Sec24c
Citrus clementina Sec24b
Solanum lycopersicum Sec24b
Mimulus guttatus Sec24a
Brassica rapa Sec24e
Brassica rapa Sec24c
Brassica rapa Sec24f
Capsella rubella Sec24b
Arabidopsis thaliana Sec24e
Zea mays Sec24c
Setaria italica Sec24c
Oryza sativa Sec24c
Zea mays Sec24f
Zea mays Sec24e
Setaria italica Sec24b
Oryza sativa Sec24b
Grass
Aquilegia coerulea Sec24b
Ostreococcus tauri Sec24a
Micromonas pusilla Sec24b
Micromonas pusilla Sec24a
Volvox carteri Sec24a
Chlamydomonas reinhardtii Sec24a
Physcomitrella patens Sec24e
Physcomitrella patens Sec24b
Physcomitrella patens Sec24a
Selaginella moellendorffii Sec24a
Aquilegia coerulea Sec24a
Zea mays Sec24d
Zea mays Sec24b
Zea mays Sec24a
Setaria italica Sec24a
Oryza sativa Sec24a
Medicago truncatula Sec24b
Glycine max Sec24c
Glycine max Sec24b
Glycine max Sec24e
Glycine max Sec24a
Prunus persica Sec24a
Fragaria vesca Sec24a
Cucumis sativus Sec24a
Eucalyptus grandis Sec24a
Populus tricocarpa Sec24c
Populus tricocarpa Sec24a
Manihot esculenta Sec24a
Populus tricocarpa Sec24b
Theobroma cacao Sec24a
Carica papaya Sec24a
Carica papaya Sec24b
Vitis vinifera Sec24a
Solanum lycopersicum Sec24a
Mimulus guttatus Sec24c
Mimulus guttatus Sec24b
Citrus clementina Sec24a
Brassica rapa Sec24d
Brassica rapa Sec24b
Brassica rapa Sec24a
Capsella rubella Sec24a
Arabidopsis thaliana Sec24a
Cyanophora paradoxa Sec24b
Cyanophora paradoxa Sec24a
Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50
0.6

Figure 4-17. Phylogenetic analysis identifies independent expansions of Sec31 in embryophytes. To determine whether or not the expansion of archaeplastid Sec31 sequences is limited to *P. patens* and *A. thaliana*, an expanded analysis of plant Sec31 sequences was carried out. A single instance of lineage-specific expansion has occurred in the Brassicaceae (grey box). Multiple species-specific expansions within embryophytes were also reconstructed. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (Phylobayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95, closed light circles ≥ 0.95/75, open circles ≥ 0.8/50.


Cyanidioschyzon merolae Sec31
Ostreococcus tauri Sec31
Micromonas pusilla Sec31
Volvox carteri Sec31
Chlamydomonas reinhardtii Sec31
Selaginella moellendorffii Sec31
Physcomitrella patens Sec31b
Physcomitrella patens Sec31a
Zea mays Sec31b
Zea mays Sec31c
Zea mays Sec31a
Setaria italica Sec31
Oryza sativa Sec31
Aquilegia coerulea Sec31
Solanum lycopersicum Sec31
Mimulus guttatus Sec31
Vitis vinifera Sec31
Medicago truncatula Sec31
Glycine max Sec31b
Glycine max Sec31a
Prunus persica Sec31
Fragaria vesca Sec31a
Fragaria vesca Sec31b
Eucalyptus grandis Sec31
Cucumis sativus Sec31
Citrus clementina Sec31
Theobroma cacao Sec31
Populus tricocarpa Sec31b
Populus tricocarpa Sec31a
Manihot esculenta Sec31b
Manihot esculenta Sec31a
Carica papaya Sec31
Brassica rapa Sec31b
Capsella rubella Sec31b
Arabidopsis thaliana Sec31d
Brassica rapa Sec31a
Capsella rubella Sec31a
Arabidopsis thaliana Sec31c
Arabidopsis thaliana Sec31b
Arabidopsis thaliana Sec31a
Brassicaceae
Cyanophora paradoxa Sec31a
0.5
Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50

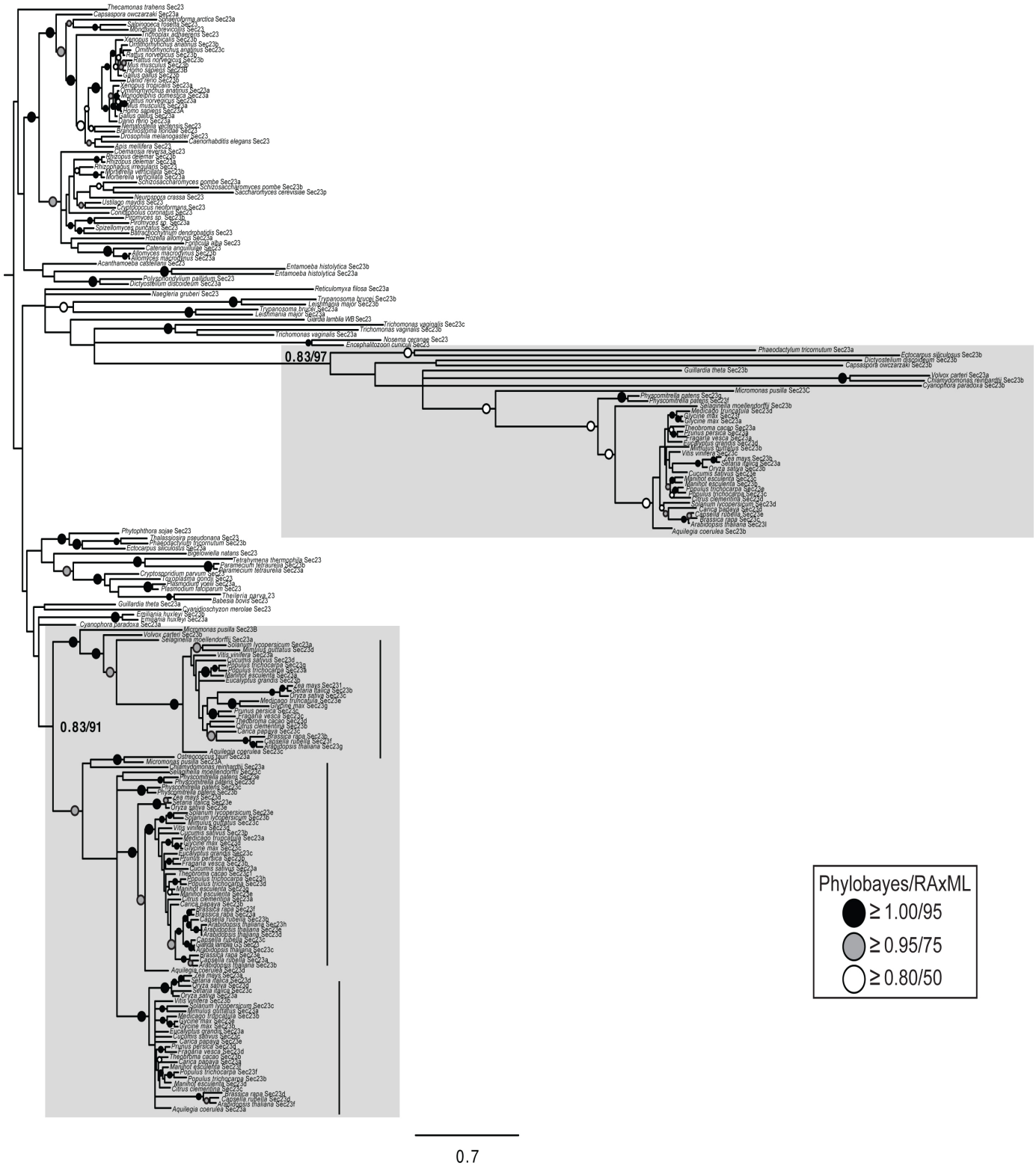
A potentially ancient duplication of Sec23 was also observed, possibly pre-dating the LECA (Figures 4-8, 4-15). To address this question, the pan-eukaryotic and expanded Archaeplastida datasets were combined and analyzed. Two distinct Sec23 clades from the Archaeplastida were uncovered, one of which is embedded in a group containing sequences from all other supergroups except the Excavata (Figure 4-18). Although all of the sequences that make up this clade are long branches, possibly contributing to phylogenetic artefact, at a minimum, the two archaeplastid clades are the product of an ancient duplication that predates the archaeplastid lineage.

4.4.3 Sed4 is a lineage-specific component present in a subset of the Saccharomycotina

In addition to Sec12, *S. cerevisiae* possesses an additional Sec12-like protein, Sed4. Originally identified as a multicopy suppressor of Δ*erd2* (encodes HDEL receptor; Hardwick et al., 1992), Sed4 is thought to aid in the recruitment of COPII components to the ER membrane by interacting with Sec16 (Gimeno et al., 1995) and act as a positive regulator of Sar1, likely by inhibiting the GTPase activity of Sec23 (Saito-Nakano and Nakano, 2000). Other analyses suggest that Sed4 possesses GAP activity and is able to stimulate GTP hydrolysis on Sar1 when Bet1 is not bound to Sar1, suggesting a method for aborting COPII vesicles with low cargo density (Kodera et al., 2011). As *S. cerevisiae* is a major model organism for the study of COPII function, it is important to address whether Sed4 is a general and ancient component of the COPII complex.

The initial survey did not identify any Sed4 orthologues, indicating that the

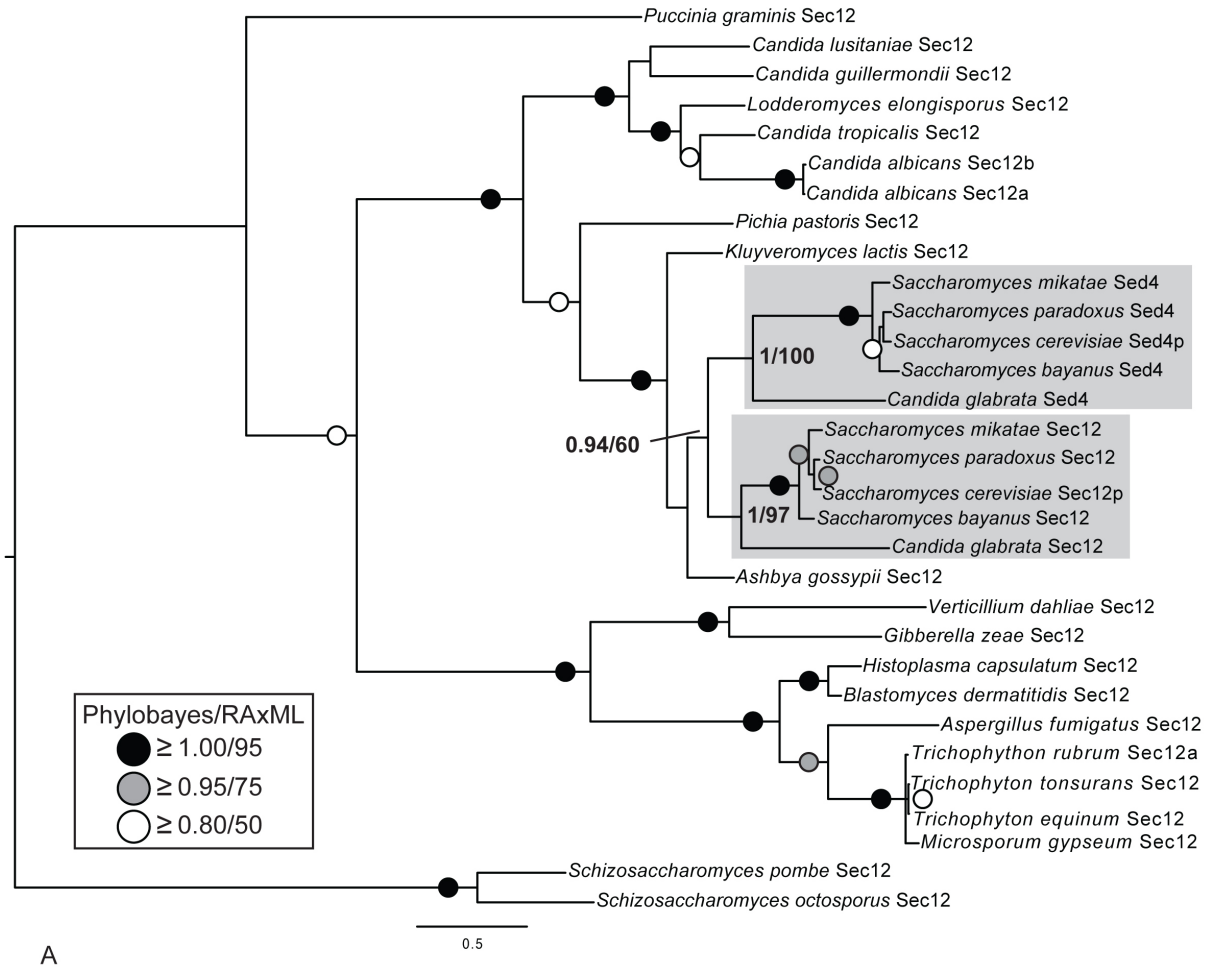
Figure 4-18. **Phylogenetic analysis identifies two ancient Sec23 paralogues in the ancestor of Archaeplastida.** To determine whether or not other eukaryotes possess orthologues of the Sec23 paralogues identified in Archaeplastida, the data sets from Figure 4-8 and Figure 4-15 were combined and analyzed. Although there is only moderate evidence to suggest the presence of a second ancient pan-eukaryotic Sec23 paralogue, the analysis confirms the presence of two Sec23 paralogues in the ancestor of Archaeplastida. Two major groups of archaeplastid sequences are visible (grey boxes). One has multiple expansions in land plants (lower grey box). The presence of *C. paradoxa* in, or near the base of each clade is suggestive that the duplication producing the two paralogues occurred prior to the differentiation of the archaeplastid lineages. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (Phylobayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95, closed light circles ≥ 0.95/75, open circles ≥ 0.8/50.


0.83/97
0.83/91
Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50
0.7

taxonomic distribution of this complex is limited compared to the other COPII subunits. In order to identify the origin and distribution of Sed4, the fungal taxon sampling was expanded. BLAST results suggested that some fungi had either Sec12 or Sed4, and very few species appeared to possess both. However, phylogenetic analysis revealed that most of these sequences are *bona fide* Sec12 orthologues, with Sed4 orthologues only present in *Saccharomyces bayanus, Saccharomyces paradoxus, Saccharomyces mikatae*, and *Candida glabrata,* suggesting that the gene duplication generating Sed4 occurred in the ancestor of the *Saccharomyces spp.* and *Candida glabrata* (Figure 4-19A).

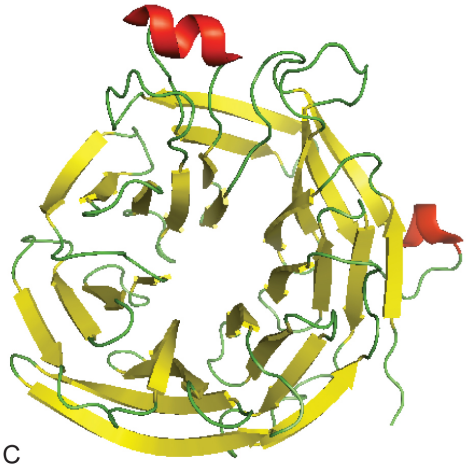
In order to gain some additional insight into the biology of Sed4, its tertiary structure was predicted. Homology modeling is generally reliable when the query and the primary sequence of a solved structure share at least 30% sequence identity (Kelley and Sternberg, 2009; Xiang, 2006). The high level of sequence identity between Sed4 and Sec12 (45%; Hardwick et al., 1992) combined with the recently solved structure of the cytosolic portion of Sec12 (McMahon et al., 2012) should provide a reliable structural prediction for Sed4. The Phyre2.0 server was used to model the structure of Sed4 (Kelley and Sternberg, 2009). As predicted, the structure of Sec12 was identified as the best homologue from which to model Sed4. Phyre2.0 modeled 32% of Sed4 (corresponding to the cytosolic portion) with 100% confidence. The low coverage is the result of homology limited to the cytosolic portions of Sec12 and Sed4, whereas, the extended luminal domain of Sed4 does not seem to share any sequence similarity to the luminal domain of Sec12. Although Sed4 has lost the ability to act as a Sar1 GEF (Saito-Nakano and Nakano, 2000), it

Figure 4-19. Sed4 is a β-propeller protein with limited taxonomic distribution within Fungi. Phylogenetic analysis shows that Sed4 is only present in a subset of the Saccharomycotina, and homology modelling predicts a tertiary structure similar to Sec12. A) Phylogenetic analysis of Sec12 and Sed4 sequences from representative fungal genomes shows that Sed4 is the product of a gene duplication in the ancestor of *Candida glabrata* and *Saccharomyces spp.,* not an ancient component of the COPII complex. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (Phylobayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95, closed light circles ≥ 0.95/75, open circles ≥ 0.8/50. B) Homology modeling of Sed4 using the Phyre2.0 server predicts that Sed4 is a β-propeller protein with a similar tertiary structure to Sec12. C) Crystal structure of the cytosolic portion of the *S. cerevisiae* Sec12 for comparison (PDB: 4H5I; McMahon et al*.,* 2012).



A

B



C

Sed4

Sec12 (PDB: 4H5I)

has retained significant structural similarity to Sec12 (Figure 4-19B, C). Sed4 is a β-propeller protein, as has been proposed previously (Chardin and Callebaut, 2002), and possesses a predicted K-loop, a short loop at the N-terminal propeller that binds a K^+ thought to be important for the interaction of Sec12 (Figure 4-19C) with Sar1, suggesting that Sed4 may interact with Sar1 by a similar mechanism. In sum, Sed4 is not an ancient, widespread COPII component, but rather is a recently added regulatory component found in a subset of fungi.

4.4.4 Multiple paralogues of Sec24 were present in the LECA

The phylogenetic analysis of the Sec24 tree (Figure 4-9) failed to show backbone resolution between potentially ancient paralogues. However, recurring clades were apparent (Holozoa, Fungi, SAR/CCTH, and Archaeplastida), suggesting that more than one paralogue of Sec24 may have been present in the LECA. To test this hypothesis, the Scrollsaw approach was used. Scrollsaw is a phylogenetic approach capable of gaining resolution between paralogues of large gene families where only short regions of sequence homology are available, by breaking large datasets into lineage-specific datasets and analyzed. Short branches are identified to act as surrogates for larger clades. These short branches are then combined in order to resolve the larger tree (Elias et al., 2012; Gabernet-Castello et al., 2013). The extensive structural (Bi et al., 2002) and sequence similarity (Yoshihisa et al., 1993) between Sec23 and Sec24 suggest a common ancestor for these two coat subunits (Tang et al., 1999). Therefore, Sec23 sequences were included in the analysis as an outgroup for Sec24. An alignment of all Sec24 and Sec23 sequences was

constructed, and was then broken into supergroup-specific datasets and analyzed (Figures 4-20 – 4-24).

Previous analyses of Sec24 have suggested ancient duplication events in the history of this component (Pagano et al., 1999; Tang et al., 1999); alignments and phylogenetic analyses showed that the human Sec24A, Sec24B, and *S. cerevisiae* Sec24p are more similar and group separately from the human Sec24C, Sec24D, and *S. cerevisiae* Sfb2 and Sfb3 sequences. This suggested that there were likely at least two paralogues of Sec24 in opisthokonts, and that these groups of paralogues represent the descendants of those lineages. In each phylogenetic analysis, the Sec24 sequences were largely resolved into two major clades. Based on the reciprocal best hit against the human genome, these corresponded to those that preferentially retrieved *H. sapiens* Sec24A and B, and those that retrieved Sec24C and D. To differentiate the two clades, they have been named: Sec24I, which corresponds to the group containing *H. sapiens* Sec24A and Sec24B, and Sec24II, which corresponds to the group containing *H. sapiens* Sec24C and Sec24D.

Next, the two shortest branches from each clade, including Sec23, were retained for use in a pan-eukaryotic phylogenetic analysis, with the selected sequences acting as surrogates for the rest of the supergroup (Figure 4-25). Most supergroups possessed two Sec24 clades along with additional unclassified Sec24 sequences. Rooting with Sec23 resulted in a paraphyletic Sec24I cluster, which was suspected to be a misplacement of the root. Removal of all Sec23 sequences from the analysis (Figure 4-26) clearly shows two distinct Sec24 clades.

Figure 4-20. Phylogenetic analysis identifies one Sec23 paralogue and two Sec24 paralogues in Opisthokonta. Phylogenetic analysis was carried out to determine the number of opisthokont Sec24 clades, and the shortest branches therein, for use in the Scrollsaw analysis. Two well-supported clades of Sec24 were reconstructed (grey boxes), labelled Sec24I which corresponds to the clade containing the human Sec24A and Sec24B sequences, and Sec24II which corresponds to the clade containing the human Sec24C and Sec24D sequences. A single strongly supported Sec23 clade was also reconstructed. Taxa used for Scrollsaw analysis are indicated in bold. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (Phylobayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95, closed light circles ≥ 0.95/75, open circles ≥ 0.8/50.

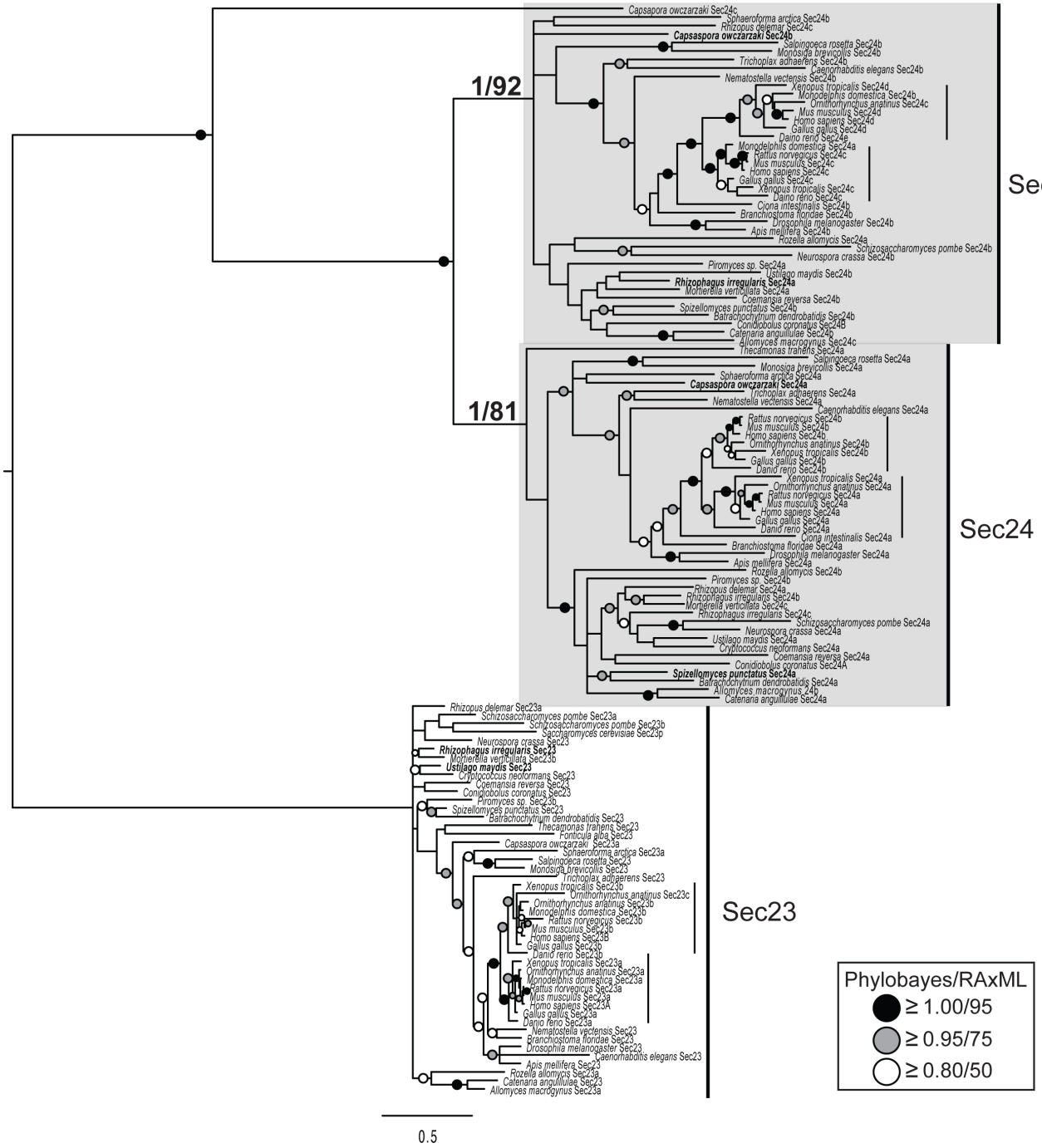

1/92
1/81
Sec24 II
Sec24 I
Sec23
Capsaspora owczarzaki Sec24c
Sphaeroforma arctica Sec24b
Rhizopus delemar Sec24c
Capsaspora owczarzaki Sec24b
Salpingoeca rosetta Sec24b
Monosiga brevicollis Sec24b
Trichoplax adhaerens Sec24b
Caenorhabditis elegans Sec24b
Nematostella vectensis Sec24b
Xenopus tropicalis Sec24d
Monodelphis domestica Sec24b
Ornithorhynchus anatinus Sec24c
Mus musculus Sec24d
Homo sapiens Sec24d
Gallus gallus Sec24d
Danio rerio Sec24e
Monodelphis domestica Sec24a
Rattus norvegicus Sec24c
Mus musculus Sec24c
Homo sapiens Sec24c
Gallus gallus Sec24c
Xenopus tropicalis Sec24c
Danio rerio Sec24c
Ciona intestinalis Sec24b
Branchiostoma floridae Sec24b
Drosophila melanogaster Sec24b
Apis mellifera Sec24b
Rozella allomycis Sec24a
Schizosaccharomyces pombe Sec24b
Neurospora crassa Sec24b
Piromyces sp. Sec24a
Ustilago maydis Sec24b
Rhizophagus irregularis Sec24a
Mortierella verticillata Sec24a
Coemansia reversa Sec24b
Spizellomyces punctatus Sec24b
Batrachochytrium dendrobatidis Sec24b
Conidiobolus coronatus Sec24B
Catenaria anguillulae Sec24b
Allomyces macrogynus Sec24c
Thecamonas trahens Sec24a
Salpingoeca rosetta Sec24a
Monosiga brevicollis Sec24a
Sphaeroforma arctica Sec24a
Capsaspora owczarzaki Sec24a
Trichoplax adhaerens Sec24a
Nematostella vectensis Sec24a
Caenorhabditis elegans Sec24a
Rattus norvegicus Sec24b
Mus musculus Sec24b
Homo sapiens Sec24b
Ornithorhynchus anatinus Sec24b
Xenopus tropicalis Sec24b
Gallus gallus Sec24b
Danio rerio Sec24b
Xenopus tropicalis Sec24a
Ornithorhynchus anatinus Sec24a
Rattus norvegicus Sec24a
Mus musculus Sec24a
Homo sapiens Sec24a
Gallus gallus Sec24a
Danio rerio Sec24a
Ciona intestinalis Sec24a
Branchiostoma floridae Sec24a
Drosophila melanogaster Sec24a
Apis mellifera Sec24a
Rozella allomycis Sec24b
Piromyces sp. Sec24b
Rhizopus delemar Sec24a
Rhizophagus irregularis Sec24b
Mortierella verticillata Sec24c
Rhizophagus irregularis Sec24c
Schizosaccharomyces pombe Sec24a
Neurospora crassa Sec24a
Ustilago maydis Sec24a
Cryptococcus neoformans Sec24a
Coemansia reversa Sec24a
Conidiobolus coronatus Sec24A
Spizellomyces punctatus Sec24a
Batrachochytrium dendrobatidis Sec24a
Allomyces macrogynus Sec24b
Catenaria anguillulae Sec24a
Rhizopus delemar Sec23a
Schizosaccharomyces pombe Sec23a
Schizosaccharomyces pombe Sec23b
Saccharomyces cerevisiae Sec23p
Neurospora crassa Sec23
Rhizophagus irregularis Sec23
Mortierella verticillata Sec23b
Ustilago maydis Sec23
Cryptococcus neoformans Sec23
Coemansia reversa Sec23
Coniglobolus coronatus Sec23
Piromyces sp. Sec23b
Spizellomyces punctatus Sec23
Batrachochytrium dendrobatidis Sec23
Thecamonas trahens Sec23
Fonticula alba Sec23
Capsaspora owczarzaki Sec23a
Sphaeroforma arctica Sec23a
Salpingoeca rosetta Sec23
Monosiga brevicollis Sec23
Trichoplax adhaerens Sec23
Xenopus tropicalis Sec23b
Ornithorhynchus anatinus Sec23c
Ornithorhynchus anatinus Sec23b
Monodelphis domestica Sec23b
Rattus norvegicus Sec23b
Mus musculus Sec23b
Homo sapiens Sec23B
Gallus gallus Sec23b
Danio rerio Sec23b
Xenopus tropicalis Sec23a
Ornithorhynchus anatinus Sec23a
Monodelphis domestica Sec23a
Rattus rlorvegicus Sec23a
Mus musculus Sec23a
Homo sapiens Sec23A
Gallus gallus Sec23a
Danio rerio Sec23a
Nematostella vectensis Sec23
Branchiostoma floridae Sec23
Drosophila melanogaster Sec23
Caenorhabditis elegans Sec23
Apis mellifera Sec23
Rozella allomycis Sec23a
Catenaria anguillulae Sec23
Allomyces macrogynus Sec23a
0.5
Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50

Figure 4-21. Phylogenetic analysis identifies one Sec23 paralogue and three Sec24 paralogues in Amoebozoa. Phylogenetic analysis was carried out to determine the number of Sec24 paralogues in the Amoebozoa and to identify the shortest branches for incorporation into the Scrollsaw analysis. Three clades of Sec24 were reconstructed (grey boxes). The clade labelled Sec24I corresponds to sequences that preferentially retrieved either the human Sec24A or Sec24B sequences during reciprocal BLAST analysis, while the clade labelled Sec24II corresponds to sequences that retrieved the human Sec24C and Sec24D sequences during reciprocal BLAST analysis. The clade labelled Sec24III contains sequences that retrieved a mix of the human Sec24A, Sec24B, Sec24C, and Sec24D sequences. Sec24I and Sec24III are strongly supported clades, whereas Sec24II only received moderate support. Only a single clade of Sec23 is recovered. Taxa used for Scrollsaw analysis are indicated in bold. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (Phylobayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95, closed light circles ≥ 0.95/75, open circles ≥ 0.8/50.

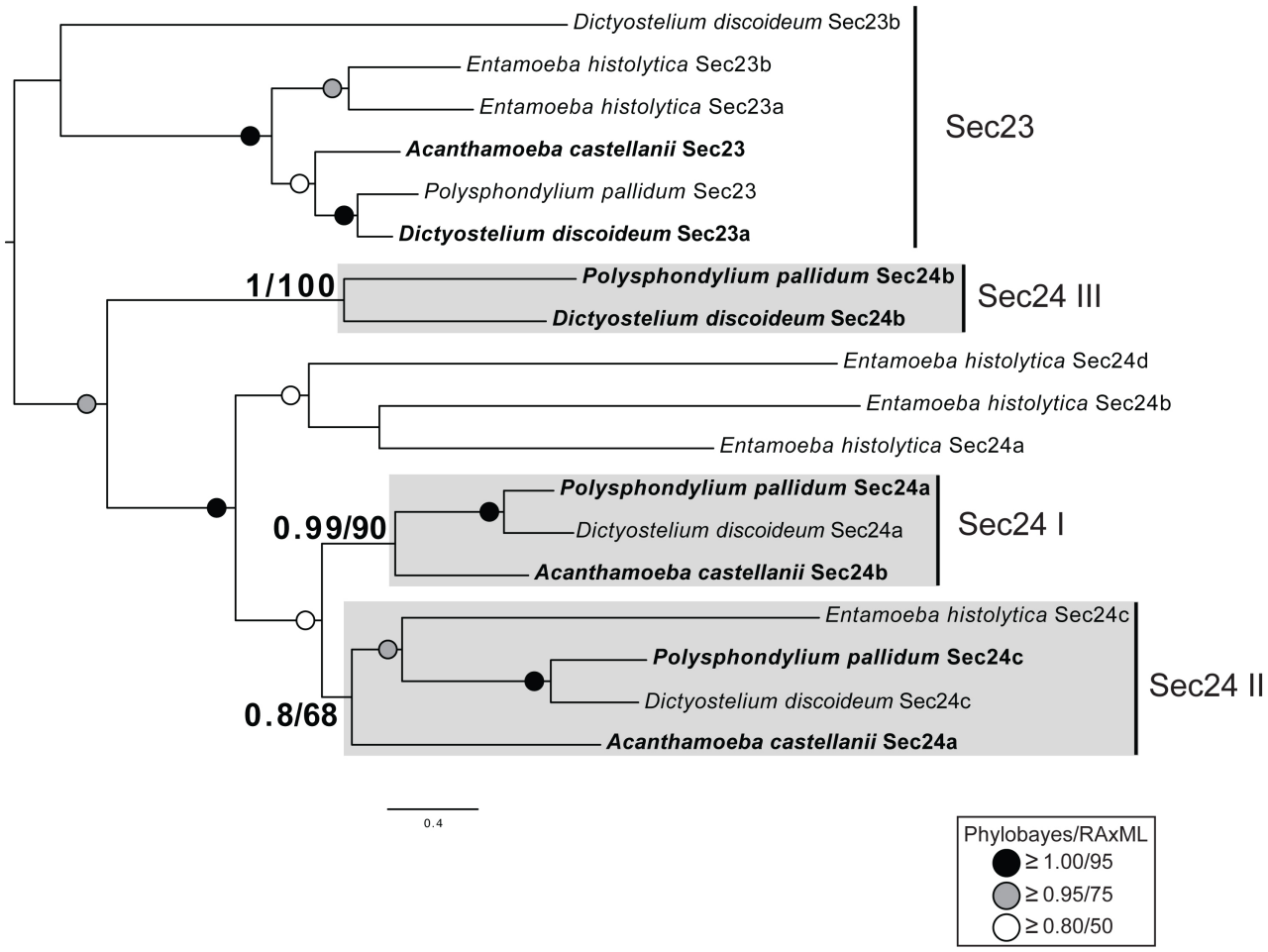

Dictyostelium discoideum Sec23b
Entamoeba histolytica Sec23b
Entamoeba histolytica Sec23a
Acanthamoeba castellanii Sec23
Polysphondylium pallidum Sec23
Dictyostelium discoideum Sec23a
Sec23
1/100
Polysphondylium pallidum Sec24b
Dictyostelium discoideum Sec24b
Sec24 III
Entamoeba histolytica Sec24d
Entamoeba histolytica Sec24b
Entamoeba histolytica Sec24a
0.99/90
Polysphondylium pallidum Sec24a
Dictyostelium discoideum Sec24a
Acanthamoeba castellanii Sec24b
Sec24 I
Entamoeba histolytica Sec24c
Polysphondylium pallidum Sec24c
Dictyostelium discoideum Sec24c
0.8/68
Acanthamoeba castellanii Sec24a
Sec24 II
0.4
Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50

Figure 4-22. Phylogenetic analysis identifies one Sec23 paralogue and two Sec24 paralogues in the Excavata. Phylogenetic analysis of excavate sequences was carried out to determine the number of Sec24 paralogues and to identify the shortest branches for inclusion in the Scrollsaw analysis. Two Sec24 clades are discernable (grey boxes). The *G. lamblia* Sec24 sequences remain unresolved; therefore, they were unable to be classified into one of the two paralogous clades. Sec23 forms a single strongly supported clade with an expansion in *T. brucei* and *L. major*. Taxa used for Scrollsaw analysis are indicated in bold. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (Phylobayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95, closed light circles ≥ 0.95/75, open circles ≥ 0.8/50.

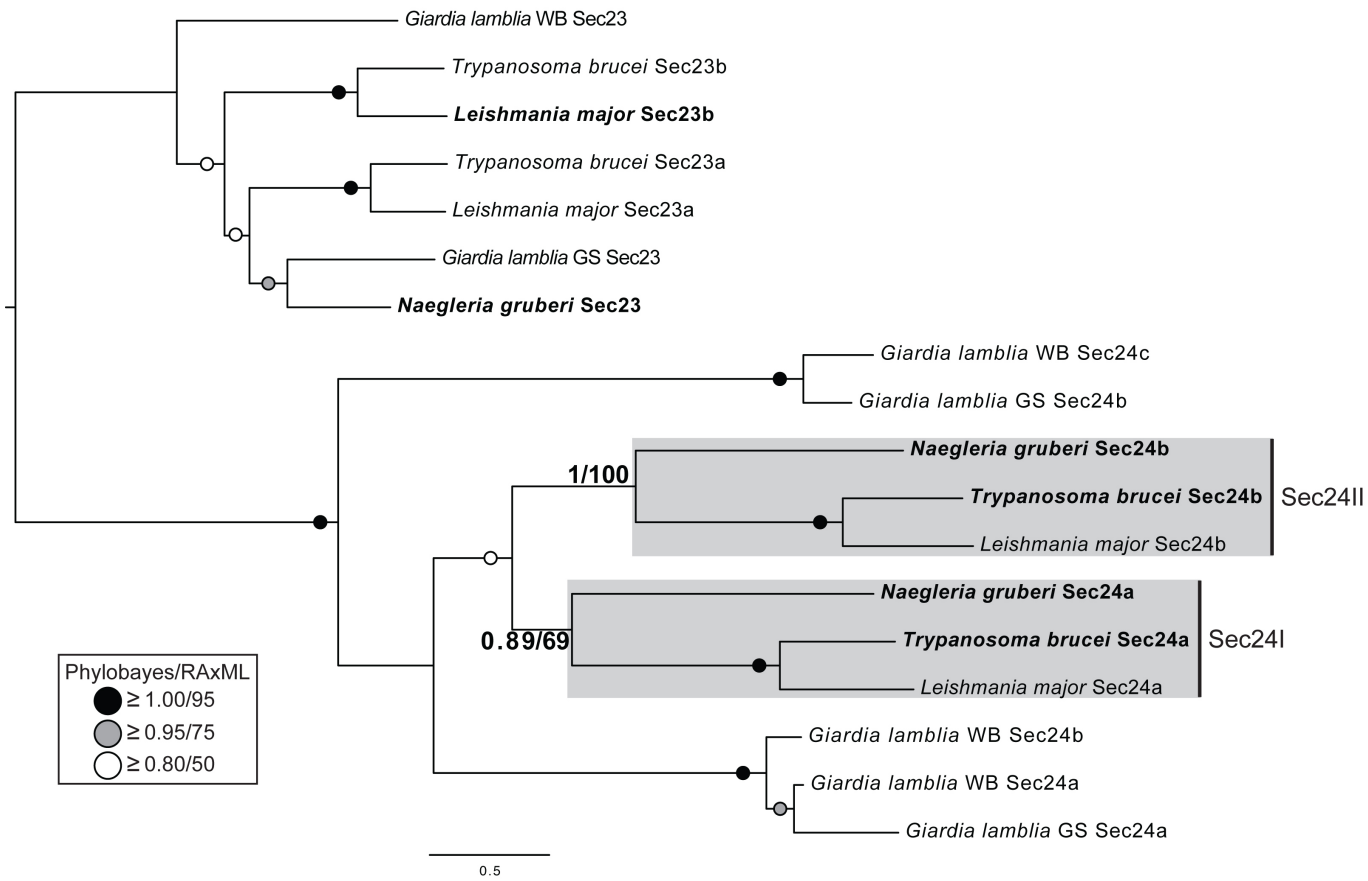

Giardia lamblia WB Sec23
Trypanosoma brucei Sec23b
Leishmania major Sec23b
Trypanosoma brucei Sec23a
Leishmania major Sec23a
Giardia lamblia GS Sec23
Naegleria gruberi Sec23
Giardia lamblia WB Sec24c
Giardia lamblia GS Sec24b
1/100
Naegleria gruberi Sec24b
Trypanosoma brucei Sec24b
Leishmania major Sec24b
Sec24II
0.89/69
Naegleria gruberi Sec24a
Trypanosoma brucei Sec24a
Leishmania major Sec24a
Sec24I
Giardia lamblia WB Sec24b
Giardia lamblia WB Sec24a
Giardia lamblia GS Sec24a
Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50
0.5

Figure 4-23. Phylogenetic analysis identifies two Sec23 paralogues and three Sec24 paralogues in the Archaeplastida. Phylogenetic analysis was carried out determine the number of Sec24 paralogues in the Archaeplastida and to identify the shortest branches in each clade for use in the Scrollsaw analysis. Sec24I and Sec24II are weakly recovered, while Sec24III is strongly supported (grey boxes). Two clades of Sec23 were also resolved, labelled Sec23I and Sec23II. Taxa used for Scrollsaw analysis are indicated in bold. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (Phylobayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95, closed light circles ≥ 0.95/75, open circles ≥ 0.8/50.

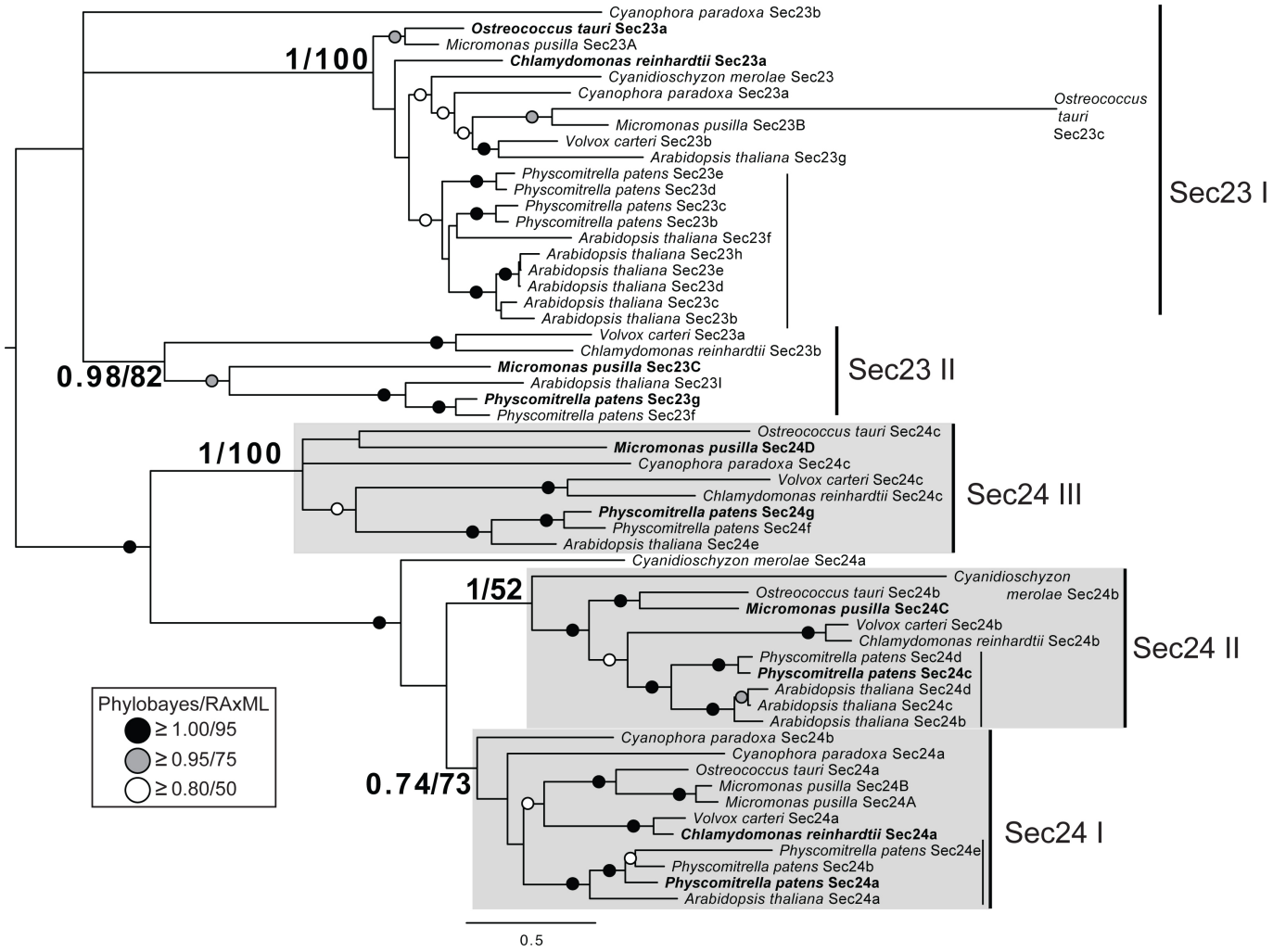

Cyanophora paradoxa Sec23b
Ostreococcus tauri Sec23a
Micromonas pusilla Sec23A
1/100
Chlamydomonas reinhardtii Sec23a
Cyanidioschyzon merolae Sec23
Cyanophora paradoxa Sec23a
Ostreococcus tauri Sec23c
Micromonas pusilla Sec23B
Volvox carteri Sec23b
Arabidopsis thaliana Sec23g
Physcomitrella patens Sec23e
Physcomitrella patens Sec23d
Physcomitrella patens Sec23c
Physcomitrella patens Sec23b
Arabidopsis thaliana Sec23f
Arabidopsis thaliana Sec23h
Arabidopsis thaliana Sec23e
Arabidopsis thaliana Sec23d
Arabidopsis thaliana Sec23c
Arabidopsis thaliana Sec23b
Sec23 I
Volvox carteri Sec23a
Chlamydomonas reinhardtii Sec23b
0.98/82
Micromonas pusilla Sec23C
Arabidopsis thaliana Sec23I
Physcomitrella patens Sec23g
Physcomitrella patens Sec23f
Sec23 II
Ostreococcus tauri Sec24c
1/100
Micromonas pusilla Sec24D
Cyanophora paradoxa Sec24c
Volvox carteri Sec24c
Chlamydomonas reinhardtii Sec24c
Physcomitrella patens Sec24g
Physcomitrella patens Sec24f
Arabidopsis thaliana Sec24e
Sec24 III
Cyanidioschyzon merolae Sec24a
Cyanidioschyzon merolae Sec24b
1/52
Ostreococcus tauri Sec24b
Micromonas pusilla Sec24C
Volvox carteri Sec24b
Chlamydomonas reinhardtii Sec24b
Physcomitrella patens Sec24d
Physcomitrella patens Sec24c
Arabidopsis thaliana Sec24d
Arabidopsis thaliana Sec24c
Arabidopsis thaliana Sec24b
Sec24 II
Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50
Cyanophora paradoxa Sec24b
Cyanophora paradoxa Sec24a
0.74/73
Ostreococcus tauri Sec24a
Micromonas pusilla Sec24B
Micromonas pusilla Sec24A
Volvox carteri Sec24a
Chlamydomonas reinhardtii Sec24a
Physcomitrella patens Sec24e
Physcomitrella patens Sec24b
Physcomitrella patens Sec24a
Arabidopsis thaliana Sec24a
Sec24 I
0.5

Figure 4-24. Phylogenetic analysis identifies one Sec23 paralogue and three Sec24 paralogues in the SAR and CCTH. Phylogenetic analysis was carried out to determine the number of Sec24 paralogues in this supergroup and to identify the shortest branches in each clade. Sec24I was moderately supported while Sec24III was weakly recovered. Sec24II is paraphyletic. One clade of Sec23 was recovered. All clades were weakly to moderately supported. Taxa used for Scrollsaw analysis are indicated in bold. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (Phylobayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95, closed light circles ≥ 0.95/75, open circles ≥ 0.8/50.

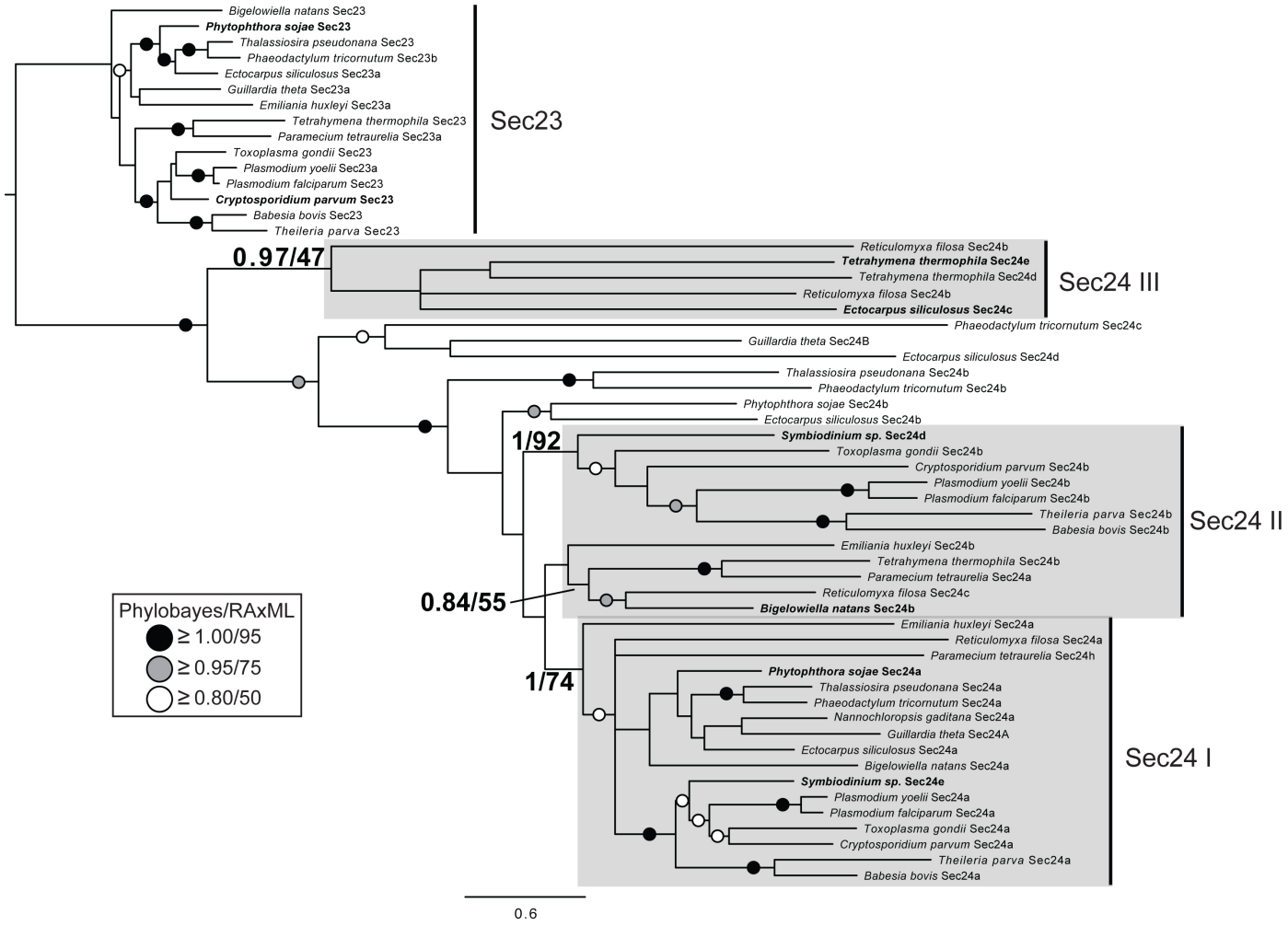

Bigelowiella natans Sec23
Phytophthora sojae Sec23
Thalassiosira pseudonana Sec23
Phaeodactylum tricornutum Sec23b
Ectocarpus siliculosus Sec23a
Guillardia theta Sec23a
Emiliania huxleyi Sec23a
Tetrahymena thermophila Sec23
Paramecium tetraurelia Sec23a
Toxoplasma gondii Sec23
Plasmodium yoelii Sec23a
Plasmodium falciparum Sec23
Cryptosporidium parvum Sec23
Babesia bovis Sec23
Theileria parva Sec23
Sec23
0.97/47
Reticulomyxa filosa Sec24b
Tetrahymena thermophila Sec24e
Tetrahymena thermophila Sec24d
Reticulomyxa filosa Sec24b
Ectocarpus siliculosus Sec24c
Sec24 III
Phaeodactylum tricornutum Sec24c
Guillardia theta Sec24B
Ectocarpus siliculosus Sec24d
Thalassiosira pseudonana Sec24b
Phaeodactylum tricornutum Sec24b
Phytophthora sojae Sec24b
Ectocarpus siliculosus Sec24b
1/92
Symbiodinium sp. Sec24d
Toxoplasma gondii Sec24b
Cryptosporidium parvum Sec24b
Plasmodium yoelii Sec24b
Plasmodium falciparum Sec24b
Theileria parva Sec24b
Babesia bovis Sec24b
Sec24 II
Emiliania huxleyi Sec24b
Tetrahymena thermophila Sec24b
Paramecium tetraurelia Sec24a
0.84/55
Reticulomyxa filosa Sec24c
Bigelowiella natans Sec24b
Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50
Emiliania huxleyi Sec24a
Reticulomyxa filosa Sec24a
Paramecium tetraurelia Sec24h
1/74
Phytophthora sojae Sec24a
Thalassiosira pseudonana Sec24a
Phaeodactylum tricornutum Sec24a
Nannochloropsis gaditana Sec24a
Guillardia theta Sec24A
Ectocarpus siliculosus Sec24a
Bigelowiella natans Sec24a
Symbiodinium sp. Sec24e
Plasmodium yoelii Sec24a
Plasmodium falciparum Sec24a
Toxoplasma gondii Sec24a
Cryptosporidium parvum Sec24a
Theileria parva Sec24a
Babesia bovis Sec24a
Sec24 I
0.6

Figure 4-25. Round one of Scrollsaw analysis identifies one Sec23 and at least two Sec24 paralogues in the LECA. To reconstruct the number of Sec24 paralogues in the LECA, surrogate (bolded) taxa were incorporated into a single data set for analysis. One clade of Sec23 is reconstructed. At least one paralogue of Sec24 was present in the LECA, as Sec24II is reconstructed with moderate support whereas Sec24I is paraphyletic. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (Phylobayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95, closed light circles ≥ 0.95/75, open circles ≥ 0.8/50.

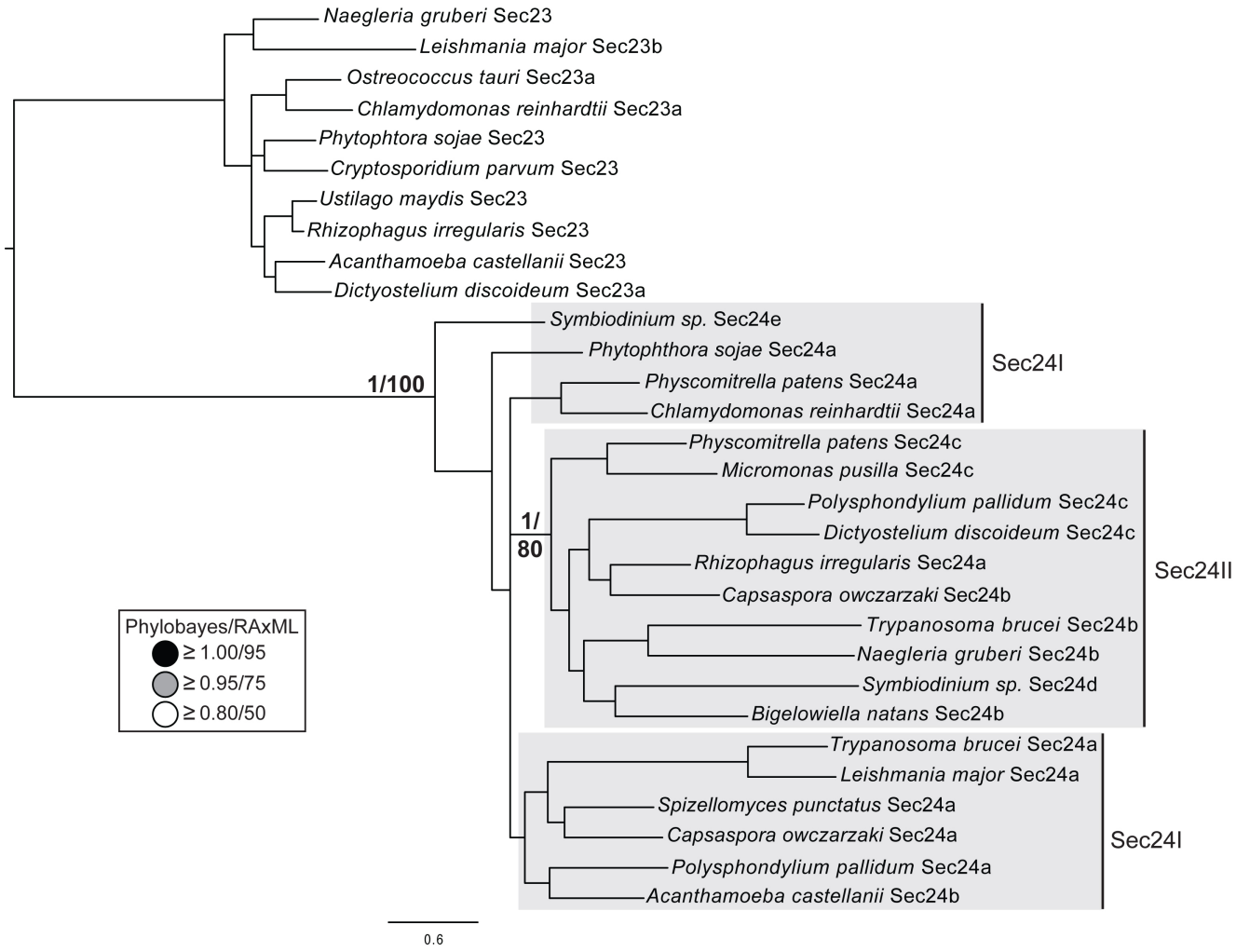
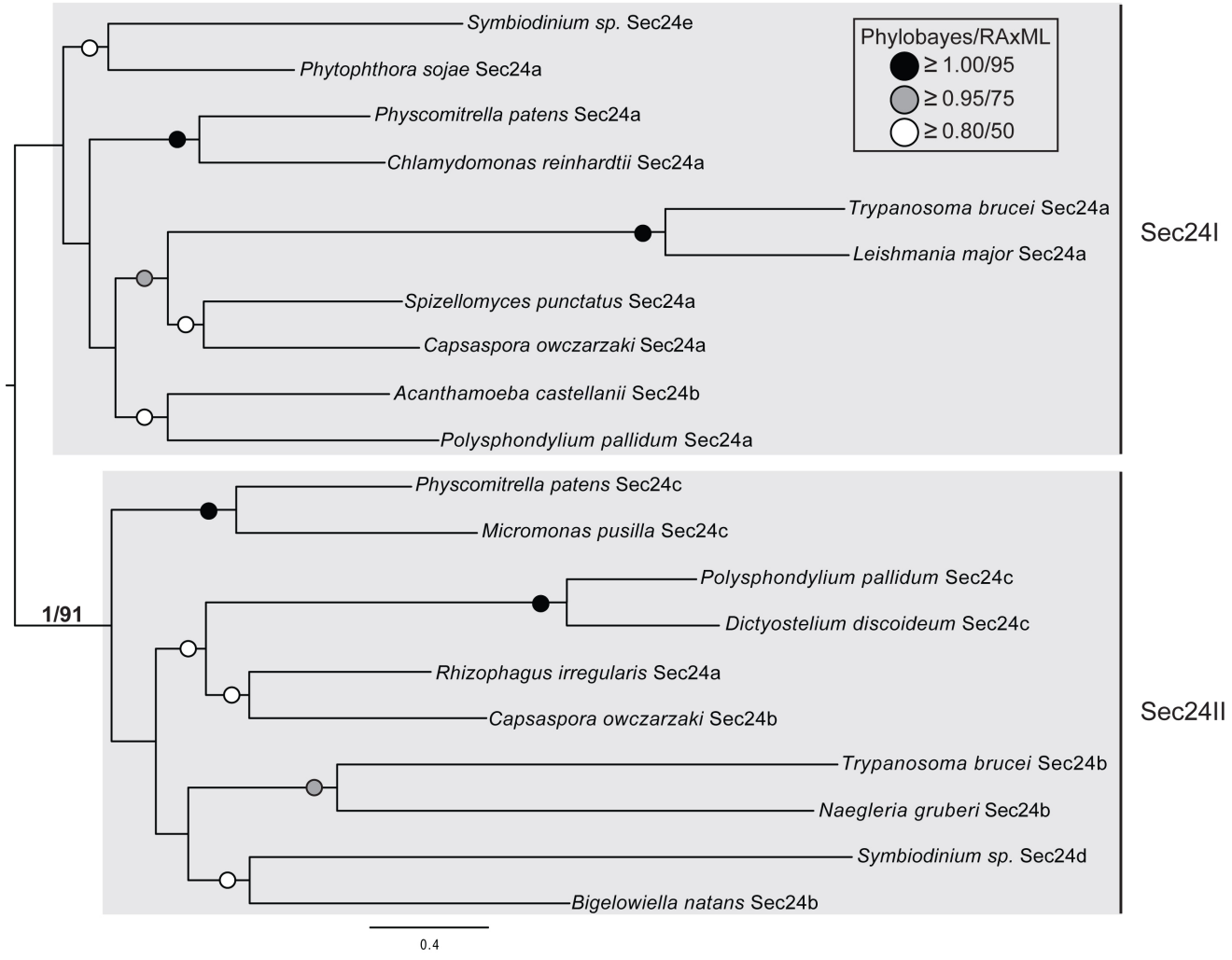

Naegleria gruberi Sec23
Leishmania major Sec23b
Ostreococcus tauri Sec23a
Chlamydomonas reinhardtii Sec23a
Phytophtora sojae Sec23
Cryptosporidium parvum Sec23
Ustilago maydis Sec23
Rhizophagus irregularis Sec23
Acanthamoeba castellanii Sec23
Dictyostelium discoideum Sec23a
Symbiodinium sp. Sec24e
Phytophthora sojae Sec24a
Physcomitrella patens Sec24a
Chlamydomonas reinhardtii Sec24a
Sec24I
1/100
Physcomitrella patens Sec24c
Micromonas pusilla Sec24c
Polysphondylium pallidum Sec24c
Dictyostelium discoideum Sec24c
1/
80
Rhizophagus irregularis Sec24a
Capsaspora owczarzaki Sec24b
Trypanosoma brucei Sec24b
Naegleria gruberi Sec24b
Symbiodinium sp. Sec24d
Bigelowiella natans Sec24b
Sec24II
Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50
Trypanosoma brucei Sec24a
Leishmania major Sec24a
Spizellomyces punctatus Sec24a
Capsaspora owczarzaki Sec24a
Polysphondylium pallidum Sec24a
Acanthamoeba castellanii Sec24b
Sec24I
0.6

Figure 4-26. Round two of Scrollsaw identifies two clades of Sec24 in the LECA. To reduce any phylogenetic artefact contributing to the paraphyly of Sec24I, all Sec23 sequences were removed from the analysis. Both Sec24I and Sec24II are reconstructed as strongly supported groups. This result indicates that at least two Sec24 paralogues were present in the LECA. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (Phylobayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95, closed light circles ≥ 0.95/75, open circles ≥ 0.8/50.


Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50
Symbiodinium sp. Sec24e
Phytophthora sojae Sec24a
Physcomitrella patens Sec24a
Chlamydomonas reinhardtii Sec24a
Trypanosoma brucei Sec24a
Leishmania major Sec24a
Spizellomyces punctatus Sec24a
Capsaspora owczarzaki Sec24a
Acanthamoeba castellanii Sec24b
Polysphondylium pallidum Sec24a
Sec24I
Physcomitrella patens Sec24c
Micromonas pusilla Sec24c
Polysphondylium pallidum Sec24c
Dictyostelium discoideum Sec24c
1/91
Rhizophagus irregularis Sec24a
Capsaspora owczarzaki Sec24b
Trypanosoma brucei Sec24b
Naegleria gruberi Sec24b
Symbiodinium sp. Sec24d
Bigelowiella natans Sec24b
Sec24II
0.4

Upon analyzing the supergroup-specific dataset for the Archaeplastida, a third clade of Sec24 sequences that did not correspond to one of the paralogous sets found in opisthokonts became apparent. Each of the other supergroup-specific datasets was re-examined (Figures 4-20 – 4-24) to determine if they also possessed extra clades of Sec24 paralogues. Additional clades were identified in both the SAR/CCTH and Amoebozoa datasets. Results from BLAST searches indicate that these paralogues retrieve each other as top BLAST hits rather than paralogues from Sec24I or Sec24II, suggesting that these sequences may represent a third ancient Sec24 paralogue. To confirm that these sequences do not represent multiple convergent, lineage-specific expansions, a phylogenetic analysis of Sec24 sequences from only these taxa was carried out (Figure 4-27). These sequences formed a group to the exclusion of all other Sec24s, indicating that they represent a third paralogue. This clade was labelled Sec24III. Given that the taxa in this clade span the diversity of the eukaryotic tree, this group likely represents yet another Sec24 paralogue that was present in the LECA.

The two shortest branches from each Sec24III clade of each supergroup were added into the Scrollsaw dataset. This analysis recovered a weakly supported Sec24II clade and paraphyletic Sec24I and Sec24III clades (Figure 4-28). To avoid any impact of LBA resulting from large evolutionary distances between Sec23 and Sec24, all Sec23 sequences were removed from the analysis (Figure 4-29). In doing so, a moderately supported Sec24III clade was recovered, but with no resolution between Sec24I and Sec24II. This analysis confirmed the presence of three Sec24 paralogues present in the LECA. However, the lack of backbone support in outgroup

Figure 4-27. Phylogenetic analysis of archaeplastid, amoebozoan, and SAR/CCTH Sec24 sequences identifies a third ancient Sec24 paralogue. To determine whether the additional Sec24III clades identified in the Amoebozoa, Archaeplastida, and SAR/CCTH represent a third ancient Sec24 paralogue or are independent lineage-specific expansions, a phylogenetic analysis combining the Sec24 sequences from these supergroups was carried out. The analysis reconstructed a single clade containing all of the extra Sec24 sequences, labelled Sec24III, suggesting that a third Sec24 paralogue was present in the LECA. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (Phylobayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95, closed light circles ≥ 0.95/75, open circles ≥ 0.8/50.

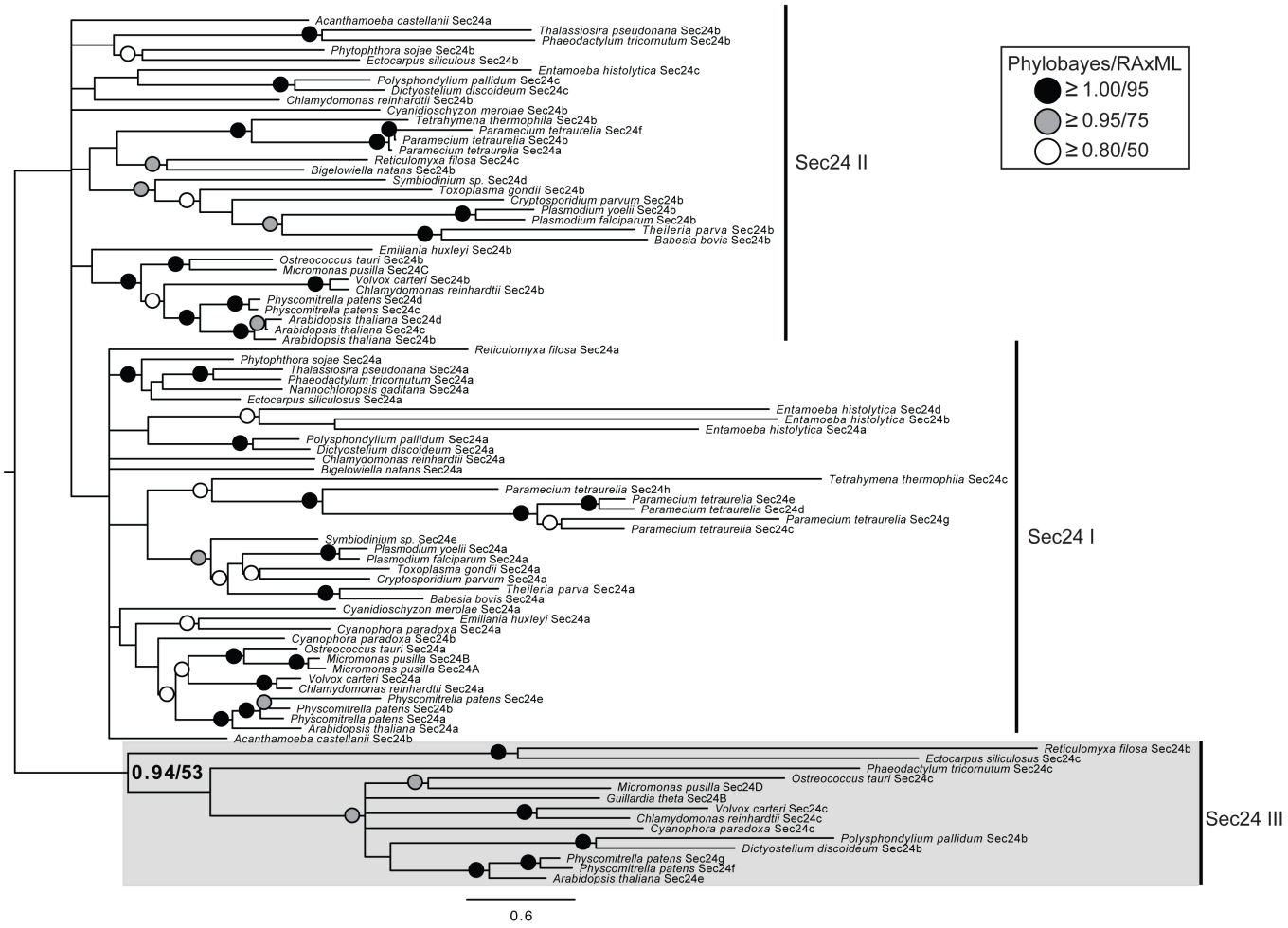

Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50
Sec24 II
Sec24 I
Sec24 III
0.94/53
0.6
Acanthamoeba castellanii Sec24a
Thalassiosira pseudonana Sec24b
Phaeodactylum tricornutum Sec24b
Phytophthora sojae Sec24b
Ectocarpus siliculous Sec24b
Entamoeba histolytica Sec24c
Polysphondylium pallidum Sec24c
Dictyostelium discoideum Sec24c
Chlamydomonas reinhardtii Sec24b
Cyanidioschyzon merolae Sec24b
Tetrahymena thermophila Sec24b
Paramecium tetraurelia Sec24f
Paramecium tetraurelia Sec24b
Paramecium tetraurelia Sec24a
Reticulomyxa filosa Sec24c
Bigelowiella natans Sec24b
Symbiodinium sp. Sec24d
Toxoplasma gondii Sec24b
Cryptosporidium parvum Sec24b
Plasmodium yoelii Sec24b
Plasmodium falciparum Sec24b
Theileria parva Sec24b
Babesia bovis Sec24b
Emiliania huxleyi Sec24b
Ostreococcus tauri Sec24b
Micromonas pusilla Sec24C
Volvox carteri Sec24b
Chlamydomonas reinhardtii Sec24b
Physcomitrella patens Sec24d
Physcomitrella patens Sec24c
Arabidopsis thaliana Sec24d
Arabidopsis thaliana Sec24c
Arabidopsis thaliana Sec24b
Reticulomyxa filosa Sec24a
Phytophthora sojae Sec24a
Thalassiosira pseudonana Sec24a
Phaeodactylum tricornutum Sec24a
Nannochloropsis gaditana Sec24a
Ectocarpus siliculosus Sec24a
Entamoeba histolytica Sec24d
Entamoeba histolytica Sec24b
Entamoeba histolytica Sec24a
Polysphondylium pallidum Sec24a
Dictyostelium discoideum Sec24a
Chlamydomonas reinhardtii Sec24a
Bigelowiella natans Sec24a
Tetrahymena thermophila Sec24c
Paramecium tetraurelia Sec24h
Paramecium tetraurelia Sec24e
Paramecium tetraurelia Sec24d
Paramecium tetraurelia Sec24g
Paramecium tetraurelia Sec24c
Symbiodinium sp. Sec24e
Plasmodium yoelii Sec24a
Plasmodium falciparum Sec24a
Toxoplasma gondii Sec24a
Cryptosporidium parvum Sec24a
Theileria parva Sec24a
Babesia bovis Sec24a
Cyanidioschyzon merolae Sec24a
Emiliania huxleyi Sec24a
Cyanophora paradoxa Sec24a
Cyanophora paradoxa Sec24b
Ostreococcus tauri Sec24a
Micromonas pusilla Sec24B
Micromonas pusilla Sec24A
Volvox carteri Sec24a
Chlamydomonas reinhardtii Sec24a
Physcomitrella patens Sec24e
Physcomitrella patens Sec24b
Physcomitrella patens Sec24a
Arabidopsis thaliana Sec24a
Acanthamoeba castellanii Sec24b
Reticulomyxa filosa Sec24b
Ectocarpus siliculosus Sec24c
Phaeodactylum tricornutum Sec24c
Ostreococcus tauri Sec24c
Micromonas pusilla Sec24D
Guillardia theta Sec24B
Volvox carteri Sec24c
Chlamydomonas reinhardtii Sec24c
Cyanophora paradoxa Sec24c
Polysphondylium pallidum Sec24b
Dictyostelium discoideum Sec24b
Physcomitrella patens Sec24g
Physcomitrella patens Sec24f
Arabidopsis thaliana Sec24e

Figure 4-28. Round three of Scrollsaw identifies one Sec23 paralogue and multiple Sec24 paralogues. In order to determine the relationship between the three ancient Sec24 paralogues the Sec24III sequences were incorporated into the original Scrollsaw data set (Figure 4-25). Although Sec24III is paraphyletic, it diverges before Sec24I and Sec24II. Sec24I is also paraphyletic but groups with Sec24II to the exclusion of Sec24III. Only Sec24II is recovered as a single supported clade. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (Phylobayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95, closed light circles ≥ 0.95/75, open circles ≥ 0.8/50.

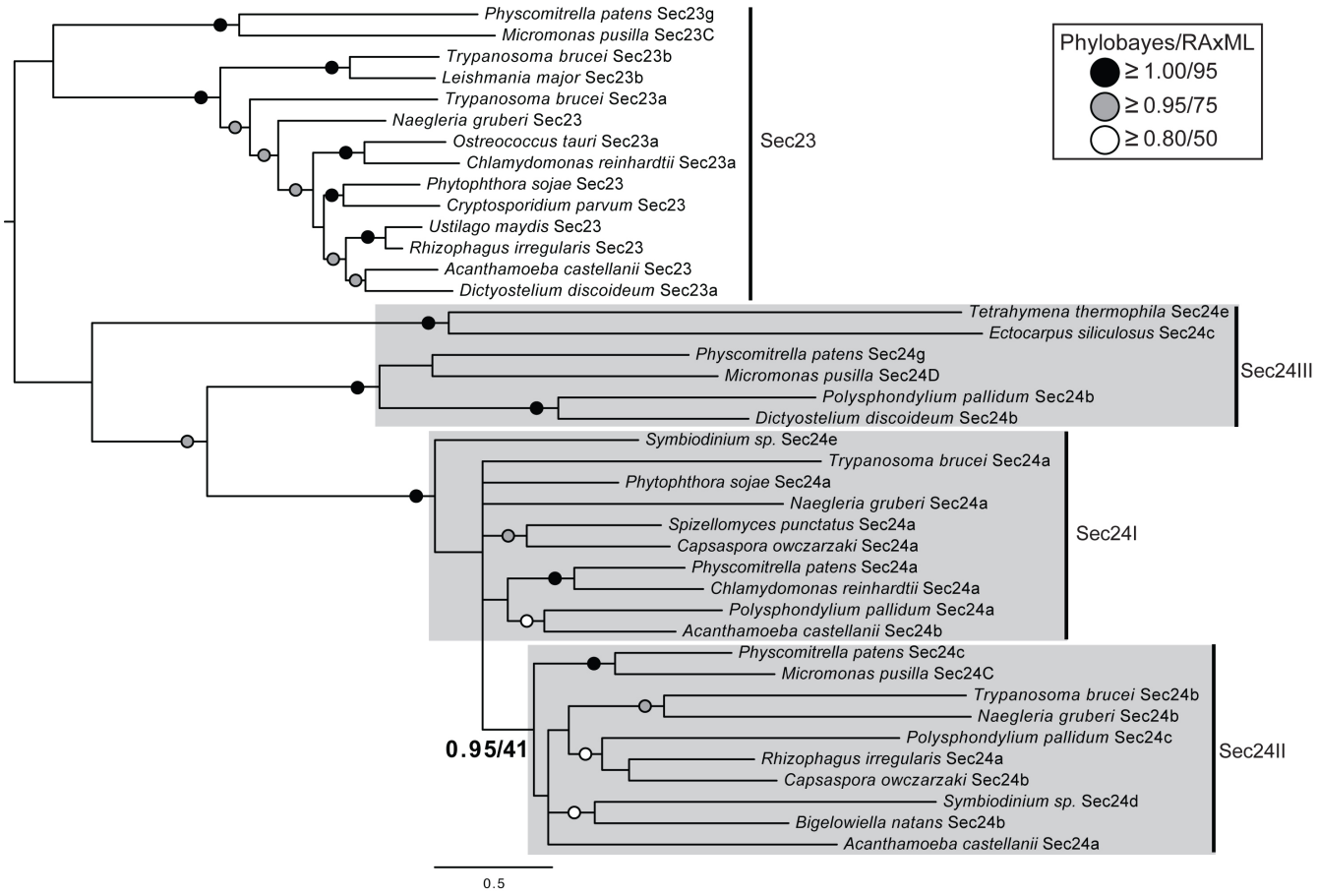
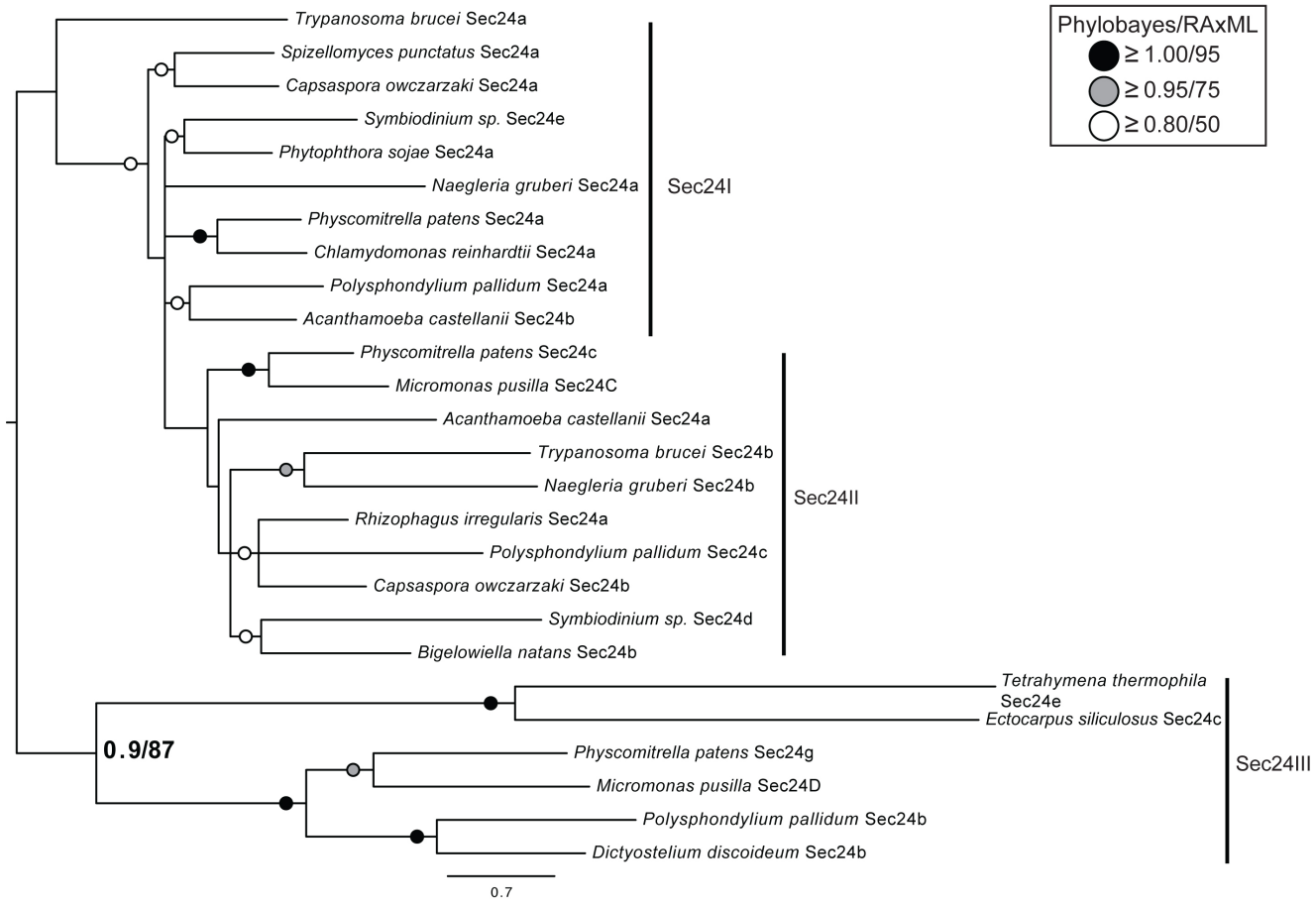

Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50
Physcomitrella patens Sec23g
Micromonas pusilla Sec23C
Trypanosoma brucei Sec23b
Leishmania major Sec23b
Trypanosoma brucei Sec23a
Naegleria gruberi Sec23
Ostreococcus tauri Sec23a
Chlamydomonas reinhardtii Sec23a
Phytophthora sojae Sec23
Cryptosporidium parvum Sec23
Ustilago maydis Sec23
Rhizophagus irregularis Sec23
Acanthamoeba castellanii Sec23
Dictyostelium discoideum Sec23a
Sec23
Tetrahymena thermophila Sec24e
Ectocarpus siliculosus Sec24c
Physcomitrella patens Sec24g
Micromonas pusilla Sec24D
Polysphondylium pallidum Sec24b
Dictyostelium discoideum Sec24b
Sec24III
Symbiodinium sp. Sec24e
Trypanosoma brucei Sec24a
Phytophthora sojae Sec24a
Naegleria gruberi Sec24a
Spizellomyces punctatus Sec24a
Capsaspora owczarzaki Sec24a
Physcomitrella patens Sec24a
Chlamydomonas reinhardtii Sec24a
Polysphondylium pallidum Sec24a
Acanthamoeba castellanii Sec24b
Sec24I
Physcomitrella patens Sec24c
Micromonas pusilla Sec24C
Trypanosoma brucei Sec24b
Naegleria gruberi Sec24b
0.95/41
Polysphondylium pallidum Sec24c
Rhizophagus irregularis Sec24a
Capsaspora owczarzaki Sec24b
Symbiodinium sp. Sec24d
Bigelowiella natans Sec24b
Acanthamoeba castellanii Sec24a
Sec24II
0.5

Figure 4-29. Round four of Scrollsaw shows that Sec24III forms a distinct clade from Sec24I and Sec24II. In order to remove any phylogenetic artefact and to increase the signal-to-noise ratio in the data set, all Sec23 sequences were removed. Re-analyzing the data set resulted in the reconstruction of a monophyletic Sec24III clade. Sec24II was reconstructed, but unsupported, and Sec24I was paraphyletic. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (Phylobayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95, closed light circles ≥ 0.95/75, open circles ≥ 0.8/50.


Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50
Trypanosoma brucei Sec24a
Spizellomyces punctatus Sec24a
Capsaspora owczarzaki Sec24a
Symbiodinium sp. Sec24e
Phytophthora sojae Sec24a
Naegleria gruberi Sec24a
Physcomitrella patens Sec24a
Chlamydomonas reinhardtii Sec24a
Polysphondylium pallidum Sec24a
Acanthamoeba castellanii Sec24b
Sec24I
Physcomitrella patens Sec24c
Micromonas pusilla Sec24C
Acanthamoeba castellanii Sec24a
Trypanosoma brucei Sec24b
Naegleria gruberi Sec24b
Rhizophagus irregularis Sec24a
Polysphondylium pallidum Sec24c
Capsaspora owczarzaki Sec24b
Symbiodinium sp. Sec24d
Bigelowiella natans Sec24b
Sec24II
Tetrahymena thermophila Sec24e
Ectocarpus siliculosus Sec24c
0.9/87
Physcomitrella patens Sec24g
Micromonas pusilla Sec24D
Polysphondylium pallidum Sec24b
Dictyostelium discoideum Sec24b
Sec24III
0.7

rooted analyses prevented the recovery of branching order between the three paralogues (Figure 4-28).

4.5 Discussion

The COPII complex was one of the last coat complexes not characterized by comparative genomic and phylogenetic methods, and represents one of the last members of the protocoatomer fold-containing proteins for which these data were not available. The COPII coat is also one of the only membrane deformation complexes strictly involved in secretion, representing an important sampling point in order to gain a complete understanding of the evolution of the vesicle formation machinery. Therefore, comparative genomic and phylogenetic analyses were carried out in order to assess the distribution and evolution of subunits belonging to the COPII coat complex.

4.5.1 COPII and the evolution of a non-heterotetrameric coat complex

Comparative genomic analyses revealed that the machinery involved in the formation of COPII coats are all broadly conserved across eukaryotes with Sar1, Sec23, Sec24, Sec13, and Sec31 found in all organisms analyzed, while Sec12 and Sec16 are missing from multiple lineages. The broad distribution of these subunits suggests that this form of the COPII complex was present in LECA (Neumann et al., 2010). Phylogenetic analysis also allowed the reconstruction of paralogue numbers for each subunit present at that time. With the exception of three paralogues of Sec24, only one copy of each subunit was present, suggesting that Sec24 may have

been one of the first drivers of complexity in this coat system. Sec24 is the primary cargo binding subunit of the complex; therefore, multiple paralogues would have allowed for a greater diversity and specificity of cargo to be transported by COPII. It has been observed that some Sec24 paralogues have multiple binding sites each with specificity for different sorting signals or cargoes (Miller et al., 2003; Mossessova et al., 2003; Sucic et al., 2011; Wendeler et al., 2007), suggesting that the cargo specificity of Sec24 paralogues evolved early on in the eukaryotic lineage. Multiple paralogues, each with multiple binding sites, would have drastically increased the number of different cargo molecules capable of binding and would have permitted fine tuning of cargo specificity. The LECA is thought to have been a biflagellated organism (Cavalier-Smith et al., 2014); however, it remains unclear what other lifecycle stages (*i.e.*, amoeboid, *etc.*) it may have had. Should it be the case that the LECA underwent multiple lifecycle stages, encoding multiple differentially expressed forms of Sec24 would have enabled tighter regulation of the coat complex, as well as provide an additional regulatory mechanism for the various cargoes to be exported. This additional ancient Sec24 paralogue is yet another example of ancient, patchy proteins that are found in diverse eukaryotic taxa, but that have been lost from opisthokonts (Elias et al., 2012; Gabernet-Castello et al., 2013; Herman et al., 2011; Schlacht et al., 2013; Schlacht et al., 2014). The apparent asymmetry in the distribution of these proteins is suggestive of novel cell biology not found in typical model systems (*i.e.,* mammals, yeast). As Sec24III is found in stramenopiles, plants, and amoebozoans, taxa with ecological, agricultural, and medical importance, this protein may represent a useful target for exploitation.

4.5.2 Sec12 and Sec16 are frequently missing

The observation that both Sec16 and Sec12 are widely distributed, but seemingly missing from a variety of organisms, requires some reconciling with the literature. Sec12 is a Sar1 GEF and is responsible for activating Sar1 by swapping GDP for GTP, recruiting it to the ER membrane (Barlowe and Schekman, 1993; Weissman et al., 2001). Sec16 on the other hand, is a multifunctional scaffolding protein involved in both the recruitment of COPII subunits and has been implicated in regulating GTP hydrolysis (Kung et al., 2012). However variability in its precise function, and when and how Sec16 is incorporated into COPII vesicles has been observed (Iinuma et al., 2007; Ivan et al., 2008). The most likely explanation is that many of the observed absences of Sec12 and Sec16 represent cases of extreme sequence divergence and that *in vivo* analyses will likely identify divergent orthologues of these proteins.

On the other hand, if these absences are in fact gene losses, then they have occurred multiple times independently, with Sec12 being lost much more frequently than Sec16. This is surprising since Sec12 essentially acts to initiate COPII coat formation. However, Sec16 has been shown to be essential for vesicle formation in *S. cerevisiae* (Kaiser and Schekman, 1990). Very few organisms are missing both subunits, which could suggest that lower levels of ER to Golgi trafficking are necessary in these organisms or that, given the appropriate cellular conditions, these factors are not necessary for the formation of COPII coated vesicles. Alternatively, other GEFs or scaffold proteins may have functionally replaced Sec12 or Sec16. Promiscuity between GAPs and their GTPases has previously been

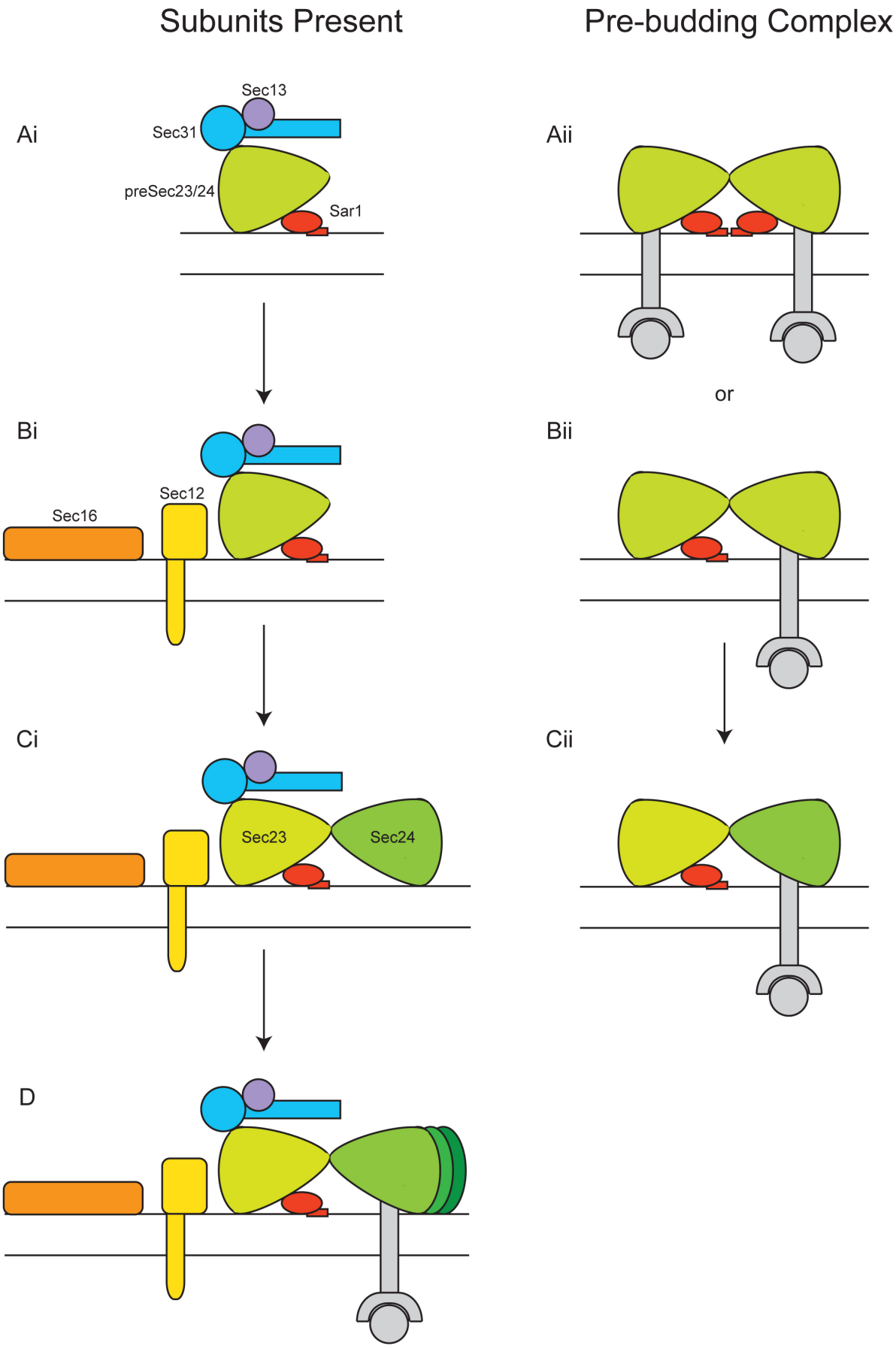
observed *in vitro,* with ELMOD2, an Arf-like protein GAP, being able to replace ArfGAPs and stimulate GTP hydrolysis on Arf (East et al., 2012).

4.5.3 Model for the evolution of the COPII complex

Phylogenetic analysis and reconstruction of the COPII coat complex in the LECA allows us to propose a model for the evolution of the coat from its state in early eukaryotes to its current incarnation in extant lineages; however, this model does not intend to imply any precise stoichiometry or quaternary structure, but simply propose a set of steps that may have occurred to give rise to this trafficking complex. The earliest form of COPII was likely made up of Sar1, Sec13, Sec31, and a preduplicate of Sec23 and Sec24 (preSec23/24; Figure 4-30Ai, Aii) that worked as a heteromeric complex. The preSec23/24 may have possessed both the Sar1-binding/GAP activity of Sec23 and the cargo binding capability of Sec24, suggesting that the preSec23/24 could bind both Sar1 and cargo. Alternatively, the preSec23/24 may have bound either Sar1 or cargo if these binding sites overlapped. Sar1 is also capable of binding cargo and may also have contributed to cargo binding in this ancient complex (Springer and Schekman, 1998). Eventually, Sec12 and Sec16 would be added to the coat-forming process (Figure 4-30Bi). Next, Sar1-binding/GAP activity and cargo binding were separated by the duplication of the preSec23/24 producing distinct Sec23 and Sec24 subunits (Figure 4-30Ci), possibly fixing their functions as the GAP and cargo binding subunits through subfunctionalization. Finally, iterative gene duplications would increase the cargo

specificity and capacity of COPII by giving rise to the three paralogues of Sec24 present in the LECA (Figure 4-30D).

Figure 4-30. Model for the evolution of the COPII complex from its earliest beginnings to the LECA. The left column represents the evolving complement of COPII coat components present in the eukaryotic lineage up to the LECA. The right column represents the evolution and hypothetical pre-budding complex across the same timeline. Note: this model is not meant to represent the stoichiometry or quaternary structure of individual subunits of the complex, but rather, is a hypothesis for the evolution of the complex itself. (Left column, Subunits Present) The earliest COPII coat was composed of Sar1, Sec13, Sec31, and a preduplicate of Sec23 and Sec24 (preSec23/24; Ai). Next, Sec12 and Sec16 would appear (Bi). Following this, a gene duplication of the preSec23/24 would have given rise to distinct Sec23 and Sec24 subunits (Ci). Finally, Sec24 would have undergone sequential gene duplications producing the three paralogues present in the LECA (D). (Right column, Pre-budding Complex) Two copies of preSec23/24 likely interacted during coat formation with two possibilities for protein binding: both copies of preSec23/24 may have been able to bind both Sar1 and cargo (Aii). Alternatively, one bound Sar1 and the other bound cargo (Bii). The precise configuration would have depended on the location of Sar1 and cargo binding sites in the preSec23/24 subunit. From here, the duplication of preSec23/24 produced Sec23 and Sec24, resulting in the subfunctionalization and fixation of GAP activity and cargo binding into two distinct subunits (Cii).


Subunits Present
Pre-budding Complex
Ai
Sec13
Sec31
preSec23/24
Sar1
Aii
or
Bi
Sec16
Sec12
Bii
Ci
Sec23
Sec24
Cii
D

Chapter 5: Comparative genomic and phylogenetic analysis of the newly discovered TSET complex

This chapter has been published as:

Hirst, J.*, Schlacht, A.*, Norcott, J.P., Traynor, D., Bloomfield, G., Antrobus, R., Kay, R.R., Dacks, J.B., Robinson, M.S. 2014. Characterization of TSET, an ancient and widespread membrane trafficking complex. eLife 3: e02866

*These authors contributed equally to this work

5.1 Overview

In the previous chapter, we examined the evolution of the COPII complex, important for trafficking in the early secretory system, and its strong conservation throughout eukaryotes. Chapter 5 similarly describes a novel coat complex from a comparative genomic and phylogenetic perspective. Although current evidence points to a function in the endocytic system, the TSET complex represents another example of a "patchy protein," proteins with a broad distribution across eukaryotes, but that are frequently lost. It should be noted that initial identification and all functional characterization of the complex was carried out by collaborators at the University of Cambridge, namely Dr. Jennifer Hirst and Dr. Margaret Robinson, who approached the Dacks lab to characterize this complex from an evolutionary perspective. This chapter will describe an evolutionary analysis of the newly identified TSET complex. Comparative genomics will be used to determine its distribution across eukaryotes and phylogenetic analysis will determine its relationship to other coat proteins, providing a model for its evolution.

5.2 The adaptins, heterotetrameric coat complexes

Post-TGN transport, both to the plasma membrane and within the endolysosomal system, is carried out by a related set of coat proteins called adaptor protein (AP) complexes. The endolysosomal system is responsible for a variety of functions, including degradation of proteins and lipids from within the cell, in addition to material brought into the cell either through endocytosis or through phagocytosis. This arm of the membrane trafficking system is also important for the

genesis of regulated secretory granules, such as dense core granules found in animal cells, or mucocysts and trichocysts found in ciliates (Elde et al., 2007).

Five distinct AP complexes have been identified, each of which is found across the diversity of eukaryotes and were present in the LECA (Hirst *et al.,* 2011). Each complex is made up of two large (β1–5 and γ/α/δ/ε/ζ, respectively), one medium (μ), and one small (σ) subunit (Boehm and Bonifacino, 2001). Both large subunits are primarily composed of α-solenoids, and are the result of an ancient pre-LECA gene duplication producing the β and γαδεζ clades (Duden et al., 1991; Schledzewski et al., 1999). The μ subunit is composed of an N-terminal longin domain and a C-terminal Mu-homology domain, and is primarily responsible for binding cargo (Owen and Evans, 1998). The σ subunit is a solitary longin domain (Rossi et al., 2004). The four subunits form a complex with the N-terminal regions of the large subunits, and the medium and small subunits forming the core of the complex, and the linker and ear domains of the large subunits extending beyond the rest of the complex (Robinson and Bonifacino, 2001).

The AP-1 complex is primarily localized to the TGN, and is responsible for bidirectional clathrin-dependent trafficking between the TGN and endosomes (Ren et al., 2013; Zhu et al., 1999). Localization to the TGN is the result of its interaction with Arf1 and phosphatidylinositol 4-phosphate (PI4P; Wang et al., 2003). AP-1 has also been shown to traffic cargo to the cell surface (Fölsch et al., 1999). Studies in mice have shown AP-1 knockouts to be embryonic lethal (Meyer et al., 2000; Zizioli et al., 1999).

The AP-2 complex mediates clathrin-dependent endocytosis at the plasma membrane of a variety of cargoes, including cell surface receptors and adhesion molecules, as well as their transport to early endosomes (Kamiguchi et al., 1998; Rappoport and Simon, 2009; Usami et al., 2014). Unlike the other AP complexes, AP-2 is primarily recruited by the presence of specific phosphoinositides, $PI(4,5)P_2$, rather than an activated GTPase (Gaidarov and Keen, 1999; Jackson et al., 2010). Knockout of AP-2 has also been shown to be embryonic lethal.

The AP-3 complex is involved in endolysosomal trafficking that may or may not depend on clathrin (Peden et al., 2002). Although initially identified as clathrin independent (Simpson et al., 1997), the potential for clathrin binding does exist, as the β-subunit possesses a clathrin-binding motif that is able to interact with clathrin *in vitro* (Dell'Angelica et al., 1998). However, mutation or deletion of the clathrin binding site does not impact complex formation, trafficking of Lamp1 to lysosomes, or the localization of AP-3 with respect to clathrin, suggesting that, even if an interaction with clathrin is possible, it is not necessary for trafficking of cargo (Peden et al., 2002). Instead, AP-3 has been proposed to interact with Vps41, a β-propeller/α-solenoid containing member of the HOPS multisubunit tethering complex (Asensio et al., 2013; Plemel et al., 2011; Rehling et al., 1999). AP-3 is also involved in the biogenesis of lysosome-related organelles, evidenced by its role in Hermansky Pudlak Syndrome type 2, a disorder characterized by oculocutaneous albinism and platelet abnormalities, resulting in excessive bleeding, generally resulting from the malformation of lysosome-related organelles (Dell'Angelica et al., 1999b).

The AP-4 complex appears to be a clathrin independent AP complex that localizes primarily to the TGN (Dell'Angelica et al., 1999a; Hirst et al., 1999). There are conflicting reports concerning the acceptor compartment of AP-4 coated transport intermediates. In polarized epithelial cells, AP-4 is important for sorting cargo to basolateral membranes (Simmen et al., 2002). In HeLa cells, amyloid precursor protein (APP) is distributed between the Golgi complex, endosomes and the plasma membrane (Caporaso et al., 1994; Haass et al., 1992). Depletion of AP-4 results in the redistribution of APP to the TGN from endosomes, suggesting a trafficking route to endosomes (Burgos et al., 2010). This conflicting evidence suggests that AP-4 may be involved in trafficking cargo to both endosomes and the plasma membrane (Hirst et al., 2013a).

Less is known about the more recently discovered AP-5 complex (Hirst et al., 2011). AP-5 is thought to play a role in endosomal trafficking, likely at the late endosome or lysosome. The available evidence suggests that AP-5 does not interact with clathrin; immunoprecipitation experiments using AP-5β did not identify clathrin as an interaction partner (Hirst et al., 2011), moreover β5 also lacks the clathrin binding box normally located in the linker region of the β-subunit. The secondary clathrin binding motifs, LLDLL and YQW (Dell'Angelica et al., 1998), generally present in the β-subunit are also missing (Hirst et al., 2011). Instead, AP-5 is thought to interact with SPG11 and SPG15, two proteins involved in hereditary spastic paraplegia, of which SPG11 shares structural similarity to both clathrin heavy chain and β'-COPI and α-COPI (Hirst et al., 2011), suggesting that they may act

as an additional membrane deformation complex (Hirst et al., 2013a; Hirst et al., 2013b).

Multiple groups have analyzed the evolution of the AP complexes. The earliest analysis, carried out by Duden et al., (1991) showed COPI-β branching earlier than all other adaptin large-subunit sequences. As more adaptin-relate sequences were identified in more species, phylogenetic analyses carried out by Schledzewski et al., (1999), identified an early branching order for the APs and the related COPI complex. As it was known that the large subunits [β and γαδ (AP-4 and AP-5 had not yet been discovered)] were part of a large multigene family (Simpson et al., 1997), likely having evolved from ancient gene duplications, they combined the datasets for both subunits. This would allow the β subunits to root the γαδ tree, and *vice versa*. Similarly, it was known that the N-terminal domain of the medium subunits and the entirety of the small subunits show clear sequence similarity (Cosson et al., 1996), and were therefore included in the same analysis in order to root the other subunit. Each subunit revealed the same branching order of COPI, followed by AP-3, then AP-1, and AP-2. The exception was the β-subunit of AP-1 and AP-2, which formed a single group, revealing that the β1 and β2 subunits found in humans are the result of a gene duplication after the divergence of arthropods (*D. melanogaster*). This was also observed by Dacks et al., (2008); using a more expansive set of taxa, at least three duplications producing independent β1 and β2 subunits have occurred throughout eukaryotes. More recently, an additional gene duplication of the β subunit in Fungi has been identified (Barlow et al., 2014). A recent concatenated phylogenetic analysis incorporating AP-4 and AP-5 revealed a

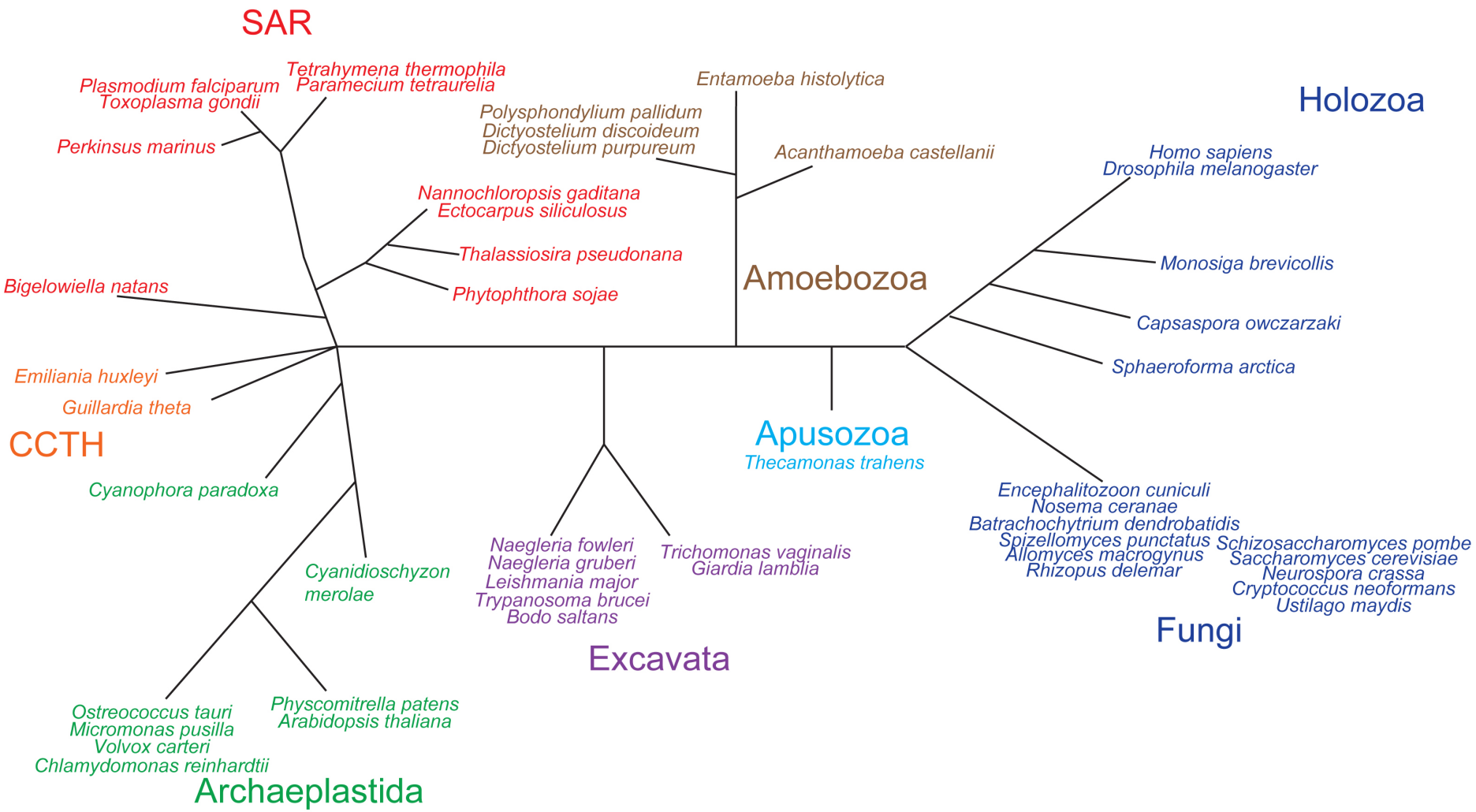
similar topology as Schledzewski et al., (1999); Hirst et al., (2011) showed that AP-3 is the earliest branching AP, followed by AP-5, AP-4, AP-1, and AP-2, with COPI branching outside of the AP clade.

In this chapter, I will carry out a comparative genomic and phylogenetic analysis of the recently discovered TSET complex. TSET is thought to be related to the AP complexes and COPI. Therefore, phylogenetic analysis will be carried out to determine the relationships between these three sets of coats. TSET will also be incorporated into an existing framework for the evolution of AP and COPI complexes in order to determine how these complexes, and the membrane trafficking system more broadly, evolved.

5.3 Abbreviated materials and methods

As the initial protein sequences of TSET were identified using an HHpred-based approach that incorporates structural information into HMMs, the forward BLAST step in comparative genomics was forgone in favour of HMM-based searching using HMMer. Taxa sampled are shown in Figure 5-1. HMMER searches were carried out as in section 2.2.2, with reciprocal BLAST experiments for validation of orthology carried out against the *A. thaliana, D. discoideum*, and *N. gruberi* genome databases. Phylogenetic analyses were carried out as in section 2.3. Table 5-1 contains details of each phylogenetic analysis, including: number of taxa, length of masked alignment, and model parameters for each method. Phylogenetic analyses were carried out using the CIPRES web server (Miller et al., 2010).

Figure 5-1. Eukaryotic taxa used in comparative genomic analysis. Summary of all taxa used to search for subunits of TSET and their relative phylogenetic positions. Taxa are coloured by supergroup.


SAR
Tetrahymena thermophila
Paramecium tetraurelia
Plasmodium falciparum
Toxoplasma gondii
Perkinsus marinus
Nannochloropsis gaditana
Ectocarpus siliculosus
Thalassiosira pseudonana
Phytophthora sojae
Bigelowiella natans
Emiliania huxleyi
Guillardia theta
CCTH
Cyanophora paradoxa
Cyanidioschyzon
merolae
Ostreococcus tauri
Micromonas pusilla
Volvox carteri
Chlamydomonas reinhardtii
Physcomitrella patens
Arabidopsis thaliana
Archaeplastida
Entamoeba histolytica
Polysphondylium pallidum
Dictyostelium discoideum
Dictyostelium purpureum
Acanthamoeba castellanii
Amoebozoa
Naegleria fowleri
Naegleria gruberi
Leishmania major
Trypanosoma brucei
Bodo saltans
Trichomonas vaginalis
Giardia lamblia
Excavata
Apusozoa
Thecamonas trahens
Holozoa
Homo sapiens
Drosophila melanogaster
Monosiga brevicollis
Capsaspora owczarzaki
Sphaeroforma arctica
Encephalitozoon cuniculi
Nosema ceranae
Batrachochytrium dendrobatidis
Spizellomyces punctatus
Allomyces macrogynus
Rhizopus delemar
Schizosaccharomyces pombe
Saccharomyces cerevisiae
Neurospora crassa
Cryptococcus neoformans
Ustilago maydis
Fungi

Structural modeling was carried out using the Phyre2.0 server as described in section 2.4.

Table 5-1. Parameters of phylogenetic analysis, corresponding dataset, and figure number for each TSET subunit. Phylobayes and PhyML were only implemented for the concatenated phylogenetic analysis.

Figure	Dataset name	Number of taxa	Length of alignment (a.a.)	Evolutionary model used			
				MrBayes	RAxML	Phylobayes	PhyML
5-5	TPLATE.R2	121	560	mixed + gamma	LG+CAT+F	-	-
5-6	TPLATE.R4	91	402	mixed + gamma	LG+CAT+F	-	-
5-7	TSAUCER.R2	154	562	mixed + gamma	LG+CAT+F	-	-
5-8	TSAUCER.R4	123	284	mixed + gamma	LG+CAT+F	-	-
5-9	TCUP.R2	159	379	mixed + gamma	LG+CAT+F	-	-
5-10	TCUP.R4	133	187	mixed + gamma	LG+CAT+F	-	-
5-11	TSPOON.R2	139	141	mixed + gamma	LG+CAT+F	-	-
5-12	Concat.R6	112	1466	mixed + gamma	LG+CAT+F	LG+CAT	LG+I+G+F
5-13	TTRAY.R4	60	437	mixed + gamma	LG+CAT+F	-	-
5-14	TTRAY.R5	28	376	mixed + gamma	LG+CAT+F	-	-

5.4 Results

5.4.1 Identification of TSET, a novel heterotetrameric coat complex

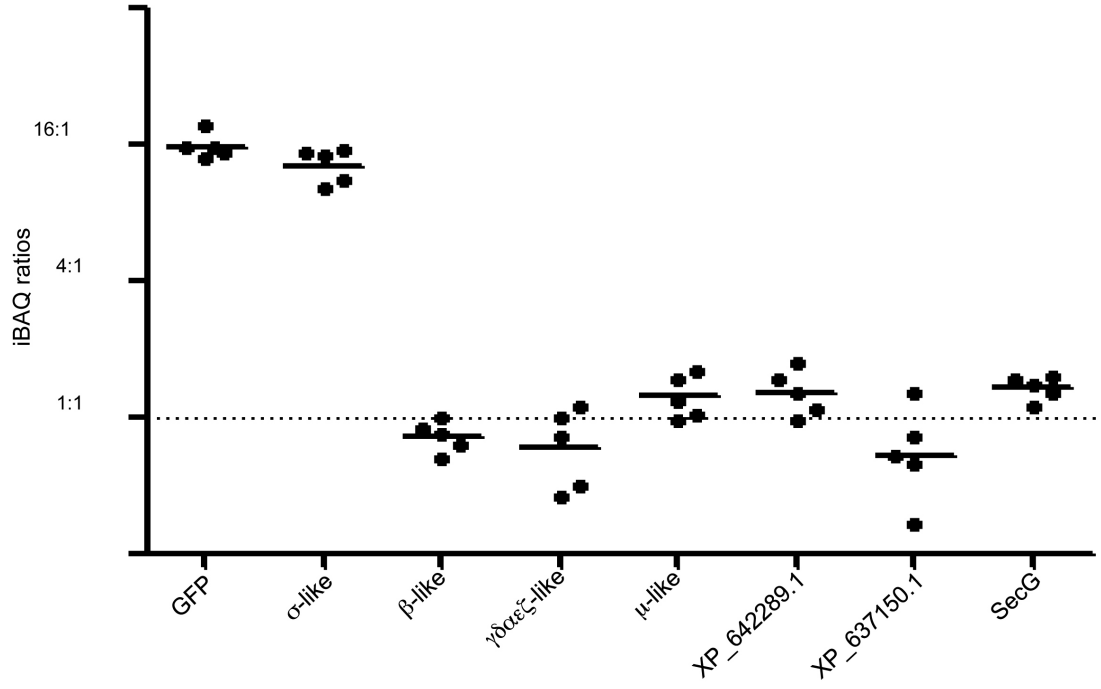
Following on their discovery of AP5, Dr. Hirst and Dr. Robinson set out to determine if additional undetected AP-like coat complexes exist. They constructed and validated a new approach to identify highly divergent sequences. They initially

identified four sequences corresponding to subunits of an AP complex in three unrelated organisms: *A. thaliana, D. discoideum,* and *N. gruberi*, suggesting the presence of a previously unreported coat complex (Hirst et al., 2014). However, the mere presence of four subunits does not guarantee complex formation; immunoprecipitation in the model organism *D. discoideum* and mass spectrometry were carried out in an attempt to show interaction. As expected, each subunit precipitated at roughly equimolar ratios (Figure 5-2), indicating that these subunits do indeed form a complex *in vivo* (Hirst et al., 2014). In addition, they identified two additional subunits that bind the AP-like core to form the complete coat (Hirst et al., 2014). The β-adaptin-like subunit had previously been identified in *A. thaliana* by screening for mutants in cell plate formation during cell division, and was named TPLATE (Van Damme et al., 2011). The other subunits were named using the same pattern: the γ-like is called 'TSAUCER', the μ-like is called 'TCUP', the σ-like is called 'TSPOON', and the outer components identified by immunoprecipitation were named 'TTRAY1' and 'TTRAY2', respectively. The analyses represented by Figure 5-2 were carried out by Dr. Hirst, but are shown here as they represent an important step in the validation of the TSET complex.

5.4.2 Structural predictions indicate that TSET is structurally similar to the AP complexes

To confirm that the sequences identified by Dr. Hirst and Dr. Robinson are indeed structurally similar to other known adaptin and COPI subunits, tertiary structure predictions were carried out using the Phyre2.0 server on the sequences

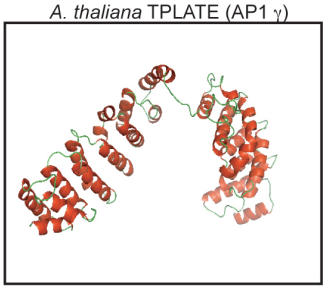
Figure 5-2. TSET subunits interact to form a complex. Immunoprecipiation using GFP-tagged TSPOON expressed in *D. discoideum* was carried out in order to determine whether the four TSET subunits interact *in vivo*. Shown are iBAQ ratios (an estimate of molar ratios) for proteins that consistently precipitate with GFP-TSPOON. All subunits appear to be equimolar. Higher ratios of GFP and GFP-TSPOON are the result of overexpression.


iBAQ ratios
16:1
4:1
1:1
GFP
σ-like
β-like
γδαεζ-like
μ-like
XP_642289.1
XP_637150.1
SecG

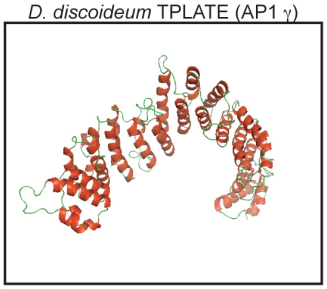
from *A. thaliana, D. discoideum,* and *N. gruberi*. Members of the TSET complex are predicted to have structures similar to those predicted for the corresponding AP/COPI subunit (Figure 5-3). All three TPLATE sequences are predicted as α-solenoid proteins, consistent with known AP structures (Heldwein et al., 2004). Similarly, TSAUCER is also predicted to be an α-solenoid protein, based on modeling against the structures of clathrin adaptor (AP) proteins. However, the *A. thaliana* structure was poorly predicted, likely the result of a C-terminal SH3 domain confounding the signal during the PSI-BLAST step of the Phyre search, resulting in the inclusion of unrelated SH3-domain-containing proteins. Removal of the SH3 domain resulted in a coiled-coil structure modeled from an alpha helicoidal repeat protein; however, the second best template was AP-1γ, suggesting that the *A. thaliana* TSAUCER does indeed form an α-solenoid. The medium subunit of AP/COPI complexes is composed of an N-terminal longin domain, and a C-terminal Mu-homology domain (Rossi et al., 2004). None of the structures predicted a complete medium subunit-like fold for any TCUP sequence; however, elements of that fold were recovered in each sequence analyzed. The *A. thaliana* and *N. gruberi* sequences were modeled from *S. cerevisiae* Syp1p protein, a mu-homology domain-containing protein not known to be part of any extant coat complex. Neither of these two sequences were modeled as possessing the N-terminal longin domain, but the mu-homology domain of the *A. thaliana* sequences was much better predicted than that of the *N. gruberi* sequence. By contrast, the *D. discoideum* TCUP sequence was modeled as possessing the N-terminal longin domain, but not the C-terminal mu-homology domain. Finally, for all three organisms, the small subunit, TSPOON, was

Figure 5-3. Predicted tertiary structures of TSET subunits from *A. thaliana, D. discoideum,* and *N. gruberi*. Structural predictions for each TSET subunit from *A. thaliana, D. discoideum, and N. gruberi* were carried out using the Phyre2.0 server (see section 2.4), with the model selected by the program to predict the structure in parentheses above each panel. Structural predictions are consistent with known structures of AP subunits. The TPLATE and TSAUCER sequences are modelled as α-solenoids. The TCUP sequences are modelled as a combination of longin and mu-homology domain-containing proteins. The three TSPOONs are all modelled as longin domains and the TTRAYs are modelled as β-propeller/α-solenoid domain-containing proteins. In all but two cases, a homologous protein was used for structural modeling. The *A. thaliana* TSAUCER was modeled without the C-terminal SH3 domain because it resulted in the use of an SH2 domain containing protein, creating a poor quality prediction. The *D. discoideum* TCUP was modeled using COPI-ζ rather than δ, likely because of an extremely divergent, or missing, mu-homology domain. No model is shown for *N. gruberi* TTRAY1 as no homologue was identified.

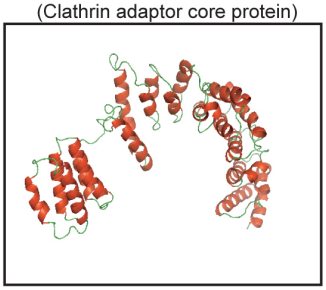
A. thaliana TPLATE (AP1 γ)



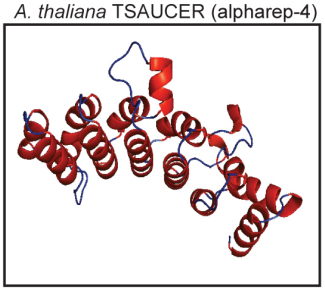
D. discoideum TPLATE (AP1 γ)



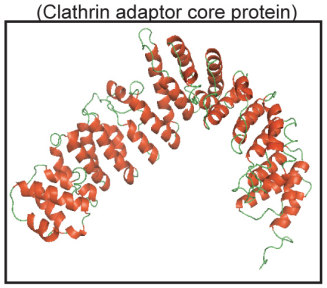
N. gruberi TPLATE
(Clathrin adaptor core protein)



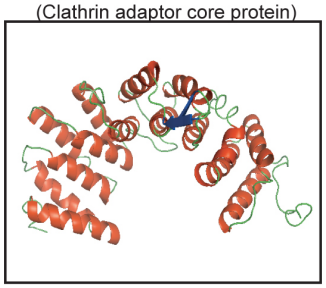
A. thaliana TSAUCER (alpharep-4)



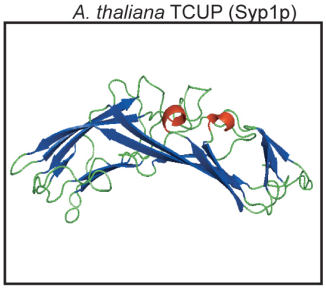
D. discoideum TSAUCER
(Clathrin adaptor core protein)



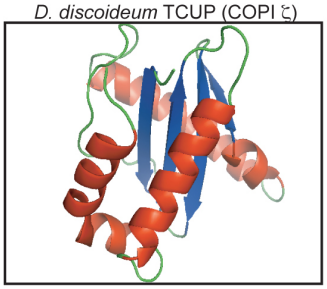
N. gruberi TSAUCER
(Clathrin adaptor core protein)



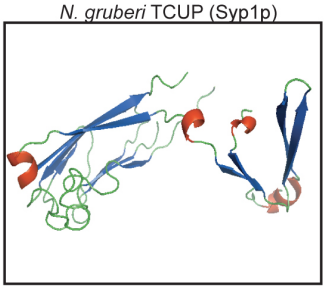
A. thaliana TCUP (Syp1p)



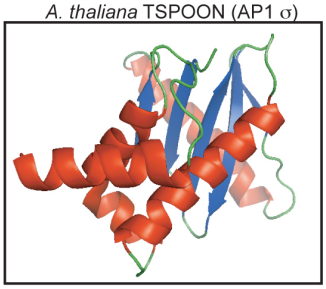
D. discoideum TCUP (COPI ζ)



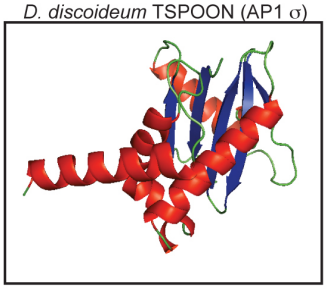
N. gruberi TCUP (Syp1p)



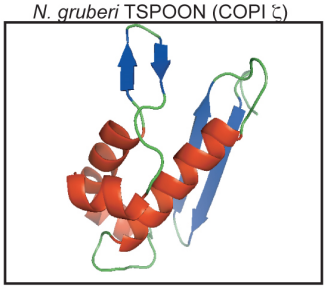
A. thaliana TSPOON (AP1 σ)



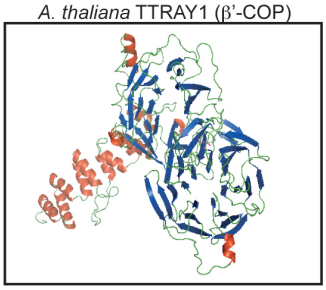
D. discoideum TSPOON (AP1 σ)



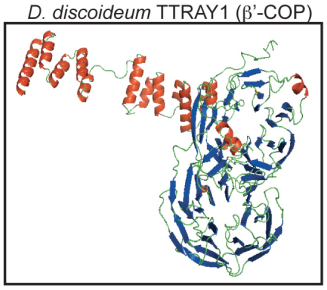
N. gruberi TSPOON (COPI ζ)



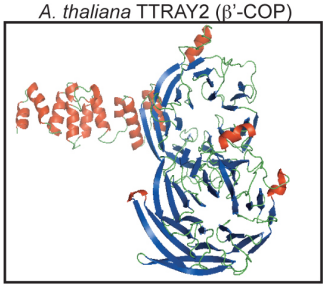
A. thaliana TTRAY1 (β'-COP)



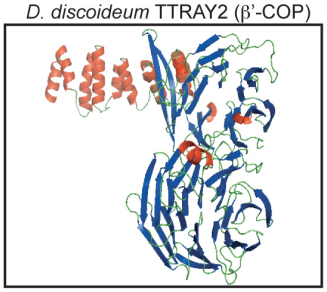
D. discoideum TTRAY1 (β'-COP)



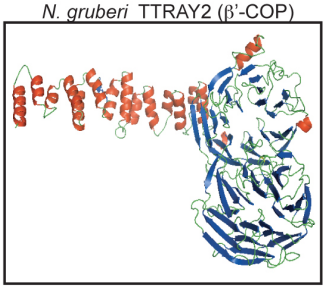
A. thaliana TTRAY2 (β'-COP)



D. discoideum TTRAY2 (β'-COP)



N. gruberi TTRAY2 (β'-COP)



modeled as a longin domain, as would be predicted from structures of the σ subunits of APs or the ζ subunit of COPI.

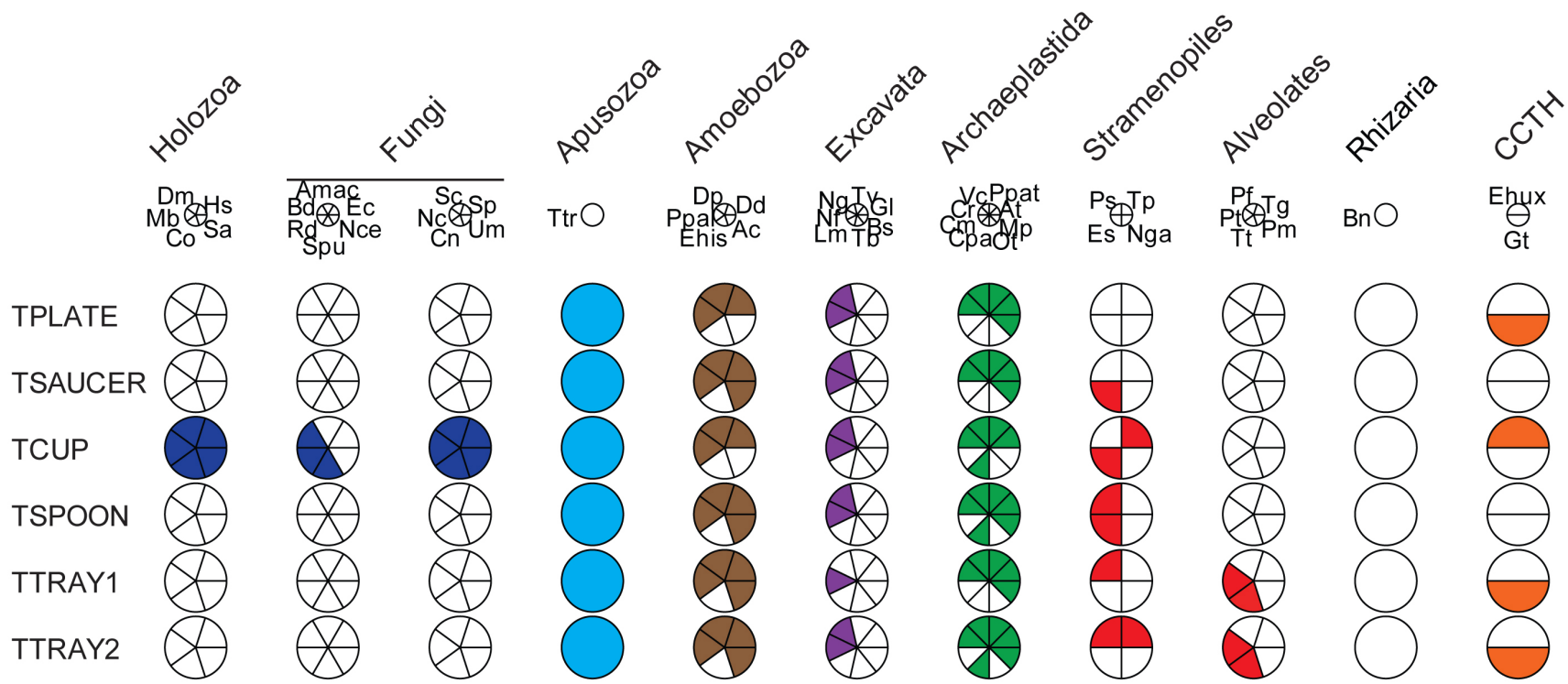
As mentioned above, TTRAY1 and TTRAY2 are not part of the heterotetrameric core, but do interact with it. Homology modeling of these subunits revealed identical folds: two β-propeller domains, followed by an α-solenoid (Figure 5-3), very similar to the structures of β'- and α-COPI on which they were modeled, suggesting a likely common ancestry of these two sets of subunits.

5.4.3 Comparative genomics indicates that TSET is a broadly distributed, but patchy complex

Starting HMMs were built for each subunit using the sequences from *A. thaliana, D. discoideum,* and *N. gruberi*, and were used to search the genomes indicated in Figure 5-1. Candidate sequences identified using HMMER were verified by reciprocal BLAST against the *A. thaliana, D. discoideum,* and *N. gruberi* genomes, and were considered positive hits if the candidate retrieved the appropriate orthologue as the top hit with an E-value at least two orders of magnitude smaller than the next best sequence in at least one of the three reciprocal BLAST experiments. Newly identified sequences were incorporated into the HMM to increase the specificity and selectivity of the model, prior to searching the next genome.

TSET displays a broad but patchy distribution; it is found in diverse eukaryotes, but is frequently missing (Figure 5-4). However, the broad distribution

Figure 5-4. TSET is broadly, but sparsely distributed. Summary of comparative genomic analyses indicate that TSET is found in a diverse set of representative eukaryotes. Presence of the complete complex in at least four supergroups suggests its presence in the LECA with frequent secondary loss. Solid sectors indicate sequences identified and classified using BLAST and HMMER. Empty sectors indicate taxa in which no significant orthologues were identified. Solid sectors in the Holozoa and Fungi represent F-BAR domain-containing FCHo and Syp1, respectively. The key to taxon name abbreviations is inset. Names in bold indicate taxa with all six subunits.



Holozoa
Hs = *Homo sapiens*
Dm = *Drosophila melanogaster*
Mb = *Monosiga brevicollis*
Co = *Capsaspora owczarzaki*
Sa = *Sphaeroforma arctica*

Fungi
Amac = *Allomyces macrogynus*
Bd = *Batrachochytrium dendrobatidis*
Rd = *Rhizopus delemar*
Spu = *Spizellomyces punctatus*
Ec = *Encephalitozoon cuniculi*
Nce = *Nosema ceranae*
Sc = *Saccharomyces cerevisiae*
Nc = *Neurospora crassa*
Cn = *Cryptococcus neoformans*
Um = *Ustilago maydis*
Sp = *Schizosaccharomyces pombe*

Apusozoa
Ttr = *Thecamonas trahens*

Amoebozoa
Dd = *Dictyostelium discoideum*
Dp = *Dicytostelium purpureum*
Ppal = *Polysphondylium pallidum*
Ehis = *Entamoeba histolytica*
Ac = *Acanthamoeba castellanii*

Excavata
Ng = *Naegleria gruberi*
Nf = *Naegleria fowleri*
Lm = *Leishmania major*
Tb = *Trypanosoma brucei*
Bs = *Bodo saltans*
Gl = *Giardia lamblia*
Tv = *Trichomonas vaginalis*

Archaeplastida
At = *Arabidopsis thaliana*
Ppat = *Physcomitrella patens*
Vc = *Volvox carteri*
Cr = *Chlamydomonas reinhardtii*
Cm = *Cyanidioschyzon merolae*
Cpa = *Cyanophora paradoxa*
Ot = *Ostreococcus tauri*
Mp = *Micromonas pusilla*

Stramenopiles
Ps = *Phytophthora sojae*
Es = *Ectocarpus siliculosus*
Nga = *Nannochloropsis gaditana*
Tp = *Thalassiosira pseudonana*

Alveolates
Tg = *Toxoplasma gondii*
Pf = *Plasmodium falciparum*
Pt = *Paramecium tetraurelia*
Tt = *Tetrahymena thermophila*
Pm = *Perkinsus marinus*

Rhizaria
Bn = *Bigelowiella natans*

CCTH
Ehux = *Emiliania huxleyi*
Gt = *Guillardia theta*

suggests that this complex was present in the LECA. Complete complexes (all six subunits) are identifiable in representatives of four supergroups (bold in Figure 5-4): *T. trahens* (Opisthokonta); *D. discoideum, Dictyostelium purpureum,* and *Polysphondylium pallidum* (Amoebozoa); *N. fowleri* (Excavata); *C. reinhardtii, V. carteri, P. patens,* and *A. thaliana* (Archaeplastida). Most other eukaryotes sampled have either a partial TSET complex, explained as the product of extreme sequence divergence, rendering homologues unidentifiable by bioinformatics, or explained through secondary loss. The latter explanation is most likely the case for the opisthokont lineage, where only homologues of TCUP were identified, or in the alveolates, where ciliates only possess TTRAY1 and TTRAY2 (Figure 5-4). Surprisingly, homology searching revealed that the opisthokont TCUP orthologues belong to a previously known family of proteins collectively known as the muniscins, mu-homology domain-containing proteins that possess an N-terminal F-BAR domain in place of a longin domain (Reider et al., 2009). Muniscins include FCHo and SGIP in *H. sapiens*, and Syp1 in *S. cerevisiae*. FCHo and Syp1 are thought to be involved in the formation of AP2-clathrin vesicles by inducing membrane curvature (Henne et al., 2010), and promoting the growth of AP-2 by stabilizing the open conformation, freeing the cargo-binding site (Hollopeter et al., 2014; Umasankar et al., 2014). It is worthwhile noting that since this analysis was carried out, subunits of TSET have been identified by lab mates in comparative genomic analyses of other eukaryotes: Ms. L. Lee has identified TCUP in the transcriptomes of the haptophytes *Isochrysis galbana*, *Gephyrocapsa oceanica*, and three additional strains of *Emiliania huxleyi*. Additionally, Mr. L. Barlow has identified TPLATE,

TTRAY1, and TTRAY2 in the genome of the amoebozoan *Mastigamoeba balamuthii* and TPLATE, TSPOON, and TTRAY1 in the genome of *Monocercomonoides sp.* (personal communications).

5.4.4 Phylogenetic analysis of TSET

Phylogenetic analysis of individual TSET subunits were carried out to determine whether these sequences represent an additional AP complex, a duplication of the COPI complex, or an additional complex distinct from both the APs and COPI. Single-gene trees were generated as detailed in section 2.3, using the phylogenetic dataset generated by Hirst et al., (2011) to analyze the position AP-5 to relative to the other AP complexes. Past analyses have shown generally poor resolution for single gene trees of AP complexes (Hirst et al., 2011). Resolution between these complexes is generally obtained by stringing together the gene sequences of the four subunits into a supergene or 'concatenated gene' that utilizes the phylogenetic signal from all four subunits in a single analysis (Hirst et al., 2011). One requirement for concatenation is that all genes being incorporated into the analysis must have largely the same phylogenetic signal; there can be no strong dissonance between the trees of the individual genes being strung together. Therefore, phylogenetic analyses of single genes (subunits) were carried out.

As only the AP β- and μ-subunits were used for concatenation (COPI-β and -δ, respectively) by Hirst et al., (2011), the analysis was supplemented by searching for the γ- and ζ-subunits of COPI in the relevant taxa for inclusion into both single-gene trees and the concatenated analysis (see below). For each single-gene analysis,

multiple rounds of phylogenetic analysis were carried out, but only the tree with the strongest support for the node of interest is shown. Previous rounds were used to identify exceedingly long branches, and to identify instances of species-specific duplication. In the case of long branches, sequences were removed, and in the case of multiple paralogues, only the paralogue with the shortest branch length was retained. Both actions were carried out to mitigate the potential effects of LBA, and to reduce the complexity of the data set.

5.4.4.1 TPLATE

As mentioned above, TPLATE was previously identified as a β-adaptin like protein important for the formation of the cell plate after division (Van Damme et al., 2011). Comparative genomic analysis suggested that TPLATE is most similar to the β-APs and β-COP, and was therefore included into the β-subunit dataset. Phylogenetic analysis identified a weakly supported albeit monophyletic TPLATE clade, a strongly monophyletic COPI-β clade, and a polyphyletic AP-β clade (Figure 5-5). Because TPLATE was clearly excluded from the COPI clade, all COPI-β sequences were removed, assuming that reducing the size and complexity of the alignment would help resolve the relative position of TPLATE and the AP complexes. Although still very weak, support for distinctly monophyletic TPLATE and AP clades increased, suggesting that TPLATE likely does not branch within the clade of known AP complexes (Figure 5-6).

Figure 5-5. Phylogenetic analysis of TPLATE indicates that TSET is distinct from COPI. Phylogenetic analysis of TPLATE, β-APs, and COPI-β was carried out to determine the relationships between these subunits. Phylogenetic analysis resulted in a very strongly supported COPI clade, excluding a weakly monophyletic TPLATE clade, nested within the APs, although with no backbone support. The results indicate that TPLATE does not branch within the COPI-β clade. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (MrBayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95, closed light circles ≥ 0.95/75, open circles ≥ 0.8/50.

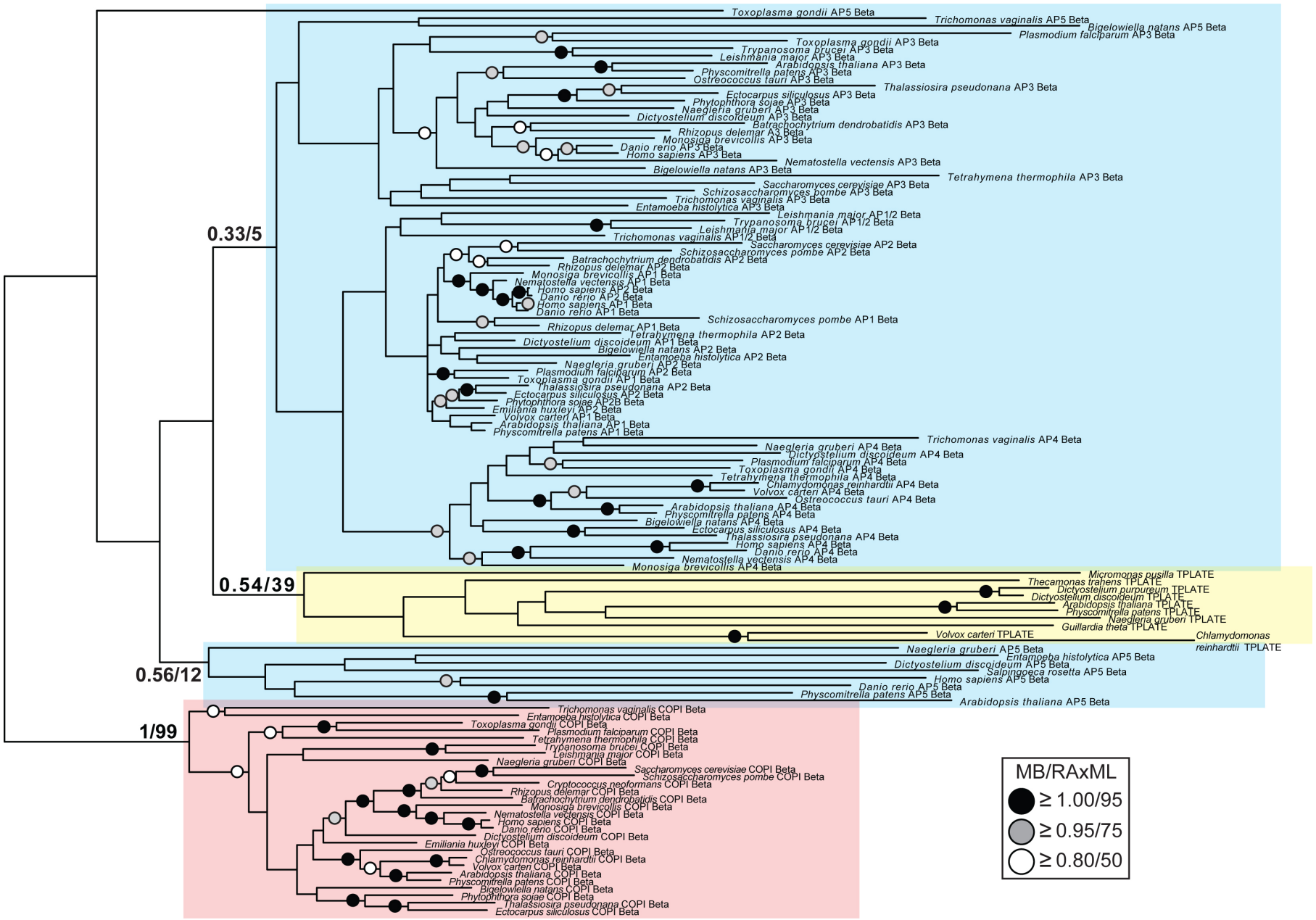

Toxoplasma gondii AP5 Beta
Trichomonas vaginalis AP5 Beta
Bigelowiella natans AP5 Beta
Plasmodium falciparum AP3 Beta
Toxoplasma gondii AP3 Beta
Trypanosoma brucei AP3 Beta
Leishmania major AP3 Beta
Arabidopsis thaliana AP3 Beta
Physcomitrella patens AP3 Beta
Ostreococcus tauri AP3 Beta
Thalassiosira pseudonana AP3 Beta
Ectocarpus siliculosus AP3 Beta
Phytophthora sojae AP3 Beta
Naegleria gruberi AP3 Beta
Dictyostelium discoideum AP3 Beta
Batrachochytrium dendrobatidis AP3 Beta
Rhizopus delemar A3 Beta
Monosiga brevicollis AP3 Beta
Danio rerio AP3 Beta
Homo sapiens AP3 Beta
Nematostella vectensis AP3 Beta
Bigelowiella natans AP3 Beta
Tetrahymena thermophila AP3 Beta
Saccharomyces cerevisiae AP3 Beta
Schizosaccharomyces pombe AP3 Beta
Trichomonas vaginalis AP3 Beta
Entamoeba histolytica AP3 Beta
0.33/5
Leishmania major AP1/2 Beta
Trypanosoma brucei AP1/2 Beta
Leishmania major AP1/2 Beta
Trichomonas vaginalis AP1/2 Beta
Saccharomyces cerevisiae AP2 Beta
Schizosaccharomyces pombe AP2 Beta
Batrachochytrium dendrobatidis AP2 Beta
Rhizopus delemar AP2 Beta
Monosiga brevicollis AP1 Beta
Nematostella vectensis AP1 Beta
Homo sapiens AP2 Beta
Danio rerio AP2 Beta
Homo sapiens AP1 Beta
Danio rerio AP1 Beta
Schizosaccharomyces pombe AP1 Beta
Rhizopus delemar AP1 Beta
Tetrahymena thermophila AP2 Beta
Dictyostelium discoideum AP1 Beta
Bigelowiella natans AP2 Beta
Entamoeba histolytica AP2 Beta
Naegleria gruberi AP2 Beta
Plasmodium falciparum AP2 Beta
Toxoplasma gondii AP1 Beta
Thalassiosira pseudonana AP2 Beta
Ectocarpus siliculosus AP2 Beta
Phytophthora sojae AP2B Beta
Emiliania huxleyi AP2 Beta
Volvox carteri AP1 Beta
Arabidopsis thaliana AP1 Beta
Physcomitrella patens AP1 Beta
Trichomonas vaginalis AP4 Beta
Naegleria gruberi AP4 Beta
Dictyostelium discoideum AP4 Beta
Plasmodium falciparum AP4 Beta
Toxoplasma gondii AP4 Beta
Tetrahymena thermophila AP4 Beta
Chlamydomonas reinhardtii AP4 Beta
Volvox carteri AP4 Beta
Ostreococcus tauri AP4 Beta
Arabidopsis thaliana AP4 Beta
Physcomitrella patens AP4 Beta
Bigelowiella natans AP4 Beta
Ectocarpus siliculosus AP4 Beta
Thalassiosira pseudonana AP4 Beta
Homo sapiens AP4 Beta
Danio rerio AP4 Beta
Nematostella vectensis AP4 Beta
Monosiga brevicollis AP4 Beta
0.54/39
Micromonas pusilla TPLATE
Thecamonas trahens TPLATE
Dictyostelium purpureum TPLATE
Dictyostelium discoideum TPLATE
Arabidopsis thaliana TPLATE
Physcomitrella patens TPLATE
Naegleria gruberi TPLATE
Guillardia theta TPLATE
Volvox carteri TPLATE
Chlamydomonas reinhardtii TPLATE
Naegleria gruberi AP5 Beta
Entamoeba histolytica AP5 Beta
Dictyostelium discoideum AP5 Beta
Salpingoeca rosetta AP5 Beta
Homo sapiens AP5 Beta
Danio rerio AP5 Beta
Physcomitrella patens AP5 Beta
Arabidopsis thaliana AP5 Beta
0.56/12
Trichomonas vaginalis COPI Beta
Entamoeba histolytica COPI Beta
Toxoplasma gondii COPI Beta
Plasmodium falciparum COPI Beta
Tetrahymena thermophila COPI Beta
Trypanosoma brucei COPI Beta
Leishmania major COPI Beta
Naegleria gruberi COPI Beta
Saccharomyces cerevisiae COPI Beta
Schizosaccharomyces pombe COPI Beta
Cryptococcus neoformans COPI Beta
Rhizopus delemar COPI Beta
Batrachochytrium dendrobatidis COPI Beta
Monosiga brevicollis COPI Beta
Nematostella vectensis COPI Beta
Homo sapiens COPI Beta
Danio rerio COPI Beta
Dictyostelium discoideum COPI Beta
Emiliania huxleyi COPI Beta
Ostreococcus tauri COPI Beta
Chlamydomonas reinhardtii COPI Beta
Volvox carteri COPI Beta
Arabidopsis thaliana COPI Beta
Physcomitrella patens COPI Beta
Bigelowiella natans COPI Beta
Phytophthora sojae COPI Beta
Thalassiosira pseudonana COPI Beta
Ectocarpus siliculosus COPI Beta
1/99
MB/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50
0.4

Figure 5-6. Phylogenetic analysis suggests that TPLATE is distinct from the AP-β subunits. To determine whether TPLATE branches within the AP clade, all COPI-β sequences were removed from the data set in Figure 5-5. Phylogenetic analysis reconstructed a weakly monophyletic TPLATE clade branching separately from the AP-β sequences, suggesting that TSET likely branches outside of the clade containing the AP complexes. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (MrBayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95, closed light circles ≥ 0.95/75, open circles ≥ 0.8/50.


0.47/12
Micromonas pusilla TPLATE
Chlamydomonas reinhardtii TPLATE
Thecamonas trahensTPLATE
Volvox carteri TPLATE
Guillardia theta TPLATE
Naegleria gruberi TPLATE
Arabidopsis thaliana TPLATE
Physcomitrella patens TPLATE
Dictyostelium purpureum TPLATE
Dictyostelium discoideum TPLATE
Physcomitrella patens AP5 Beta
Arabidopsis thaliana AP5 Beta
Danio rerio AP5 Beta
Bigelowiella natans AP5 Beta
Dictyostelium discoideum AP5 Beta
Homo sapiens AP5 Beta
Toxoplasma gondii AP5 Beta
Leishmania major AP1/2 Beta
Trypanosoma brucei AP1/2 Beta
Leishmania major AP1/2 Beta
Saccharomyces cerevisiae AP2 Beta
Trichomonas vaginalis AP2 Beta
Tetrahymena thermophila AP2 Beta
Bigelowiella natans AP2 Beta
Naegleria gruberi AP2 Beta
Plasmodium falciparum AP2 Beta
Toxoplasma gondii AP1 Beta
Thalassiosira pseudonana AP2 Beta
Ectocarpus siliculosus AP2 Beta
Phytophthora sojae AP2B Beta
Emiliania huxleyi AP2 Beta
Volvox carteri AP1 Beta
Arabidopsis thaliana AP1 Beta
Physcomitrella patens AP1 Beta
Dictyostelium discoideum AP1 Beta
Monosiga brevicollis AP1 Beta
Nematostella vectensis AP1 Beta
Homo sapiens AP2 Beta
Danio rerio AP2 Beta
Homo sapiens AP1 Beta
Danio rerio AP1 Beta
Schizosaccharomyces pombe AP1 Beta
Rhizopus delemar AP1 Beta
Entamoeba histolytica AP2 Beta
Batrachochytrium dendrobatidis AP2 Beta
Rhizopus delemar AP2 Beta
Schizosaccharomyces pombe 2 Beta
Trichomonas vaginalis AP4 Beta
Naegleria gruberi AP4 Beta
Dictyostelium discoideum AP4 Beta
Tetrahymena thermophila AP4 Beta
Plasmodium falciparum AP4 Beta
Toxoplasma gondii AP4 Beta
Chlamydomonas reinhardtii AP4 Beta
Volvox carteri AP4 Beta
Ostreococcus tauri AP4 Beta
Arabidopsis thaliana AP4 Beta
Physcomitrella patens AP4 Beta
Bigelowiella natans AP4 Beta
Ectocarpus siliculosus AP4 Beta
Thalassiosira pseudonana AP4 Beta
Homo sapiens AP4 Beta
Danio rerio AP4 Beta
Nematostella vectensis AP4 Beta
Monosiga brevicollis AP4 Beta
Naegleria gruberi AP5 Beta
Tetrahymena thermophila AP3 Beta
Plasmodium falciparum AP3 Beta
Toxoplasma gondii AP3 Beta
Saccharomyces cerevisiae AP3 Beta
Trypanosoma brucei AP3 Beta
Leishmania major AP3 Beta
Entamoeba histolytica AP3 Beta
Arabidopsis thaliana AP3 Beta
Physcomitrella patens AP3 Beta
Ostreococcus tauri AP3 Beta
Thalassiosira pseudonana AP3 Beta
Ectocarpus siliculosus AP3 Beta
Phytophthora sojae AP3 Beta
Naegleria gruberi AP3 Beta
Dictyostelium discoideum AP3 Beta
Batrachochytrium dendrobatidis AP3 Beta
Rhizopus delemar AP3 Beta
Monosiga brevicollis AP3 Beta
Danio rerio AP3 Beta
Homo sapiens AP3 Beta
Nematostella vectensis AP3 Beta
Bigelowiella natans AP3 Beta
Trichomonas vaginalis AP3 Beta
Schizosaccharomyces pombe AP3 Beta
Salpingoeca rosetta AP5 Beta
MB/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50
0.5

5.4.4.2 TSAUCER

Comparative genomic analysis identified TSAUCER as a homologue of the γ-COPI/γαδεζ-AP subunits, confirming the results obtained through homology modeling. Similar to TPLATE, TSAUCER was strongly excluded from the γ-COPI clade, but only weakly excluded from clade containing the AP-γαδεζ subunits (Figure 5-7), and was nearly monophyletic, with the *Micromonas pusilla* TSAUCER branching outside of the TSAUCER group. Removal of the COPI-γ clade resulted in a tree topology similar to TPLATE with a weakly monophyletic TSAUCER clade that is weakly excluded from the AP-γαδεζ clade (Figure 5-8).

5.4.4.3 TCUP

TCUP is most similar to AP-μ and COPI-δ subunits, as confirmed by both structural prediction and comparative genomics. Phylogenetic analysis recovered TCUP as a paraphyletic clade, but was strongly excluded from the COPI-δ clade, and weakly excluded from the AP-μ clade (Figure 5-9). Removal of all COPI-δ sequences did not result in the formation of a single TCUP clade, as some TCUP sequences (*i.e., M. brevicollis, D. discoideum,* and *D. purpureum*) grouped within the AP5-μ clade, and the *C. reinhardtii* TCUP groups with AP4-μ, although with no support in either case (Figure 5-10). The curious topology reflected by these sequences is likely the result of low sequence conservation paired with LBA as TCUP and AP5-μ represent the longest branches in the analysis.

Figure 5-7. Phylogenetic analysis of TSAUCER indicates that TSET is distinct from COPI. Phylogenetic analysis of TSAUCER, γαδεζ-APs, and COPI-γ was carried out to determine the relationships between these subunits. The analysis identified a strongly supported COPI-γ clade that excludes TSAUCER and the APs. The APs are weakly monophyletic and exclude TSAUCER. TSAUCER sequences are weakly monophyletic except for the *M. pusilla* sequence which branches outside of the TSAUCER-AP clade. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (MrBayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95, closed light circles ≥ 0.95/75, open circles ≥ 0.8/50.

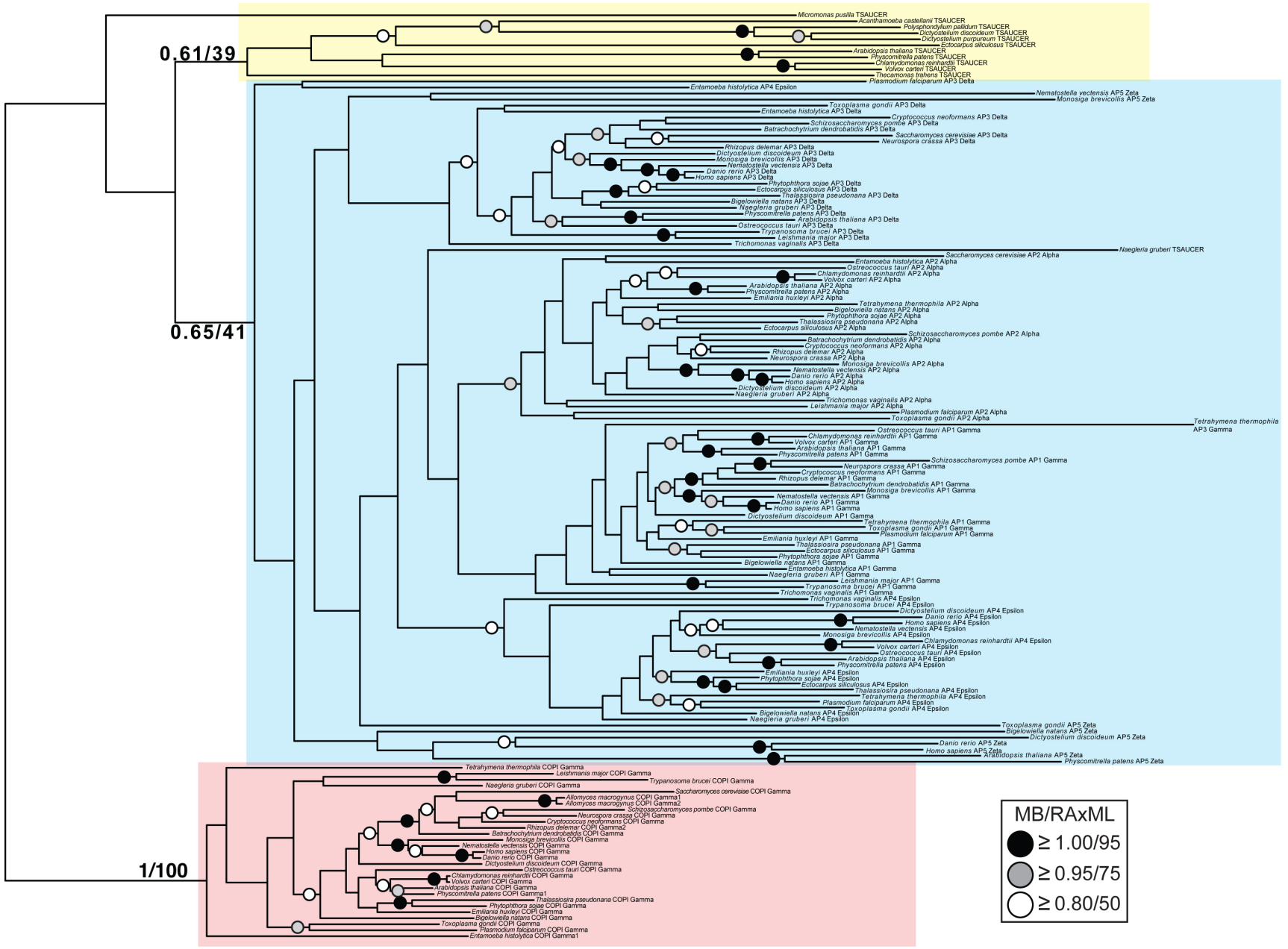

0.61/39
0.65/41
1/100
Micromonas pusilla TSAUCER
Acanthamoeba castellanii TSAUCER
Polysphondylium pallidum TSAUCER
Dictyostelium discoideum TSAUCER
Dictyostelium purpureum TSAUCER
Ectocarpus siliculosus TSAUCER
Arabidopsis thaliana TSAUCER
Physcomitrella patens TSAUCER
Chlamydomonas reinhardtii TSAUCER
Volvox carteri TSAUCER
Thecamonas trahens TSAUCER
Plasmodium falciparum AP3 Delta
Entamoeba histolytica AP4 Epsilon
Nematostella vectensis AP5 Zeta
Monosiga brevicollis AP5 Zeta
Toxoplasma gondii AP3 Delta
Entamoeba histolytica AP3 Delta
Cryptococcus neoformans AP3 Delta
Schizosaccharomyces pombe AP3 Delta
Batrachochytrium dendrobatidis AP3 Delta
Saccharomyces cerevisiae AP3 Delta
Neurospora crassa AP3 Delta
Rhizopus delemar AP3 Delta
Dictyostelium discoideum AP3 Delta
Monosiga brevicollis AP3 Delta
Nematostella vectensis AP3 Delta
Danio rerio AP3 Delta
Homo sapiens AP3 Delta
Phytophthora sojae AP3 Delta
Ectocarpus siliculosus AP3 Delta
Thalassiosira pseudonana AP3 Delta
Bigelowiella natans AP3 Delta
Naegleria gruberi AP3 Delta
Physcomitrella patens AP3 Delta
Arabidopsis thaliana AP3 Delta
Ostreococcus tauri AP3 Delta
Trypanosoma brucei AP3 Delta
Leishmania major AP3 Delta
Trichomonas vaginalis AP3 Delta
Naegleria gruberi TSAUCER
Saccharomyces cerevisiae AP2 Alpha
Entamoeba histolytica AP2 Alpha
Ostreococcus tauri AP2 Alpha
Chlamydomonas reinhardtii AP2 Alpha
Volvox carteri AP2 Alpha
Arabidopsis thaliana AP2 Alpha
Physcomitrella patens AP2 Alpha
Emiliania huxleyi AP2 Alpha
Tetrahymena thermophila AP2 Alpha
Bigelowiella natans AP2 Alpha
Phytophthora sojae AP2 Alpha
Thalassiosira pseudonana AP2 Alpha
Ectocarpus siliculosus AP2 Alpha
Schizosaccharomyces pombe AP2 Alpha
Batrachochytrium dendrobatidis AP2 Alpha
Cryptococcus neoformans AP2 Alpha
Rhizopus delemar AP2 Alpha
Neurospora crassa AP2 Alpha
Monosiga brevicollis AP2 Alpha
Nematostella vectensis AP2 Alpha
Danio rerio AP2 Alpha
Homo sapiens AP2 Alpha
Dictyostelium discoideum AP2 Alpha
Naegleria gruberi AP2 Alpha
Trichomonas vaginalis AP2 Alpha
Leishmania major AP2 Alpha
Plasmodium falciparum AP2 Alpha
Toxoplasma gondii AP2 Alpha
Tetrahymena thermophila AP3 Gamma
Ostreococcus tauri AP1 Gamma
Chlamydomonas reinhardtii AP1 Gamma
Volvox carteri AP1 Gamma
Arabidopsis thaliana AP1 Gamma
Physcomitrella patens AP1 Gamma
Schizosaccharomyces pombe AP1 Gamma
Neurospora crassa AP1 Gamma
Cryptococcus neoformans AP1 Gamma
Rhizopus delemar AP1 Gamma
Batrachochytrium dendrobatidis AP1 Gamma
Monosiga brevicollis AP1 Gamma
Nematostella vectensis AP1 Gamma
Danio rerio AP1 Gamma
Homo sapiens AP1 Gamma
Dictyostelium discoideum AP1 Gamma
Tetrahymena thermophila AP1 Gamma
Toxoplasma gondii AP1 Gamma
Plasmodium falciparum AP1 Gamma
Emiliania huxleyi AP1 Gamma
Thalassiosira pseudonana AP1 Gamma
Ectocarpus siliculosus AP1 Gamma
Phytophthora sojae AP1 Gamma
Bigelowiella natans AP1 Gamma
Entamoeba histolytica AP1 Gamma
Naegleria gruberi AP1 Gamma
Leishmania major AP1 Gamma
Trypanosoma brucei AP1 Gamma
Trichomonas vaginalis AP1 Gamma
Trichomonas vaginalis AP4 Epsilon
Trypanosoma brucei AP4 Epsilon
Dictyostelium discoideum AP4 Epsilon
Danio rerio AP4 Epsilon
Homo sapiens AP4 Epsilon
Nematostella vectensis AP4 Epsilon
Monosiga brevicollis AP4 Epsilon
Chlamydomonas reinhardtii AP4 Epsilon
Volvox carteri AP4 Epsilon
Ostreococcus tauri AP4 Epsilon
Arabidopsis thaliana AP4 Epsilon
Physcomitrella patens AP4 Epsilon
Emiliania huxleyi AP4 Epsilon
Phytophthora sojae AP4 Epsilon
Ectocarpus siliculosus AP4 Epsilon
Thalassiosira pseudonana AP4 Epsilon
Tetrahymena thermophila AP4 Epsilon
Plasmodium falciparum AP4 Epsilon
Toxoplasma gondii AP4 Epsilon
Bigelowiella natans AP4 Epsilon
Naegleria gruberi AP4 Epsilon
Toxoplasma gondii AP5 Zeta
Bigelowiella natans AP5 Zeta
Dictyostelium discoideum AP5 Zeta
Danio rerio AP5 Zeta
Homo sapiens AP5 Zeta
Arabidopsis thaliana AP5 Zeta
Physcomitrella patens AP5 Zeta
Tetrahymena thermophila COPI Gamma
Leishmania major COPI Gamma
Trypanosoma brucei COPI Gamma
Naegleria gruberi COPI Gamma
Saccharomyces cerevisiae COPI Gamma
Allomyces macrogynus COPI Gamma1
Allomyces macrogynus COPI Gamma2
Schizosaccharomyces pombe COPI Gamma
Neurospora crassa COPI Gamma
Cryptococcus neoformans COPI Gamma
Rhizopus delemar COPI Gamma2
Batrachochytrium dendrobatidis COPI Gamma
Monosiga brevicollis COPI Gamma
Nematostella vectensis COPI Gamma
Homo sapiens COPI Gamma
Danio rerio COPI Gamma
Dictyostelium discoideum COPI Gamma
Ostreococcus tauri COPI Gamma
Chlamydomonas reinhardtii COPI Gamma
Volvox carteri COPI Gamma
Arabidopsis thaliana COPI Gamma
Physcomitrella patens COPI Gamma1
Thalassiosira pseudonana COPI Gamma
Phytophthora sojae COPI Gamma
Emiliania huxleyi COPI Gamma
Bigelowiella natans COPI Gamma
Toxoplasma gondii COPI Gamma
Plasmodium falciparum COPI Gamma
Entamoeba histolytica COPI Gamma1
MB/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50
0.3

Figure 5-8. Phylogenetic analysis suggests that TSAUCER is excluded from the AP clade. To determine if TSAUCER and the γαδεζ-APs form mutually exclusive monophyletic groups, all COPI-γ sequences were removed from the data set in Figure 5-7. Re-analyzing the data set increased the support for a monophyletic TSAUCER and γαδεζ-AP clades, although node support was still very weak for both groups. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (MrBayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95, closed light circles ≥ 0.95/75, open circles ≥ 0.8/50.

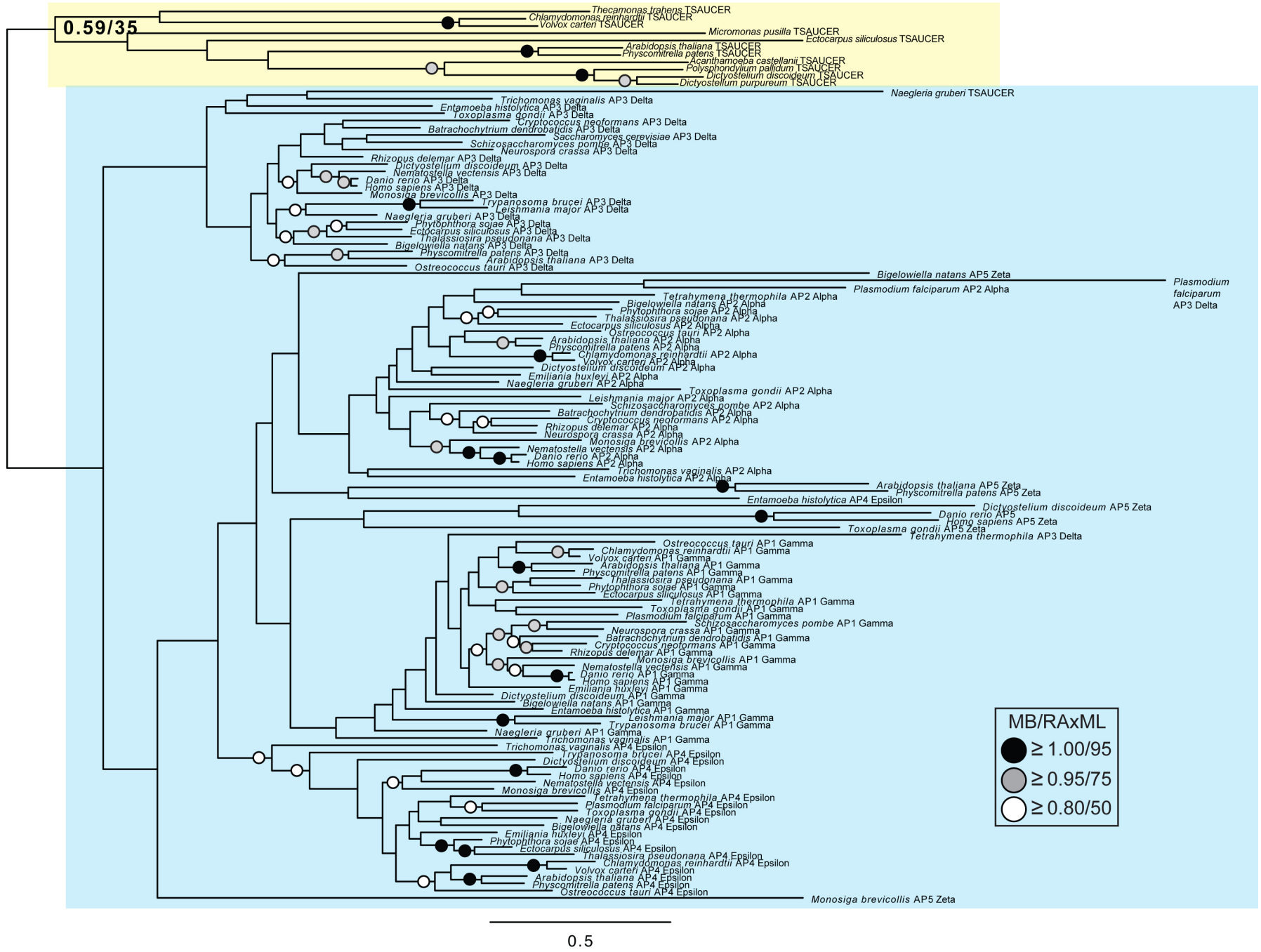

0.59/35
Thecamonas trahens TSAUCER
Chlamydomonas reinhardtii TSAUCER
Volvox carteri TSAUCER
Micromonas pusilla TSAUCER
Ectocarpus siliculosus TSAUCER
Arabidopsis thaliana TSAUCER
Physcomitrella patens TSAUCER
Acanthamoeba castellanii TSAUCER
Polysphondylium pallidum TSAUCER
Dictyostelium discoideum TSAUCER
Dictyostelium purpureum TSAUCER
Naegleria gruberi TSAUCER
Trichomonas vaginalis AP3 Delta
Entamoeba histolytica AP3 Delta
Toxoplasma gondii AP3 Delta
Cryptococcus neoformans AP3 Delta
Batrachochytrium dendrobatidis AP3 Delta
Saccharomyces cerevisiae AP3 Delta
Schizosaccharomyces pombe AP3 Delta
Neurospora crassa AP3 Delta
Rhizopus delemar AP3 Delta
Dictyostelium discoideum AP3 Delta
Nematostella vectensis AP3 Delta
Danio rerio AP3 Delta
Homo sapiens AP3 Delta
Monosiga brevicollis AP3 Delta
Trypanosoma brucei AP3 Delta
Leishmania major AP3 Delta
Naegleria gruberi AP3 Delta
Phytophthora sojae AP3 Delta
Ectocarpus siliculosus AP3 Delta
Thalassiosira pseudonana AP3 Delta
Bigelowiella natans AP3 Delta
Physcomitrella patens AP3 Delta
Arabidopsis thaliana AP3 Delta
Ostreococcus tauri AP3 Delta
Bigelowiella natans AP5 Zeta
Plasmodium falciparum AP2 Alpha
Plasmodium falciparum AP3 Delta
Tetrahymena thermophila AP2 Alpha
Bigelowiella natans AP2 Alpha
Phytophthora sojae AP2 Alpha
Thalassiosira pseudonana AP2 Alpha
Ectocarpus siliculosus AP2 Alpha
Ostreococcus tauri AP2 Alpha
Arabidopsis thaliana AP2 Alpha
Physcomitrella patens AP2 Alpha
Chlamydomonas reinhardtii AP2 Alpha
Volvox carteri AP2 Alpha
Dictyostelium discoideum AP2 Alpha
Emiliania huxleyi AP2 Alpha
Naegleria gruberi AP2 Alpha
Toxoplasma gondii AP2 Alpha
Leishmania major AP2 Alpha
Schizosaccharomyces pombe AP2 Alpha
Batrachochytrium dendrobatidis AP2 Alpha
Cryptococcus neoformans AP2 Alpha
Rhizopus delemar AP2 Alpha
Neurospora crassa AP2 Alpha
Monosiga brevicollis AP2 Alpha
Nematostella vectensis AP2 Alpha
Danio rerio AP2 Alpha
Homo sapiens AP2 Alpha
Trichomonas vaginalis AP2 Alpha
Entamoeba histolytica AP2 Alpha
Arabidopsis thaliana AP5 Zeta
Physcomitrella patens AP5 Zeta
Entamoeba histolytica AP4 Epsilon
Dictyostelium discoideum AP5 Zeta
Danio rerio AP5
Homo sapiens AP5 Zeta
Toxoplasma gondii AP5 Zeta
Tetrahymena thermophila AP3 Delta
Ostreococcus tauri AP1 Gamma
Chlamydomonas reinhardtii AP1 Gamma
Volvox carteri AP1 Gamma
Arabidopsis thaliana AP1 Gamma
Physcomitrella patens AP1 Gamma
Thalassiosira pseudonana AP1 Gamma
Phytophthora sojae AP1 Gamma
Ectocarpus siliculosus AP1 Gamma
Tetrahymena thermophila AP1 Gamma
Toxoplasma gondii AP1 Gamma
Plasmodium falciparum AP1 Gamma
Schizosaccharomyces pombe AP1 Gamma
Neurospora crassa AP1 Gamma
Batrachochytrium dendrobatidis AP1 Gamma
Cryptococcus neoformans AP1 Gamma
Rhizopus delemar AP1 Gamma
Monosiga brevicollis AP1 Gamma
Nematostella vectensis AP1 Gamma
Danio rerio AP1 Gamma
Homo sapiens AP1 Gamma
Emiliania huxleyi AP1 Gamma
Dictyostelium discoideum AP1 Gamma
Bigelowiella natans AP1 Gamma
Entamoeba histolytica AP1 Gamma
Leishmania major AP1 Gamma
Trypanosoma brucei AP1 Gamma
Naegleria gruberi AP1 Gamma
Trichomonas vaginalis AP1 Gamma
Trichomonas vaginalis AP4 Epsilon
Trypanosoma brucei AP4 Epsilon
Dictyostelium discoideum AP4 Epsilon
Danio rerio AP4 Epsilon
Homo sapiens AP4 Epsilon
Nematostella vectensis AP4 Epsilon
Monosiga brevicollis AP4 Epsilon
Tetrahymena thermophila AP4 Epsilon
Plasmodium falciparum AP4 Epsilon
Toxoplasma gondii AP4 Epsilon
Naegleria gruberi AP4 Epsilon
Bigelowiella natans AP4 Epsilon
Emiliania huxleyi AP4 Epsilon
Phytophthora sojae AP4 Epsilon
Ectocarpus siliculosus AP4 Epsilon
Thalassiosira pseudonana AP4 Epsilon
Chlamydomonas reinhardtii AP4 Epsilon
Volvox carteri AP4 Epsilon
Arabidopsis thaliana AP4 Epsilon
Physcomitrella patens AP4 Epsilon
Ostreococcus tauri AP4 Epsilon
Monosiga brevicollis AP5 Zeta
MB/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50
0.5

Figure 5-9. Phylogenetic analysis of TCUP suggests that TSET is distinct from COPI. To determine the relationship between TCUP, COPI-δ, and AP-μ subunits, phylogenetic analysis was undertaken. Phylogenetic analysis shows a very strong COPI-δ clade and an unsupported AP-μ clade. TCUP is not unified, with two separate groups and some TCUP sequences branching within the AP-μ clade. The results indicate that TCUP does not group within the COPI-δ clade. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (MrBayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95, closed light circles ≥ 0.95/75, open circles ≥ 0.8/50.

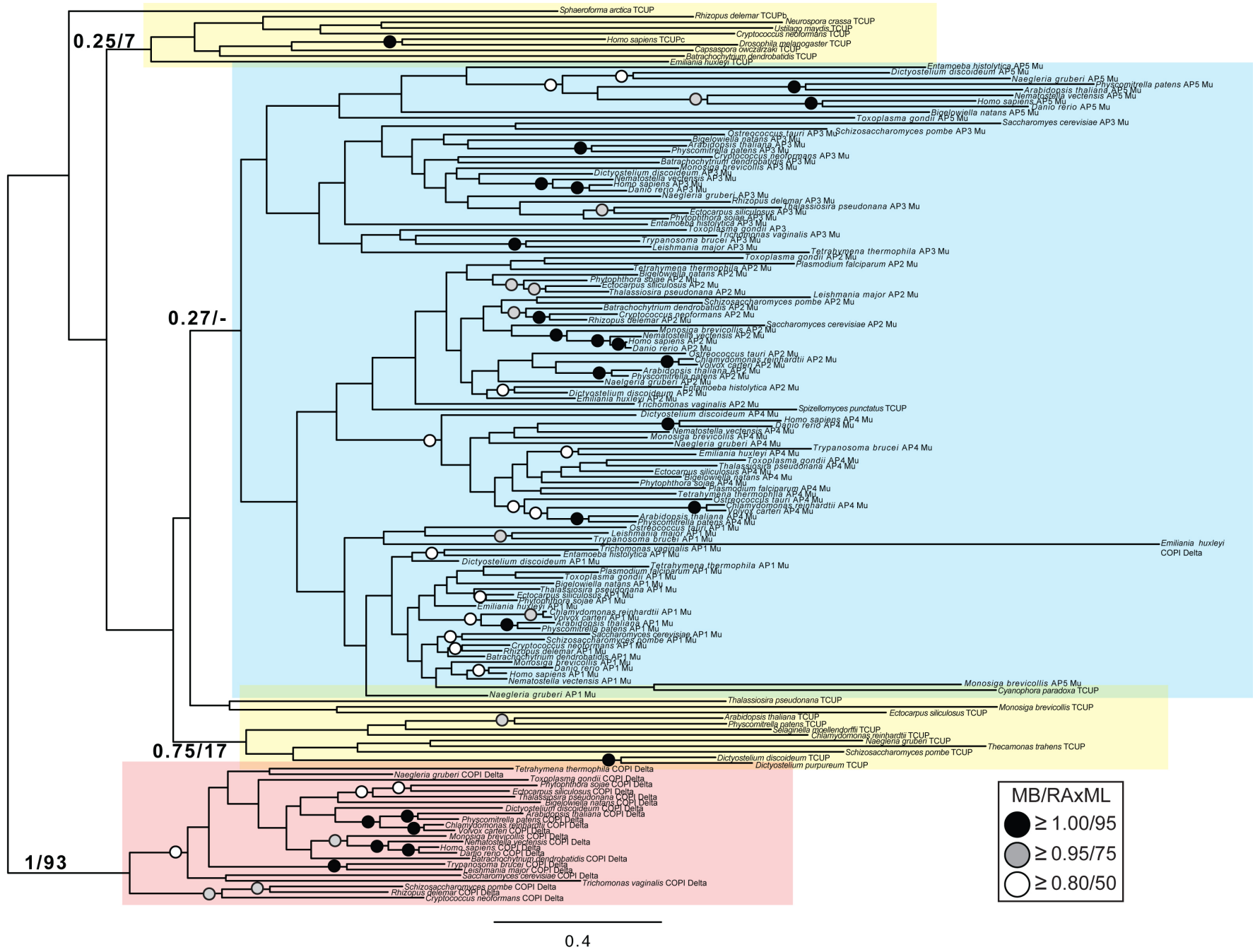
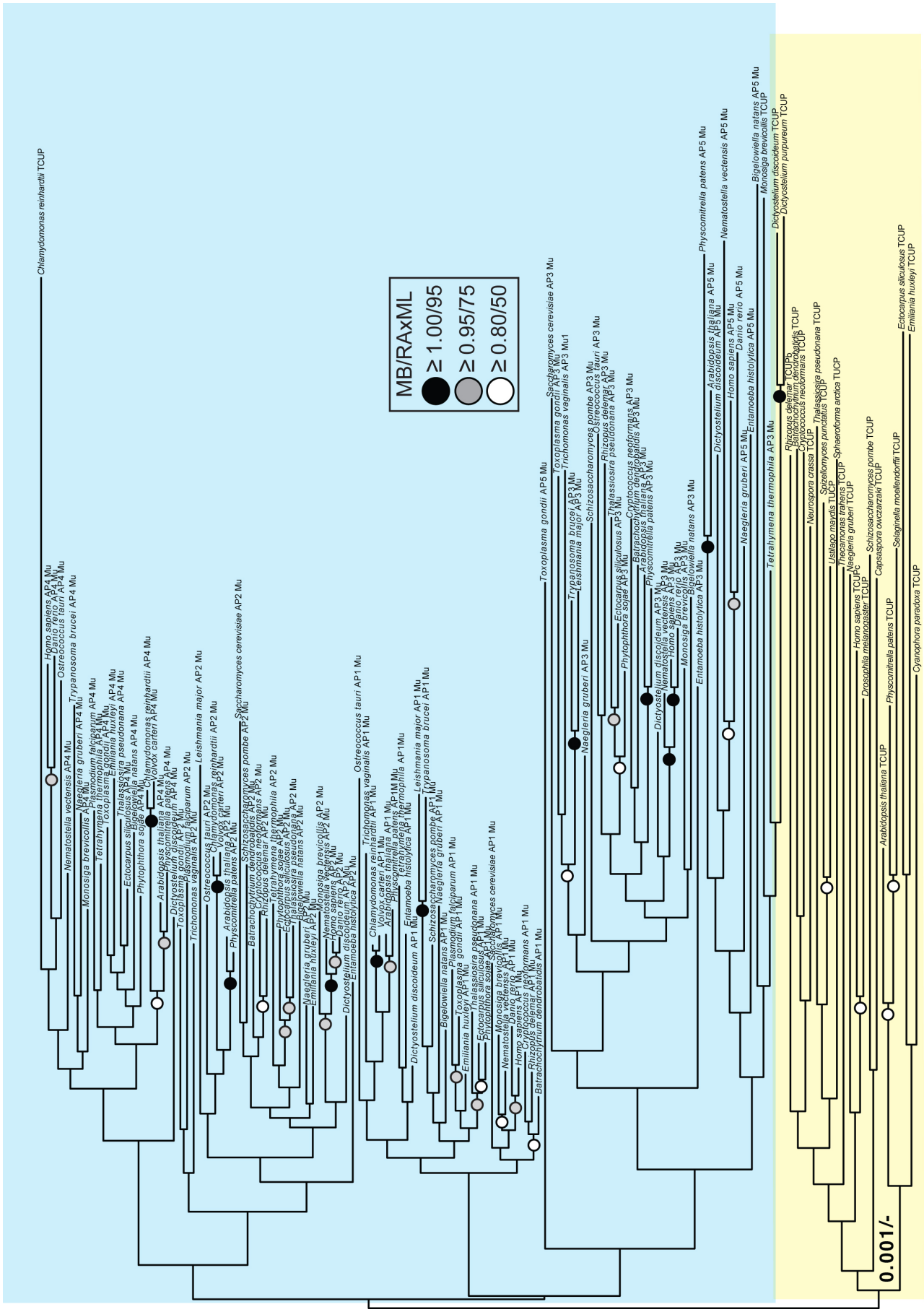

0.25/7
0.27/-
0.75/17
1/93
Sphaeroforma arctica TCUP
Rhizopus delemar TCUPb
Neurospora crassa TCUP
Ustilago maydis TCUP
Cryptococcus neoformans TCUP
Homo sapiens TCUPc
Drosophila melanogaster TCUP
Capsaspora owczarzaki TCUP
Batrachochytrium dendrobatidis TCUP
Emiliania huxleyi TCUP
Entamoeba histolytica AP5 Mu
Dictyostelium discoideum AP5 Mu
Naegleria gruberi AP5 Mu
Physcomitrella patens AP5 Mu
Arabidopsis thaliana AP5 Mu
Nematostella vectensis AP5 Mu
Homo sapiens AP5 Mu
Danio rerio AP5 Mu
Bigelowiella natans AP5 Mu
Toxoplasma gondii AP5 Mu
Saccharomyces cerevisiae AP3 Mu
Schizosaccharomyces pombe AP3 Mu
Ostreococcus tauri AP3 Mu
Bigelowiella natans AP3 Mu
Arabidopsis thaliana AP3 Mu
Physcomitrella patens AP3 Mu
Cryptococcus neoformans AP3 Mu
Batrachochytrium dendrobatidis AP3 Mu
Monosiga brevicollis AP3 Mu
Dictyostelium discoideum AP3 Mu
Nematostella vectensis AP3 Mu
Homo sapiens AP3 Mu
Danio rerio AP3 Mu
Naegleria gruberi AP3 Mu
Rhizopus delemar AP3 Mu
Thalassiosira pseudonana AP3 Mu
Ectocarpus siliculosus AP3 Mu
Phytophthora sojae AP3 Mu
Entamoeba histolytica AP3 Mu
Toxoplasma gondii AP3
Trichomonas vaginalis AP3 Mu
Trypanosoma brucei AP3 Mu
Leishmania major AP3 Mu
Tetrahymena thermophila AP3 Mu
Toxoplasma gondii AP2 Mu
Plasmodium falciparum AP2 Mu
Tetrahymena thermophila AP2 Mu
Bigelowiella natans AP2 Mu
Phytophthora sojae AP2 Mu
Ectocarpus siliculosus AP2 Mu
Thalassiosira pseudonana AP2 Mu
Leishmania major AP2 Mu
Schizosaccharomyces pombe AP2 Mu
Batrachochytrium dendrobatidis AP2 Mu
Cryptococcus neoformans AP2 Mu
Rhizopus delemar AP2 Mu
Saccharomyces cerevisiae AP2 Mu
Monosiga brevicollis AP2 Mu
Nematostella vectensis AP2 Mu
Homo sapiens AP2 Mu
Danio rerio AP2 Mu
Ostreococcus tauri AP2 Mu
Chlamydomonas reinhardtii AP2 Mu
Volvox carteri AP2 Mu
Arabidopsis thaliana AP2 Mu
Physcomitrella patens AP2 Mu
Naelgeria gruberi AP2 Mu
Entamoeba histolytica AP2 Mu
Dictyostelium discoideum AP2 Mu
Emiliania huxleyi AP2 Mu
Trichomonas vaginalis AP2 Mu
Spizellomyces punctatus TCUP
Dictyostelium discoideum AP4 Mu
Homo sapiens AP4 Mu
Danio rerio AP4 Mu
Nematostella vectensis AP4 Mu
Monosiga brevicollis AP4 Mu
Naegleria gruberi AP4 Mu
Trypanosoma brucei AP4 Mu
Emiliania huxleyi AP4 Mu
Toxoplasma gondii AP4 Mu
Thalassiosira pseudonana AP4 Mu
Ectocarpus siliculosus AP4 Mu
Bigelowiella natans AP4 Mu
Phytophthora sojae AP4 Mu
Plasmodium falciparum AP4 Mu
Tetrahymena thermophila AP4 Mu
Ostreococcus tauri AP4 Mu
Chlamydomonas reinhardtii AP4 Mu
Volvox carteri AP4 Mu
Arabidopsis thaliana AP4 Mu
Physcomitrella patens AP4 Mu
Ostreococcus tauri AP1 Mu
Leishmania major AP1 Mu
Trypanosoma brucei AP1 Mu
Emiliania huxleyi COPI Delta
Trichomonas vaginalis AP1 Mu
Entamoeba histolytica AP1 Mu
Dictyostelium discoideum AP1 Mu
Tetrahymena thermophila AP1 Mu
Plasmodium falciparum AP1 Mu
Toxoplasma gondii AP1 Mu
Bigelowiella natans AP1 Mu
Thalassiosira pseudonana AP1 Mu
Ectocarpus siliculosus AP1 Mu
Phytophthora sojae AP1 Mu
Emiliania huxleyi AP1 Mu
Chlamydomonas reinhardtii AP1 Mu
Volvox carteri AP1 Mu
Arabidopsis thaliana AP1 Mu
Physcomitrella patens AP1 Mu
Saccharomyces cerevisiae AP1 Mu
Schizosaccharomyces pombe AP1 Mu
Cryptococcus neoformans AP1 Mu
Rhizopus delemar AP1 Mu
Batrachochytrium dendrobatidis AP1 Mu
Monosiga brevicollis AP1 Mu
Danio rerio AP1 Mu
Homo sapiens AP1 Mu
Nematostella vectensis AP1 Mu
Monosiga brevicollis AP5 Mu
Cyanophora paradoxa TCUP
Naegleria gruberi AP1 Mu
Thalassiosira pseudonana TCUP
Monosiga brevicollis TCUP
Ectocarpus siliculosus TCUP
Arabidopsis thaliana TCUP
Physcomitrella patens TCUP
Selaginella moellendorffii TCUP
Chlamydomonas reinhardtii TCUP
Naegleria gruberi TCUP
Thecamonas trahens TCUP
Schizosaccharomyces pombe TCUP
Dictyostelium discoideum TCUP
Dictyostelium purpureum TCUP
Tetrahymena thermophila COPI Delta
Naegleria gruberi COPI Delta
Toxoplasma gondii COPI Delta
Phytophthora sojae COPI Delta
Ectocarpus siliculosus COPI Delta
Thalassiosira pseudonana COPI Delta
Bigelowiella natans COPI Delta
Dictyostelium discoideum COPI Delta
Arabidopsis thaliana COPI Delta
Physcomitrella patens COPI Delta
Chlamydomonas reinhardtii COPI Delta
Volvox carteri COPI Delta
Monosiga brevicollis COPI Delta
Nematostella vectensis COPI Delta
Homo sapiens COPI Delta
Danio rerio COPI Delta
Batrachochytrium dendrobatidis COPI Delta
Trypanosoma brucei COPI Delta
Leishmania major COPI Delta
Saccharomyces cerevisiae COPI Delta
Trichomonas vaginalis COPI Delta
Schizosaccharomyces pombe COPI Delta
Rhizopus delemar COPI Delta
Cryptococcus neoformans COPI Delta
MB/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50
0.4

Figure 5-10. Phylogenetic analysis weakly suggests that TCUP may be excluded from the AP clade. To determine if TCUP and the AP-μ sequences formed mutually exclusive clades, all COPI-δ sequences were removed from the data set used to generate Figure 5-9. Removal of the COPI-δ sequences resulted in most TCUP sequences forming a single group, with some outliers still branching with AP-μ sequences (*C. reinhardtii* TCUP, *M. brevicollis* TCUP), although this topology is unsupported. This topology is at least partially the product of LBA, as these sequences represent some of the longest branches in the analysis. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (MrBayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed dark circles ≥ 1.00/95, closed light circles ≥ 0.95/75, open circles ≥ 0.8/50.


MB/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50
0.001/-
0.4

5.4.4.4 TSPOON

Comparative genomics and structural prediction indicated that the closest homologues of TSPOON are the σ and ζ subunits of APs and COPI, respectively. Unlike the other TSET subunits, the TSPOON tree resolved at nearly every major node (Figure 5-11). TSPOON was clearly excluded from both the COPI and AP clades, suggesting that it might be a distinct lineage from these other coat complexes.

5.4.4.5 Concatenated Phylogeny

Phylogenetic analysis of individual subunits displayed generally the same pattern, with TSET excluded from both COPI and the AP clades but with little node support. An analysis providing stronger, more robust phylogenetic signal was desired to fully tease out the relationships between TSET, COPI, and the APs. Therefore, the four subunits of the heterotetramer were concatenated in order to utilize all of the sequence information possible. Alignments from the analyses above were concatenated and used for phylogenetic inference. The resulting tree was exceptionally well resolved, and consistent with the phylogenetic signal observed for the individual subunits; TSET was excluded from both the COPI and the AP clades (Figure 5-12). The phylogeny in Figure 5-12 is unrooted, therefore we cannot determine the branching order of COPI, TSET and the APs. However, a recent analysis of longin domain-containing proteins suggests a root with COPI and TSET together on one side and the Adaptins on the other side, (C. Klinger, personal communication) although further analyses are required to confirm this topology. Nevertheless, the tree does indicate that TSET is an ancient complex distinct from

Figure 5-11. Phylogenetic analysis indicates that TSPOON branches separately from both COPI and the APs. To determine the relationship between TSPOON, COPI-ζ, and AP-σ, these sequences were analyzed using phylogenetic methods. Analysis recovers a moderately supported, monophyletic TSPOON clade that is excluded from moderately supported COPI and AP clades, indicating that it neither branches within the COPI-ζ clade, nor does it branch with in the AP-σ clade. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (MrBayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed, dark circles ≥ 1.00/95; closed, light circles ≥ 0.95/75; open circles ≥ 0.8/50.

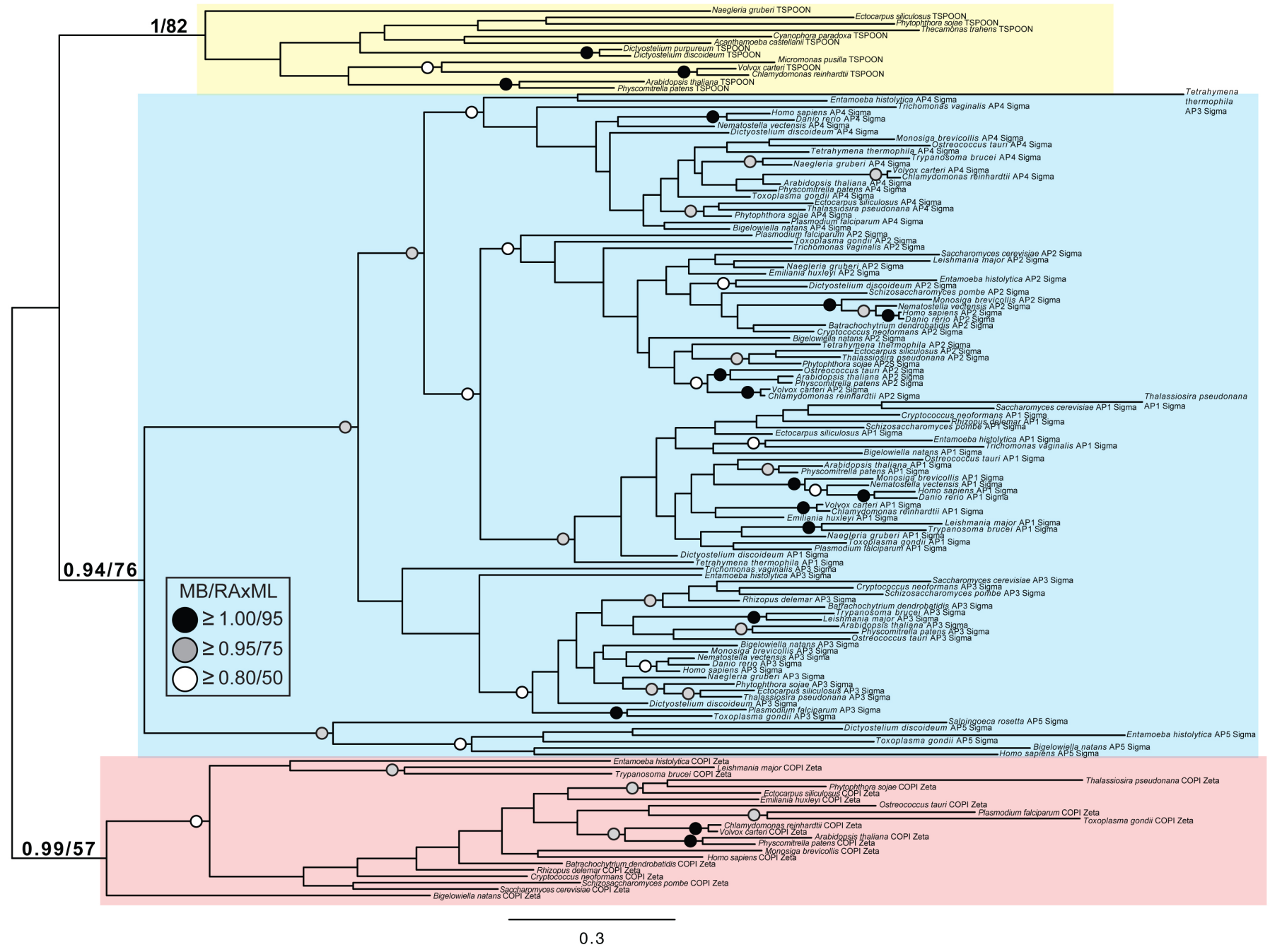
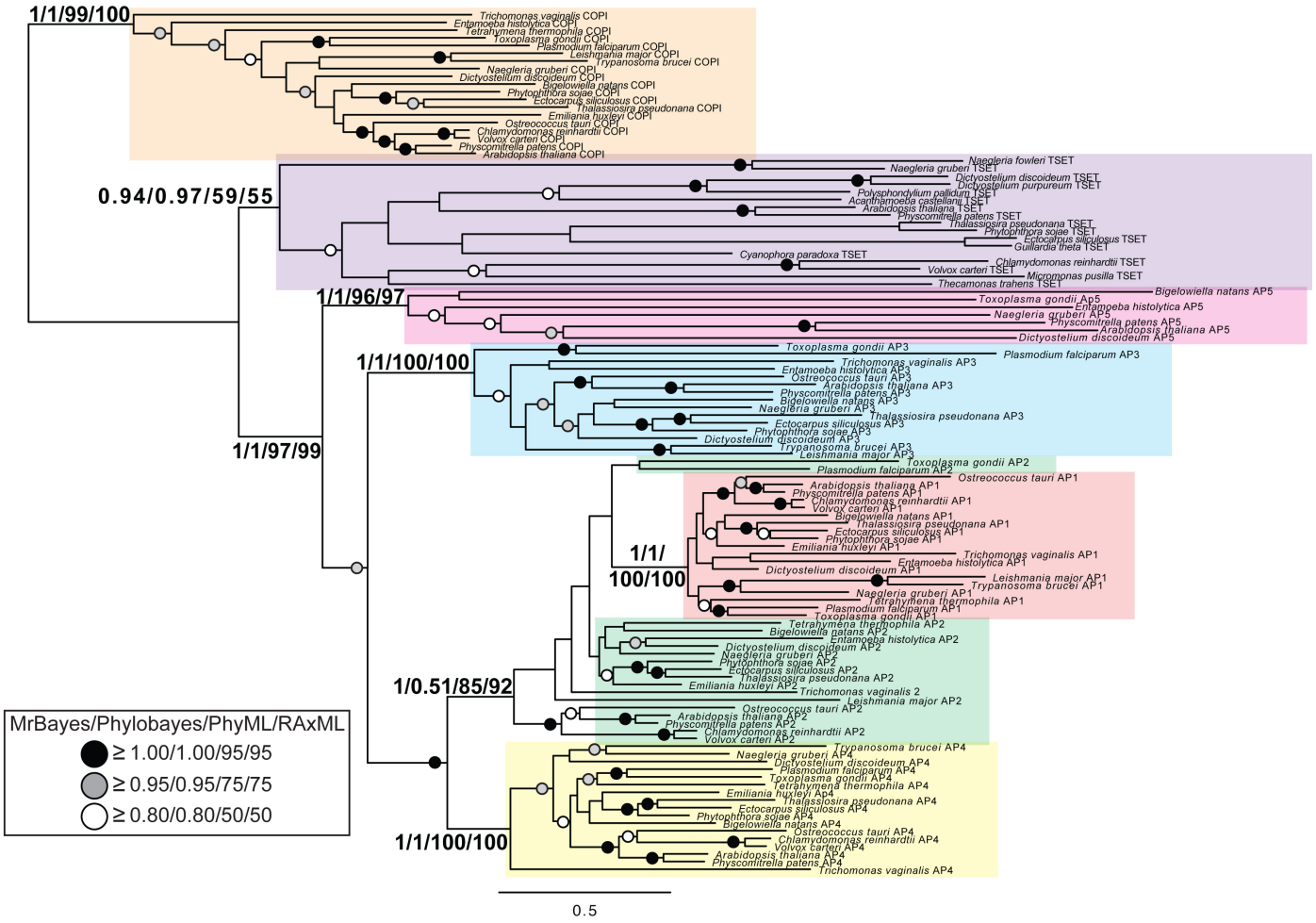

1/82
0.94/76
0.99/57
MB/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50
0.3
Naegleria gruberi TSPOON
Ectocarpus siliculosus TSPOON
Phytophthora sojae TSPOON
Thecamonas trahens TSPOON
Cyanophora paradoxa TSPOON
Acanthamoeba castellanii TSPOON
Dictyostelium purpureum TSPOON
Dictyostelium discoideum TSPOON
Micromonas pusilla TSPOON
Volvox carteri TSPOON
Chlamydomonas reinhardtii TSPOON
Arabidopsis thaliana TSPOON
Physcomitrella patens TSPOON
Tetrahymena thermophila AP3 Sigma
Entamoeba histolytica AP4 Sigma
Trichomonas vaginalis AP4 Sigma
Homo sapiens AP4 Sigma
Danio rerio AP4 Sigma
Nematostella vectensis AP4 Sigma
Dictyostelium discoideum AP4 Sigma
Monosiga brevicollis AP4 Sigma
Ostreococcus tauri AP4 Sigma
Tetrahymena thermophila AP4 Sigma
Trypanosoma brucei AP4 Sigma
Naegleria gruberi AP4 Sigma
Volvox carteri AP4 Sigma
Chlamydomonas reinhardtii AP4 Sigma
Arabidopsis thaliana AP4 Sigma
Physcomitrella patens AP4 Sigma
Toxoplasma gondii AP4 Sigma
Ectocarpus siliculosus AP4 Sigma
Thalassiosira pseudonana AP4 Sigma
Phytophthora sojae AP4 Sigma
Plasmodium falciparum AP4 Sigma
Bigelowiella natans AP4 Sigma
Plasmodium falciparum AP2 Sigma
Toxoplasma gondii AP2 Sigma
Trichomonas vaginalis AP2 Sigma
Saccharomyces cerevisiae AP2 Sigma
Leishmania major AP2 Sigma
Naegleria gruberi AP2 Sigma
Emiliania huxleyi AP2 Sigma
Entamoeba histolytica AP2 Sigma
Dictyostelium discoideum AP2 Sigma
Schizosaccharomyces pombe AP2 Sigma
Monosiga brevicollis AP2 Sigma
Nematostella vectensis AP2 Sigma
Homo sapiens AP2 Sigma
Danio rerio AP2 Sigma
Batrachochytrium dendrobatidis AP2 Sigma
Cryptococcus neoformans AP2 Sigma
Bigelowiella natans AP2 Sigma
Tetrahymena thermophila AP2 Sigma
Ectocarpus siliculosus AP2 Sigma
Thalassiosira pseudonana AP2 Sigma
Phytophthora sojae AP2S Sigma
Ostreococcus tauri AP2 Sigma
Arabidopsis thaliana AP2 Sigma
Physcomitrella patens AP2 Sigma
Volvox carteri AP2 Sigma
Chlamydomonas reinhardtii AP2 Sigma
Thalassiosira pseudonana AP1 Sigma
Saccharomyces cerevisiae AP1 Sigma
Cryptococcus neoformans AP1 Sigma
Rhizopus delemar AP1 Sigma
Schizosaccharomyces pombe AP1 Sigma
Ectocarpus siliculosus AP1 Sigma
Entamoeba histolytica AP1 Sigma
Trichomonas vaginalis AP1 Sigma
Bigelowiella natans AP1 Sigma
Ostreococcus tauri AP1 Sigma
Arabidopsis thaliana AP1 Sigma
Physcomitrella patens AP1 Sigma
Monosiga brevicollis AP1 Sigma
Nematostella vectensis AP1 Sigma
Homo sapiens AP1 Sigma
Danio rerio AP1 Sigma
Volvox carteri AP1 Sigma
Chlamydomonas reinhardtii AP1 Sigma
Emiliania huxleyi AP1 Sigma
Leishmania major AP1 Sigma
Trypanosoma brucei AP1 Sigma
Naegleria gruberi AP1 Sigma
Toxoplasma gondii AP1 Sigma
Plasmodium falciparum AP1 Sigma
Dictyostelium discoideum AP1 Sigma
Tetrahymena thermophila AP1 Sigma
Trichomonas vaginalis AP3 Sigma
Entamoeba histolytica AP3 Sigma
Saccharomyces cerevisiae AP3 Sigma
Cryptococcus neoformans AP3 Sigma
Schizosaccharomyces pombe AP3 Sigma
Rhizopus delemar AP3 Sigma
Batrachochytrium dendrobatidis AP3 Sigma
Trypanosoma brucei AP3 Sigma
Leishmania major AP3 Sigma
Arabidopsis thaliana AP3 Sigma
Physcomitrella patens AP3 Sigma
Ostreococcus tauri AP3 Sigma
Bigelowiella natans AP3 Sigma
Monosiga brevicollis AP3 Sigma
Nematostella vectensis AP3 Sigma
Danio rerio AP3 Sigma
Homo sapiens AP3 Sigma
Naegleria gruberi AP3 Sigma
Phytophthora sojae AP3 Sigma
Ectocarpus siliculosus AP3 Sigma
Thalassiosira pseudonana AP3 Sigma
Dictyostelium discoideum AP3 Sigma
Plasmodium falciparum AP3 Sigma
Toxoplasma gondii AP3 Sigma
Salpingoeca rosetta AP5 Sigma
Dictyostelium discoideum AP5 Sigma
Entamoeba histolytica AP5 Sigma
Toxoplasma gondii AP5 Sigma
Bigelowiella natans AP5 Sigma
Homo sapiens AP5 Sigma
Entamoeba histolytica COPI Zeta
Leishmania major COPI Zeta
Trypanosoma brucei COPI Zeta
Thalassiosira pseudonana COPI Zeta
Phytophthora sojae COPI Zeta
Ectocarpus siliculosus COPI Zeta
Emiliania huxleyi COPI Zeta
Ostreococcus tauri COPI Zeta
Plasmodium falciparum COPI Zeta
Toxoplasma gondii COPI Zeta
Chlamydomonas reinhardtii COPI Zeta
Volvox carteri COPI Zeta
Arabidopsis thaliana COPI Zeta
Physcomitrella patens COPI Zeta
Monosiga brevicollis COPI Zeta
Homo sapiens COPI Zeta
Batrachochytrium dendrobatidis COPI Zeta
Rhizopus delemar COPI Zeta
Cryptococcus neoformans COPI Zeta
Schizosaccharomyces pombe COPI Zeta
Saccharomyces cerevisiae COPI Zeta
Bigelowiella natans COPI Zeta

Figure 5-12. TSET is a distinct lineage from the F-COPI and the AP complexes. Concatenated phylogenetic analysis of heterotetrameric complexes F-COPI (orange), TSET (purple), and APs (AP-5 is magenta, AP-3 is blue, AP-1 is red, AP-2 is green, and AP-4 is yellow), shows strong support for COPI, weak support for TSET, and strong support for the entire AP clade, indicating that TSET does not branch with in the COPI clade nor does it branch within the APs. This analysis also resolved the branching order of each AP complex. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (MrBayes)/ Bayesian posterior probabilities (Phylobayes)/Maximum-Likelihood bootstrap values (PhyML)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed, dark circles ≥ 1.00/1.00/0.95/95; closed, light circles ≥ 0.95/0.95/75/75; open circles ≥ 0.80/0.80/50/50.


1/1/99/100
Trichomonas vaginalis COPI
Entamoeba histolytica COPI
Tetrahymena thermophila COPI
Toxoplasma gondii COPI
Plasmodium falciparum COPI
Leishmania major COPI
Trypanosoma brucei COPI
Naegleria gruberi COPI
Dictyostelium discoideum COPI
Bigelowiella natans COPI
Phytophthora sojae COPI
Ectocarpus siliculosus COPI
Thalassiosira pseudonana COPI
Emiliania huxleyi COPI
Ostreococcus tauri COPI
Chlamydomonas reinhardtii COPI
Volvox carteri COPI
Physcomitrella patens COPI
Arabidopsis thaliana COPI
0.94/0.97/59/55
Naegleria fowleri TSET
Naegleria gruberi TSET
Dictyostelium discoideum TSET
Dictyostelium purpureum TSET
Polysphondylium pallidum TSET
Acanthamoeba castellanii TSET
Arabidopsis thaliana TSET
Physcomitrella patens TSET
Thalassiosira pseudonana TSET
Phytophthora sojae TSET
Ectocarpus siliculosus TSET
Guillardia theta TSET
Cyanophora paradoxa TSET
Chlamydomonas reinhardtii TSET
Volvox carteri TSET
Micromonas pusilla TSET
Thecamonas trahens TSET
1/1/96/97
Bigelowiella natans AP5
Toxoplasma gondii AP5
Entamoeba histolytica AP5
Naegleria gruberi AP5
Physcomitrella patens AP5
Arabidopsis thaliana AP5
Dictyostelium discoideum AP5
1/1/100/100
Toxoplasma gondii AP3
Plasmodium falciparum AP3
Trichomonas vaginalis AP3
Entamoeba histolytica AP3
Ostreococcus tauri AP3
Arabidopsis thaliana AP3
Physcomitrella patens AP3
Bigelowiella natans AP3
Naegleria gruberi AP3
Thalassiosira pseudonana AP3
Ectocarpus siliculosus AP3
Phytophthora sojae AP3
Dictyostelium discoideum AP3
Trypanosoma brucei AP3
Leishmania major AP3
1/1/97/99
Toxoplasma gondii AP2
Plasmodium falciparum AP2
Ostreococcus tauri AP1
Arabidopsis thaliana AP1
Physcomitrella patens AP1
Chlamydomonas reinhardtii AP1
Volvox carteri AP1
Bigelowiella natans AP1
Thalassiosira pseudonana AP1
Ectocarpus siliculosus AP1
Phytophthora sojae AP1
Emiliania huxleyi AP1
1/1/
100/100
Trichomonas vaginalis AP1
Entamoeba histolytica AP1
Dictyostelium discoideum AP1
Leishmania major AP1
Trypanosoma brucei AP1
Naegleria gruberi AP1
Tetrahymena thermophila AP1
Plasmodium falciparum AP1
Toxoplasma gondii AP1
Tetrahymena thermophila AP2
Bigelowiella natans AP2
Entamoeba histolytica AP2
Dictyostelium discoideum AP2
Naegleria gruberi AP2
Phytophthora sojae AP2
Ectocarpus siliculosus AP2
Thalassiosira pseudonana AP2
Emiliania huxleyi AP2
1/0.51/85/92
Trichomonas vaginalis 2
Leishmania major AP2
Ostreococcus tauri AP2
Arabidopsis thaliana AP2
Physcomitrella patens AP2
Chlamydomonas reinhardtii AP2
Volvox carteri AP2
Trypanosoma brucei AP4
Naegleria gruberi AP4
Dictyostelium discoideum AP4
Plasmodium falciparum AP4
Toxoplasma gondii AP4
Tetrahymena thermophila AP4
Emiliania huxleyi AP4
Thalassiosira pseudonana AP4
Ectocarpus siliculosus AP4
Phytophthora sojae AP4
Bigelowiella natans AP4
Ostreococcus tauri AP4
Chlamydomonas reinhardtii AP4
Volvox carteri AP4
Arabidopsis thaliana AP4
Physcomitrella patens AP4
Trichomonas vaginalis AP4
1/1/100/100
MrBayes/Phylobayes/PhyML/RAxML
≥ 1.00/1.00/95/95
≥ 0.95/0.95/75/75
≥ 0.80/0.80/50/50
0.5

these other families of transporters. Resolution was also obtained between the different AP complexes, except for one problematic node separating AP-1 and AP-2. This is likely because of incongruent signal between the β-subunits of these complexes; in most taxa, one β-subunit is shared between the two complexes, but has duplicated multiple times in distantly related taxa, giving a slightly different evolutionary signal than that observed for the other subunits (Dacks et al., 2008).The concatenated phylogeny also indicates a slightly different branching order for the AP complexes than has previously been reported (Hirst et al., 2011). In particular, AP-3 was proposed to have been the deepest branching AP complex, followed by AP-5, then AP-4, then AP-1 and AP-2. Hirst et al*.,* proposed that the initial duplication of COPI and the APs represented the evolution of a Golgi compartment and a TGN/endosome-like compartment because of the involvement of AP-3 and AP-5 in endosomal transport (Hirst et al., 2011; Peden et al., 2002). The separation of the TGN and endosomes would have co-occurred with the evolution of AP-4 which is predominantly TGN localized (Dell'Angelica et al., 1999a). Finding here that AP-5 is the earliest emerging AP complex does not refute this hypothesis, but is consistent with their argument since the endosomal APs are still the first to diverge. The argument of Hirst et al., would suggest that the ancestor of heterotetrameric complexes would likely have acted at a single type of intracellular compartment in the cell, and its duplication would have produced a Golgi and TGN/endosome-like compartment.

5.4.4.6 TTRAY1/2

Predicted structures and comparative genomics indicated that the closest relative of TTRAY1 and TTRAY2 are COPI-α and –β', members of the B-COP subcomplex (Schledzewski et al., 1999). The question addressed with this analysis differed from the one asked for the subunits above; COPI-α, COPI-β', TTRAY1, and TTRAY2 are all found across the diversity of eukaryotes, and therefore presumed to have been present in the LECA. Therefore, all four of these subunits would have arisen prior to the LECA. What then, is the internal relationship between TTRAY1/2 and COPI-α/β'? In other words, are the COPI subunits more closely related to each other, or are they each sister to a different TTRAY? Phylogenetic analysis showed that COPI-α and –β' are each other's closest relative, as are TTRAY1 and TTRAY2, indicating that the duplications producing two outer coat subunits in COPI and TSET occurred convergently (Figure 5-13). This analysis was able to resolve the pre-LECA duplication producing COPI-α and COPI-β'. It was unable to pinpoint the timing of the duplication producing TTRAY1 and TTRAY2, although the presence of redundant clades suggests that the duplication is ancient and occurred prior to the LECA, given its presence in two of the three major lineages. An attempt to resolve the duplication of TTRAY1 and TTRAY2 was carried out by removing the COPI sequences (Figure 5-14). No significant backbone resolution was obtained; however, TTRAY1 and TTRAY2 sequences from the same organism did not group together, suggesting that an ancient duplication generated these two subunits.

Figure 5-13. Phylogenetic analysis indicates that dual outer coat subunits arose via independent gene duplications. Phylogenetic analysis of TTRAY1, TTRAY2, COPI-α, and COPI-β' was carried out to determine whether the gene duplication giving rise to TTRAY and COPI occurred before or after the gene duplication giving rise to TTRAY1 and TTRAY2 (*i.e.,* are TTRAY1 and 2 each other's closest relative or are they each more closely related to a different COPI subunit?). Phylogenetic analysis indicates that an ancestral gene duplication gave rise to the ancestral COPI and the TTRAY subunits. Independent duplications in both lineages then gave rise to TTRAY1 and TTRAY2 and COPI-α and COPI-β', respectively. This result indicates that the conformation of two different outer coat subunits arose convergently in the two complexes. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (MrBayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed, dark circles ≥ 1.00/95; closed, light circles ≥ 0.95/75; open circles ≥ 0.8/50.

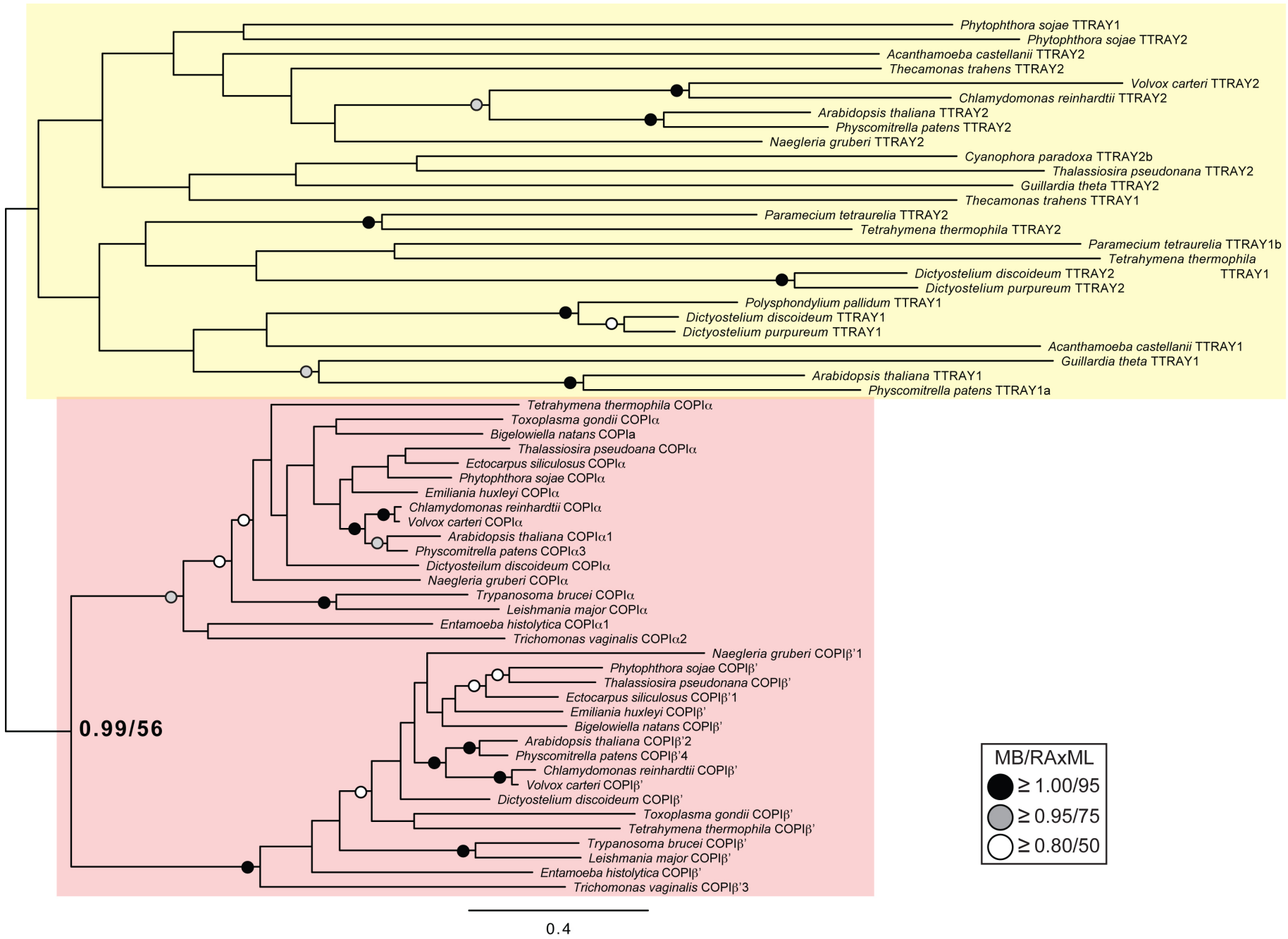
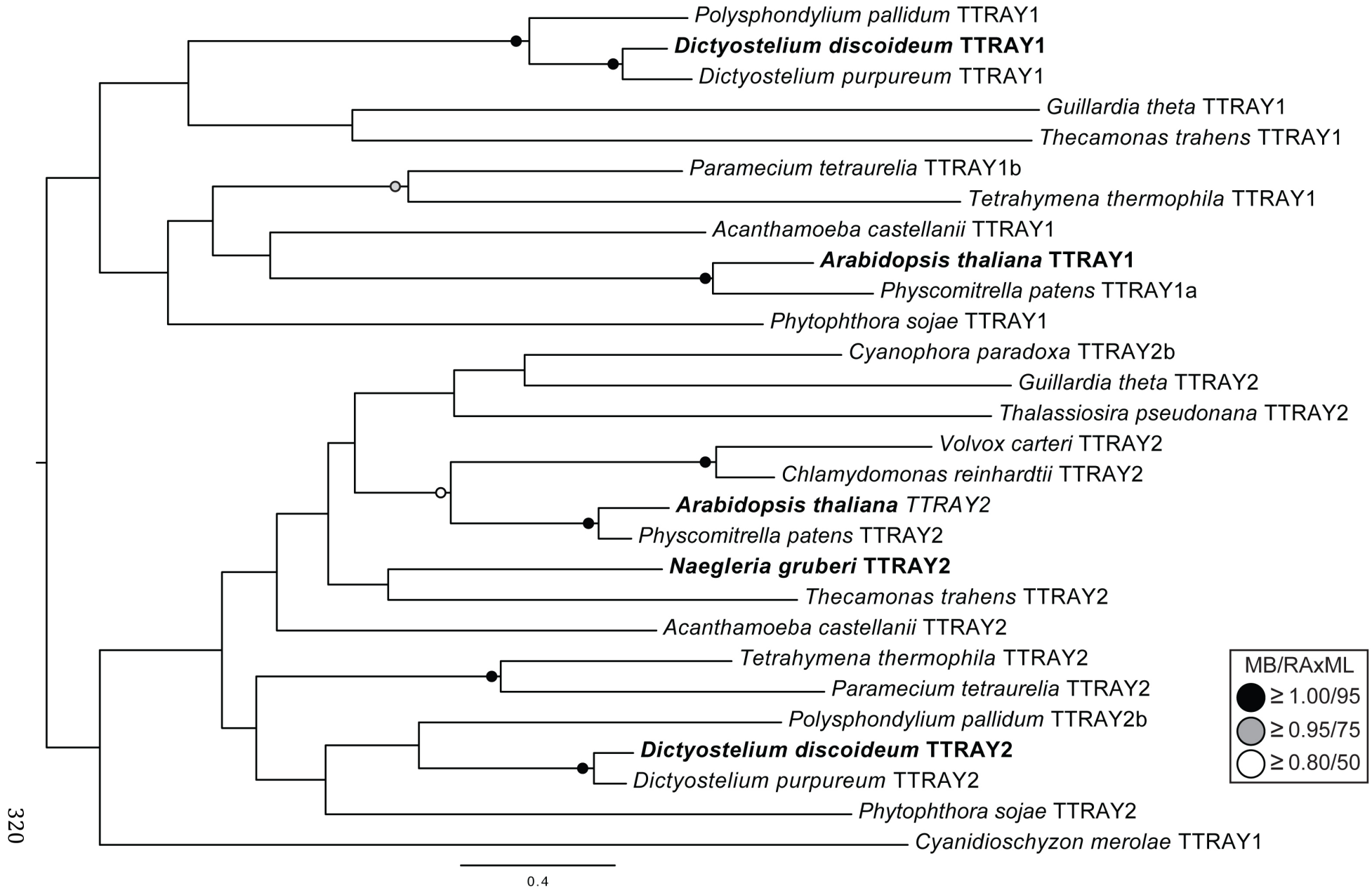

Phytophthora sojae TTRAY1
Phytophthora sojae TTRAY2
Acanthamoeba castellanii TTRAY2
Thecamonas trahens TTRAY2
Volvox carteri TTRAY2
Chlamydomonas reinhardtii TTRAY2
Arabidopsis thaliana TTRAY2
Physcomitrella patens TTRAY2
Naegleria gruberi TTRAY2
Cyanophora paradoxa TTRAY2b
Thalassiosira pseudonana TTRAY2
Guillardia theta TTRAY2
Thecamonas trahens TTRAY1
Paramecium tetraurelia TTRAY2
Tetrahymena thermophila TTRAY2
Paramecium tetraurelia TTRAY1b
Tetrahymena thermophila
Dictyostelium discoideum TTRAY2
TTRAY1
Dictyostelium purpureum TTRAY2
Polysphondylium pallidum TTRAY1
Dictyostelium discoideum TTRAY1
Dictyostelium purpureum TTRAY1
Acanthamoeba castellanii TTRAY1
Guillardia theta TTRAY1
Arabidopsis thaliana TTRAY1
Physcomitrella patens TTRAY1a
Tetrahymena thermophila COPIα
Toxoplasma gondii COPIα
Bigelowiella natans COPIa
Thalassiosira pseudoana COPIα
Ectocarpus siliculosus COPIα
Phytophthora sojae COPIα
Emiliania huxleyi COPIα
Chlamydomonas reinhardtii COPIα
Volvox carteri COPIα
Arabidopsis thaliana COPIα1
Physcomitrella patens COPIα3
Dictyosteilum discoideum COPIα
Naegleria gruberi COPIα
Trypanosoma brucei COPIα
Leishmania major COPIα
Entamoeba histolytica COPIα1
Trichomonas vaginalis COPIα2
Naegleria gruberi COPIβ'1
Phytophthora sojae COPIβ'
Thalassiosira pseudonana COPIβ'
Ectocarpus siliculosus COPIβ'1
Emiliania huxleyi COPIβ'
Bigelowiella natans COPIβ'
Arabidopsis thaliana COPIβ'2
Physcomitrella patens COPIβ'4
Chlamydomonas reinhardtii COPIβ'
Volvox carteri COPIβ'
Dictyostelium discoideum COPIβ'
Toxoplasma gondii COPIβ'
Tetrahymena thermophila COPIβ'
Trypanosoma brucei COPIβ'
Leishmania major COPIβ'
Entamoeba histolytica COPIβ'
Trichomonas vaginalis COPIβ'3
0.99/56
MB/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50
0.4

Figure 5-14. Phylogenetic analysis suggests that TTRAY1 and TTRAY2 are the result of an ancient duplication. COPI-α and β' sequences were removed to determine whether a single gene duplication event gave rise to TTRAY1 and TTRAY2, or if this occurred independently in multiple lineages. Original queries from *A. thaliana, D. discoideum,* and *N. gruberi* are in bold to illustrate separate, although unsupported, positions on the tree. Other taxa with both subunits do not group together, indicating that TTRAY1 and TTRAY2 likely arose from a single ancestral gene duplication. The best Bayesian topology is shown. Numerical values represent Bayesian posterior probabilities (MrBayes)/Maximum-Likelihood bootstrap values (RAxML). Nodes of interest are in bold. Values for other supported nodes have been replaced by symbols: closed, dark circles ≥ 1.00/95; closed, light circles ≥ 0.95/75; open circles ≥ 0.8/50.


Polysphondylium pallidum TTRAY1
Dictyostelium discoideum TTRAY1
Dictyostelium purpureum TTRAY1
Guillardia theta TTRAY1
Thecamonas trahens TTRAY1
Paramecium tetraurelia TTRAY1b
Tetrahymena thermophila TTRAY1
Acanthamoeba castellanii TTRAY1
Arabidopsis thaliana TTRAY1
Physcomitrella patens TTRAY1a
Phytophthora sojae TTRAY1
Cyanophora paradoxa TTRAY2b
Guillardia theta TTRAY2
Thalassiosira pseudonana TTRAY2
Volvox carteri TTRAY2
Chlamydomonas reinhardtii TTRAY2
Arabidopsis thaliana TTRAY2
Physcomitrella patens TTRAY2
Naegleria gruberi TTRAY2
Thecamonas trahens TTRAY2
Acanthamoeba castellanii TTRAY2
Tetrahymena thermophila TTRAY2
Paramecium tetraurelia TTRAY2
Polysphondylium pallidum TTRAY2b
Dictyostelium discoideum TTRAY2
Dictyostelium purpureum TTRAY2
Phytophthora sojae TTRAY2
Cyanidioschyzon merolae TTRAY1
MB/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50
0.4

5.5 Discussion

In this chapter, we have characterized a novel adaptin-like coat complex, composed of six subunits: TPLATE, TSAUCER, TCUP, TSPOON, TTRAY1, and TTRAY2, collectively called TSET. Structural predictions indicated that each subunit of TSET shares extensive structural similarity to AP complexes and COPI suggesting an evolutionary relationship. Comparative genomics identified a complete complex in four different supergroups, indicating that TSET was present in the LECA. Phylogenetic analysis indicated that it is distinct from both COPI and the AP complexes.

Functional analyses in *D. discoideum* carried out by J. Hirst, D. Traynor, G. Bloomfield, R. Antrobus, R.R. Kay, and M.S. Robinson identified a role for TSET at the plasma membrane. Disruption of complex formation through knockout of TSPOON did not affect cell viability, nor did it affect the ability of *D. discoideum* to form fruiting bodies (Hirst et al., 2014). Fluid-phase endocytosis was not affected in TSPOON-knockout cells. However, membrane turnover as measured by uptake of the membrane marker FM1-43 was slower than control cells, indicating a role in plasma membrane turnover.

At the same time as this analysis, another group independently identified TSET in *A. thaliana*, calling it the TPLATE complex (TPC), after the previously identified TPLATE subunit (Gadeyne et al., 2014). In *A. thaliana* the TPC is composed of eight subunits: the six found in TSET as well as two Eps15 homology domain-containing proteins, AtEH1 and AtEH2. The TPC was also found to colocalize with clathrin at the plasma membrane. Tandem affinity purification experiments

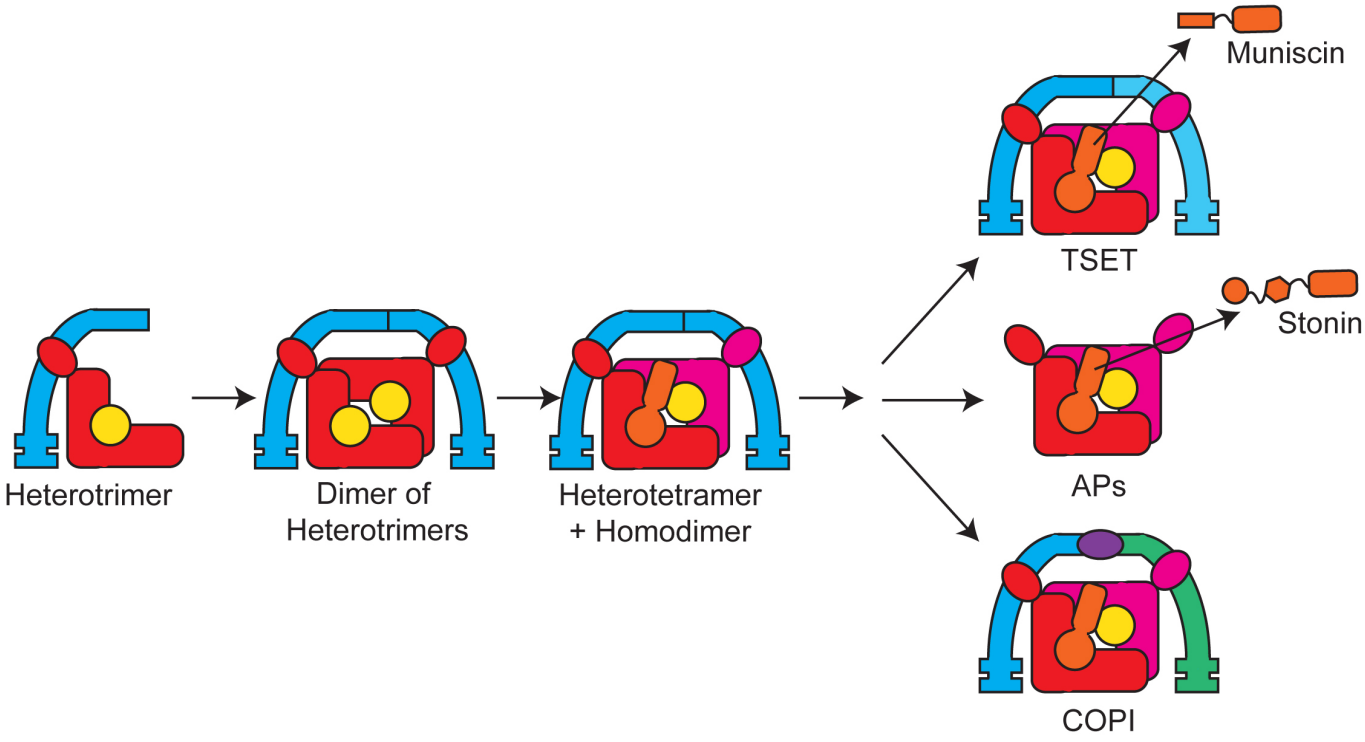
identified clathrin and AP-2 as interacting partners for the TPC, which further points to a role in endocytosis at the plasma membrane. Unlike *D. discoideum*, mutations in TPC subunits in *A. thaliana* resulted in pollen lethality and its down regulation results in seedling lethality, indicating that the TPC is essential for plant development (Gadeyne et al., 2014).

5.5.1 Model for the evolution of TSET and heterotetrameric complexes

The discovery and evolutionary analysis of the TSET complex allows its integration into the evolutionary model of heterotetrameric coat complexes. The original versions of the model (Boehm and Bonifacino, 2001; Schledzewski et al., 1999) proposed that a heterodimeric complex composed of one large and one small subunit dimerized to form a homodimer. This gene set duplicated, producing two large and two small subunits that diverged over time with one of the small subunits gaining a mu-homology domain. This heterotetramer then duplicated multiple times to give rise to COPI and the AP complexes.

With the discovery of TSET, not only does this newly identified complex need to be incorporated into the model, but so too does the outer coat configuration represented by TTRAY1 & 2 and COPI-α and –β'. A new hypothesis would suggest that, rather than simply one large and one small subunit, one outer coat unit was present as well, producing a heterotrimer (Figure 5-15). This heterotrimer homodimerized to form a dimer of trimers, generating a complete coat with both inner and outer coat complexes. Gene duplications of the large and small subunits, followed by sequence divergence and the gain of a mu-homology domain in one of

Figure 5-15. Model for the evolution of heterotetrameric complexes. All heterotetrameric complexes evolved from a common ancestor composed of one small, one large, and one outer coat component that form a heterotrimer, which forms a homodimer during vesicle formation. Duplication and sequence divergence of the large and small subunits gives rise to the heterotetramer, with one small subunit gaining a mu-homology domain. Multiple duplications of the entire complex gave rise to TSET, the APs, and COPI, with the outer component of TSET and COPI duplicating independently, and the APs losing the outer coat. In opisthokonts, the medium subunits of TSET and AP-2 acquire additional domains, giving rise to the muniscins and stonins, respectively. Modified from Hirst et al., 2014.


Heterotrimer
Dimer of
Heterotrimers
Heterotetramer
+ Homodimer
Muniscin
TSET
Stonin
APs
COPI

the small subunits, would produce the four subunits that make up the heterotetrameric core. Multiple coordinated duplications of all five subunits would have given rise to each lineage of trafficking complex, with independent duplications of the outer components in TSET and COPI giving rise to TTRAY1 & 2 and COPI-α/-β', respectively, and loss of the outer component entirely in the AP complexes. Finally, coordinated loss of the entire TSET complex except for TCUP, and the swapping of the N-terminal longin domain for an F-BAR domain would have given rise to the muniscins in opisthokonts (Figure 5-15).

5.5.2 Endocytosis has evolved separately multiple times

The available data suggest that TSET is involved in endocytosis at the plasma membrane, at least in *A. thaliana* and *D. discoideum* (Gadeyne et al., 2014; Hirst et al., 2014). It therefore appears that coat-based endocytosis (as opposed to phagocytosis) has evolved multiple times independently. In mammalian cells, where complete TSET complexes are not present, knockout of AP-2 is embryonic lethal, underpinning the importance of this trafficking complex in this lineage (Mitsunari et al., 2005). Moreover, FCHo plays a role in AP-2 vesicle biogenesis (Henne et al., 2010; Hollopeter et al., 2014; Umasankar et al., 2014). In *A. thaliana*, knockout of TSET, rather than AP-2, causes major growth defects, whereas loss of AP-2 imposes minor growth defects, but none so severe that plant development is dramatically impacted (Gadeyne et al., 2014). The role of TSET in endocytosis is made even more clear by studies in *D. discoideum*. However, disruption of this complex in this system carried a much milder phenotype (Hirst et al., 2014). Similarly, knockout of

individual AP-2 subunits in *D. discoideum* resulted in viable cells, capable of endocytosis (Macro et al., 2012).

Clearly TSET and AP-2 both have roles in endocytosis but the evidence for the independent gain of function at the plasma membrane is presented here, illustrated by the concatenated analysis of COPI, TSET, and the AP complexes (Figure 5-12). There is no support for TSET branching within the clade containing the AP complexes, let alone as sister to AP-2. This separation supports the hypothesis of functional convergence in endocytosis for TSET and AP-2. The alternative is for the ancestor of TSET and of all AP complexes to have functioned at the plasma membrane, with APs 1, 3, 4, and 5 (and perhaps even COPI) having gained new functions. While technically possible, it is far less parsimonious, leaving the two origins scenario as the most likely explanation.

5.5.3 Muniscins and stonins are examples of convergent evolution in the membrane trafficking system

It is interesting to note that the two complexes involved in endocytosis in the LECA have spawned additional adaptor molecules that continue to be involved in that process. TSET, although lost in the lineage leading to the ancestor of opisthokonts, passed on the mu-homology domain present in TCUP, forming the muniscins: FCHo1/2 and SGIP in Metazoa, and Syp1 in Fungi. Functional analyses suggest that FCHo is involved in stabilizing the formation of nascent AP-2 vesicles while also aiding in the incorporation of cargo into budding vesicles (Henne et al., 2010; McMahon and Boucrot, 2011). Stonins are metazoan-specific mu-homology

domain containing proteins thought to be derived from AP-2μ, have been shown to act as adaptors for synaptotagmin 1, and are required for its endocytosis in neuronal cells (Diril et al., 2006; Jung et al., 2007; Martina et al., 2001). Both of these proteins increase the diversity of cargoes able to be taken up by the AP-2 complex. In the case of FCHo, it is conceivable that the cargo that it binds were at one point incorporated into vesicles formed by TSET. Stochastic mutation or gene loss may have rendered the TSET complex non-functional resulting in the selection for interaction between TCUP and the AP-2 complex, or perhaps TCUP was promiscuous and was already being incorporated into AP-2 vesicles. Alternatively, a single gene duplication of TCUP, followed by the exchange of the longin domain for an F-BAR domain, may have simplified the endocytic system rendering TSET redundant and was subsequently lost, either through negative selection, or through genetic drift. A similar argument may be made for the evolution of stonins; a gene duplication or domain swap generated a protein capable of binding additional cargo, opening the door for a more complicated endocytic system. As a whole, FCHo and the stonins are examples of how complexity can evolve within a system: in one case, a change in binding specificity from one complex to another, and in the other case, the generation of a new subunit altogether. However, both carry out similar functions to achieve the same goal.

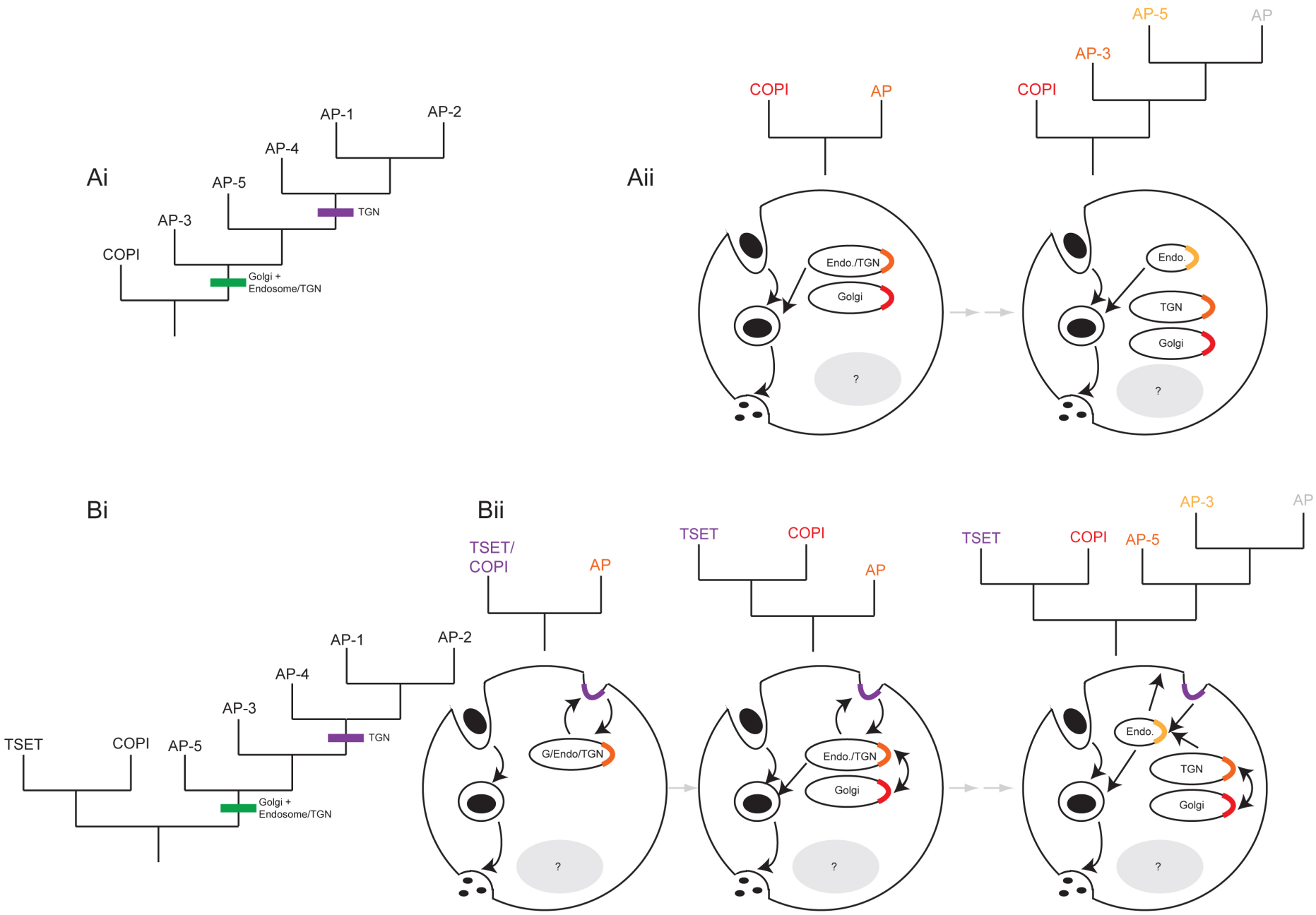
5.5.4 TSET and the evolution of the early membrane trafficking system

The discovery of TSET adds an additional step before the evolution of the TGN/endosome-like structure proposed by Hirst et al., (2011). In their hypothesis,

the gene duplication that produced COPI and the ancestral AP complex also gave rise to a Golgi and a TGN/endosome-like compartment (Figure 5-16A). The role for TSET in endocytosis suggests an additional step that connected this early compartment with the plasma membrane. An earlier duplication event likely gave rise to ancestors of TSET/COPI and the APs and, produced two coats, one acting at the plasma membrane during endocytosis, and another possibly involved in communication with the early secretory system, or transport to the plasma membrane, or both (Figure 5-16B). TSET and COPI are assumed to branch together based on recent evidence from phylogenetic analyses of longin domain-containing proteins that suggest that TSET and COPI share a more recent common ancestor with each other than either do with the AP complexes (C. Klinger, personal communication).

Evidence presented in chapter 3 and by others (Elias et al., 2012) point to an early gene duplication producing machinery involved in endocytosis and machinery involved in exocytosis. The role of TSET in endocytosis suggests that it may have gained this function early in eukaryote evolution. The alternative complex would be AP-2. However, it is clear that the evolution of AP-2 occurred after the membrane trafficking system was much more established. Duplication of one AP complex would generate the Golgi complex and the endosomes/TGN, as proposed by Hirst et al., (2011; Figure 5-16). Subsequent duplications would then produce additional compartments, eventually segregating endosomal and TGN functions into distinct compartments (Figure 5-16), allowing for further modification and tailoring of this arm of the membrane trafficking system.

Figure 5-16. Hypothesis for the early evolution of the heterotetrameric complexes and the membrane trafficking system. A) Model for the evolution of the endosomal system as put forth by Hirst et al., (2011). Ai) Evolutionary tree of the heterotetrameric coat complexes. Aii) Organelle evolution based on tree in Ai. Early duplication giving rise to COPI and the ancestor of the AP complexes produced a Golgi and endosome/TGN-like organelle, uniting the membrane trafficking system with the phagocytic system. Duplications of AP-3 and AP-5 then gave rise to endosomal compartments, separating the endosome from the TGN. Grey circle and question mark represents the early secretory system, as they are explicitly ignored here (see Figure 6-3). B) Incorporation of TSET into the Hirst et al., model. Bi) Tree of the evolution of heterotetrameric complexes rooted on the APs (see text for rationale). Bii) An early duplication of the ancestral complex produced the ancestors of the TSET/COPI complexes and the ancestor of the AP complexes. These early coat complexes could have carried out basic endocytic and exocytic trafficking. Subsequent duplications of the complex would then give rise to the organelles as in Aii. G = Golgi, Endo. = Endosomes, TGN = *trans*-Golgi Network. Grey 'AP' represents the remaining AP complexes (AP-1, AP-2, and AP3).


Ai
AP-1
AP-2
AP-4
AP-5
TGN
AP-3
COPI
Golgi +
Endosome/TGN
Aii
COPI
AP
Endo./TGN
Golgi
?
AP-5
AP
AP-3
COPI
Endo.
TGN
Golgi
?
Bi
AP-1
AP-2
AP-4
AP-3
TGN
TSET
COPI
AP-5
Golgi +
Endosome/TGN
Bii
TSET/
COPI
AP
G/Endo/TGN
?
TSET
COPI
AP
Endo./TGN
Golgi
?
TSET
COPI
AP-5
AP-3
AP
Endo.
TGN
Golgi
?

Chapter 6: Perspectives

A portion of this chapter has been published as:

Schlacht, A., Herman, E.K., Klute, M.J., Field, M.C., Dacks, J.B. 2014. Missing pieces of an ancient puzzle: evolution of the eukaryotic membrane-trafficking system. In *Cold Spring Harbor Perspectives: The Origin and Evolution of Eukaryotes* (eds. Keeling, P.J., Koonin, E.V.) Cold Spring Harbor Press

6.1 Synopsis

Early analyses of membrane trafficking pointed toward a LECA with a tremendously complex membrane trafficking system, not unlike that observed in modern eukaryotes (Koumandou et al., 2013, *inter alia*). These studies focused primarily on membrane trafficking machinery involved in vesicle fusion, as it was thought that this machinery was responsible for encoding organelle identity. To determine whether or not the machinery involved in vesicle formation is similarly conserved, I carried out comparative genomic analyses of regulatory elements of vesicle formation (ArfGAPs and ArfGEFs), and coat proteins (COPII, TSET). We saw that ArfGAPs, ArfGEFs, COPII, and TSET are broadly conserved across eukaryotes, indicating their presence in the LECA and solidifying their importance in general models of membrane trafficking. However, it became apparent that not all subfamilies, coats, or subunits thereof, are equally well conserved, pointing to examples of where evolution has shaped biological process in different eukaryotic lineages. Specifically, three patterns of conservation were consistently observed; ubiquitous and lineage-specific patterns had been observed previously. However, patchy distributions had previously been written off as either sampling error or misidentification of homologues as the product of fast evolving genomes. Work in this thesis makes it abundantly clear that this latter pattern of conservation is far more pervasive than previously thought, and suggests that patchy proteins may have, or may still, be playing a major role in shaping the cellular landscape of eukaryotic cells. The analyses also allow the integration of important membrane

trafficking subfamilies into the OPH, solidifying hypotheses about the early evolution of the membrane trafficking system.

6.2 Multiple patterns of protein conservation

While the comparative genomic analyses presented in chapters 3, 4, and 5 identified ancient membrane trafficking protein families and coat complexes, each protein family or complex appeared to have members with ubiquitous, lineage-specific, and patchy distributions. Independent analyses of Rabs and TBCs have identified protein subfamilies with similar patterns of conservation, indicating that these patterns are not limited to the proteins analyzed here or to components of vesicle formation, but rather, are prominently distributed throughout the membrane trafficking system.

6.2.1 Ubiquitously conserved proteins

Proteins that are distributed across the diversity of eukaryotes and that are seldom lost characterize the first pattern of protein conservation, ubiquitously conserved proteins. This pattern suggests that these proteins are necessary to maintain basic cellular function. Examples from previous chapters include the ArfGAPs: SMAP, ArfGAP1, ArfGAP2, ACAP, and AGFG; the ArfGEFs: BIG and GBF; and the majority of COPII coat components: Sar1, Sec23, Sec24, Sec13, and Sec31. Finding proteins conserved in this manner provides support to established models of cell biology. For example, that the five core COPII coat components are ubiquitous, lends support to the importance of its function in exit of cargo from the

ER (Barlowe et al., 1994). The conservation of ArfGAP1 and GBF supports the importance of Arf regulation at the *cis*-Golgi (Casanova, 2007; Kahn et al., 2008). The ubiquity of these proteins also points to their presence in the LECA.

Other protein families in the membrane trafficking system also share this distribution. This was the pattern of conservation originally suggesting a complex trafficking system in the LECA (Dacks and Doolittle, 2001; Dacks and Field, 2004). This pattern is generally found when analyzing key trafficking complexes, such as ESCRTs (Leung et al., 2008), the MTCs (Klinger et al., 2013; Koumandou et al., 2007), and membrane deformation machinery (Dacks and Field, 2004; Neumann et al., 2010). More recently, this pattern has been identified in highly paralogous protein families with functions spanning the trafficking system. Analysis of the Rab proteins suggested the presence of up to 23 Rab paralogues in the LECA, of which nine (1, 2, 4, 5, 6, 7, 8, 11, and 18) are seldom lost (Elias et al., 2012). The conservation of this machinery across trafficking pathways or paralogous protein families is additional support for a functionally complex LECA.

6.2.2 Lineage-specific proteins

The second pattern of protein conservation is lineage specificity. In contrast to the broadly conserved proteins above, these proteins are limited in their taxonomic distribution, indicative of recent paralogous duplication. Although these proteins may have important functions for the lineage in which they are found, they should not be incorporated into general models of the eukaryotic membrane traffic as they are not broadly distributed. Examples of proteins from previous chapters

include the opisthokont and holozoan ArfGAPs: ASAP, ARAP, and GIT, and the opisthokont ArfGEFs: EFA6 and FBX8, and the COPII subunit Sed4. However, the identification of proteins with this distribution can provide insight into organism-specific biology. For example, the regulation of focal adhesions by ASAP in metazoan cells (Randazzo et al., 2007) or the additional regulatory mechanism imposed on forming COPII coats by Sed4 (Kodera et al., 2011) are both novel additions to the cell biology of Metazoa and Fungi, respectively.

Proteins with lineage-specific distributions were also identified in early comparative genomic analyses, examples include caveolin, stonins, GGAs, and novel paralogues of highly conserved subfamilies, such as Rabs and Arfs (Boehm and Bonifacino, 2001; Diekmann et al., 2011; Field et al., 2007b; Manolea et al., 2010). Most of our understanding of membrane trafficking, and cell biological processes in general, stem from studies carried out in mammalian and yeast model systems. This has resulted in a wealth of opisthokont-specific machinery and fewer examples elsewhere on the eukaryotic tree. This asymmetry results in bias towards searching for machinery characterized in opisthokont systems, rendering lineage-specific machinery in other taxa unidentified or missing. Recognition of this bias has begun to rectify this problem; early phylogenetic analyses identified independent duplications giving rise to the β-subunit of AP 1 and 2 (Boehm and Bonifacino, 2001; Dacks et al., 2008; Schledzewski et al., 1999). Multiple expansions of Rabs and SNAREs in plants have also been well established (Rutherford and Moore, 2002; Sanderfoot, 2007). Moreover, the development of phylogenetic pipelines such as Scrollsaw, in order to identify both ancient and lineage-specific paralogues of large

protein families that have otherwise gone unidentified (*e.g.,* TBC-ExA in excavates, TBC-PlA, TBC-PlB in plants, and the archaeplastid-specific Sec23 paralogue described here; Elias et al., 2012; Gabernet-Castello et al., 2013; Schlacht and Dacks, 2015).

The development of genetically tractable systems outside of opisthokonts has also begun to rectify this problem. One such example are the trypanosomes, pathogens found in the supergroup Excavata, evade the immune system by constantly recycling surface antigens (Allen et al., 2003), a process greatly dependent on endocytosis. Novel adaptations of endocytic function have been identified in trypanosomes, including the loss of AP-2 and the presence of trypanosome-specific clathrin associated machinery that mediates endocytosis at the plasma membrane (Adung'a et al., 2013; Manna et al., 2015). Although functional characterization is still required, this machinery represents a novel mechanism for regulating endocytosis not present in other eukaryotes and likely would not have been identified by bioinformatics, highlighting the importance of studying these processes in other eukaryotes.

A second example is found in the ciliate *T. thermophila*. It has been proposed that mucocysts, secretory granule-like organelles, may have arisen convergently in ciliates and Metazoa (Elde et al., 2007). Transport of cargo to these organelles is dependent on the cargo receptor sortillin/Vps10, of which four paralogues are present in *T. thermophila* versus five in humans (Koumandou et al., 2011). These expansions occurred independently in vertebrates and in ciliates (Briguglio et al.,

2013), suggesting either functional convergence in these lineages or divergence in trafficking processes of cargo to lysosome-related organelles.

6.2.3 Patchy proteins

The third pattern of protein conservation is the "patchy" distribution. These proteins are broadly conserved, and were likely to have been present in the LECA; however, they are frequently missing. In some cases, these proteins are missing from animals and fungi, resulting in omission from general models of membrane trafficking. Examples of patchy proteins in previous chapters include AGAP, and ArfGAPC2, Sec24III, and the entire TSET complex. The identification of proteins with this distribution is somewhat perplexing; these proteins are sufficiently necessary to have been retained in some lineages since the LECA, but are clearly disposable given appropriate cellular contexts.

Patchy distributions of proteins were observed in early comparative genomic analyses (Field et al., 2007b), and persist, even with substantially more sequenced genomes and greater depth of analyses of protein families. One example is from the ESCRTs, endosomal proteins responsible for budding vesicles into the multivesicular body (Henne et al., 2011). ESCRT subcomplexes I-IV are well conserved, whereas ESCRT-0 is opisthokont-specific (Herman et al., 2011; Leung et al., 2008). The protein Tom1-esc has been suggested to possess overlapping functionality with ESCRT-0, such as binding ubiquitin and interacting with ESCRT-1 (Blanc et al., 2009; Puertollano, 2005). Tom1-esc has a much broader distribution than ESCRT-0, but is frequently missing. Another example is DSCR3, a paralogue of

the Vps26 subunit that is involved in recycling vacuolar receptors from early endosomes to the TGN as part of the retromer complex (Seaman, 2012). The function of DSCR3 is not known beyond an association with Down's Syndrome. While Vps26 is very well conserved, DSCR3 is found broadly, but not frequently (Koumandou et al., 2011).

Other examples of patchy proteins have been identified in large protein families. As mentioned in chapter 5, AP complexes mediate trafficking in the late secretory and endolysosomal systems and are differentially conserved across eukaryotes, with AP-1 and AP-5 being the best and least conserved, respectively (Hirst et al., 2011). Members of the TBC family (TBC-F through –N) are broadly, but sparsely found. Surprisingly, subfamilies of Rabs and TBCs, RabTitan and TBC-RootA respectively, have been identified that while broadly conserved, have been lost multiple times including from humans. These paralogues are expected to have important roles in other taxa, but are missing from our biology. There is a paucity of functional information regarding these paralogues, and it will be interesting to determine if they represent ancient functionality not present in opisthokonts, or if they possess redundant or convergent functions to other cellular factors.

6.3 Patchy proteins may be redundant to other cellular factors

Important functional and evolutionary information can be gained from lineage-specific and well-conserved proteins. They clarify which components can be generalized to models of cell biological processes in all eukaryotes, and indicate that many of the basic biological processes occurring in animals and fungi also occur in

other eukaryotic organisms. These similarities provide a starting point from which we can study differences between organisms to understand how natural selection affects different taxonomic lineages. By contrast, proteins with a limited distribution inform how organisms differ from the general model.

At present it is unclear what information can be gleaned from proteins displaying patchy distributions. This is largely due to the absence of functional data for many of these proteins. At the very least, it is reasonable to assume that they play a role consistent with the catalytic domain present in the protein, *i.e.,* RabTitan likely acts as a Rab GTPase, and ArfGAPC2 likely acts as an ArfGAP protein. However, this is merely pointing out the obvious. Some patchy protein are likely redundant with other cellular factors. For example, Sec24III is very likely a cargo binding subunit at ER exit sites as no other function has yet been identified for Sec24. Sec24III may possess a unique repertoire of cargo molecules in some taxa, or it may bind overlapping cargo with other Sec24 paralogues.

The best candidate for a patchy protein with functional redundancy is TSET. The TSET complex is involved in plasma membrane turnover in the amoebae *D. discoideum* and clathrin-mediated endocytosis in the flowering plant *A. thaliana* (Gadeyne et al., 2014; Hirst et al., 2014), a function canonically carried out by the AP-2 complex in mammalian cells (Cocucci et al., 2012). The ancient functional redundancy of these two complexes may have permitted tailoring of endocytic processes in different lineages. In mammalian cells, AP-2-clathrin vesicles largely carry out endocytosis, as the TSET complex is not present (Hirst et al., 2014). Knockout of AP-2 in mammalian cells is embryonic lethal (Mitsunari et al., 2005),

likely because no compensatory mechanism exists. By contrast, knockout of AP-2 in *A. thaliana* does not result in lethality, rather, plants display developmental defects but are viable (Zhang et al., 2015). However, mutation of TSET subunits results in pollen lethality and loss of viability indicating that TSET, not AP-2, is essential in plants (Gadeyne et al., 2014; Van Damme et al., 2011). In *A. thaliana*, clathrin has been shown to bind both AP-2 and TSET. Multiple subpopulations of nascent clathrin vesicles have been reported, the majority of which include both complexes (Gadeyne et al., 2014), although vesicle nucleation sites uniquely containing AP-2 or TSET have also been observed. These findings suggest that clathrin vesicles with heterogeneous adaptor complexes can form. In *D. discoideum*, neither knockout of AP-2, nor TSET results in loss of viability (Hirst et al., 2014; Macro et al., 2012). Moreover, clathrin does not appear to interact with the *D. discoideum* TSET complex, suggesting that its interaction with clathrin is unique to plants (Hirst et al., 2014). It should be kept in mind that this is based on limited functional data, and as TSET, and clathrin-dependent endocytosis generally, are studied in organisms from different supergroups, the roles that these two coat complexes play will become much more clear.

6.4 Patchy proteins may have permitted fine-tuning of the membrane trafficking system

There is clear functional redundancy provided by at least some patchy proteins. What benefit would redundancy provide to a cellular system? One argument has been made suggesting that cellular complexity may be selected for

under conditions with low selective pressures, as systemic redundancy would reduce the impact of mutations that would otherwise disrupt a complicated system (Schlacht et al., 2014).

Another hypothesis is that they permitted fine-tuning of the membrane trafficking system. Early in the evolution of the endomembrane system, particular biochemical functions such as GTPase activity, regulation of GTPases by GAPs and GEFs, coat proteins, *etc*., would have been required. The incorporation of proteins able to fill these vacant functional roles would have occurred, regardless of their catalytic efficiency. The subsequent evolution of more efficient paralogues may have functionally replaced some of these patchy proteins, allowing for more efficient or precise modes of regulation, resulting in their loss. This process may be continuing in extant eukaryotes; patchy proteins may be replaced by lineage-specific expansions that are better adapted to more specialized cellular contexts than the ancient paralogues, similar to birth and death processes that take place during organismal evolution. For example, *A. castellanii* has lost the patchy Rab RTW (Elias et al., 2012). It is possible that expansion of the *A. castellanii* Rab32 clade, which is largely associated with the ER, but has also been found in the endosomal system (Bultema et al., 2012; Friedman et al., 2011), has compensated for the missing RTW, suggesting a role for this protein at both locations.

Alternatively, loss of RTW may be the result of loss of other cellular processes or complexes, as is likely the case for the Rab IFT27, whose loss corresponds to losses of the intraflagellar transport complex (Elias et al., 2012; van Dam et al., 2013). Nonetheless, further functional characterization of these patchy

proteins in a variety of organisms will be required to fully understand the roles of these proteins in extant eukaryotes, and what their contributions to the evolution of the membrane trafficking system may have been.

6.5 Integration into the Organelle Paralogy Hypothesis

6.5.1 ArfGAPs and ArfGEFs

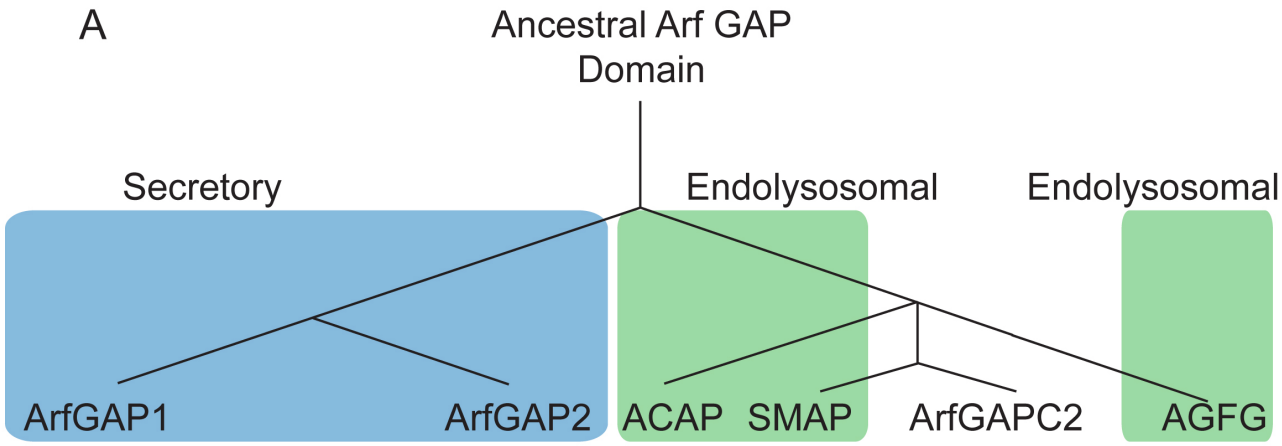
The presence of a single Arf homologue in the LECA excludes its incorporation into the organelle paralogy hypothesis, because a single paralogue is unable to provide information about the evolution of multiple organelles. However, Arfs are regulated by ArfGAPs and ArfGEFs, both of which are paralogous protein families with organelle-specific paralogues, some of which were previously proposed to have been present in the LECA (Cox et al., 2004; Kahn et al., 2008). Therefore, the evolution of the ArfGAP and ArfGEF subfamilies were analyzed in the context of the OPH.

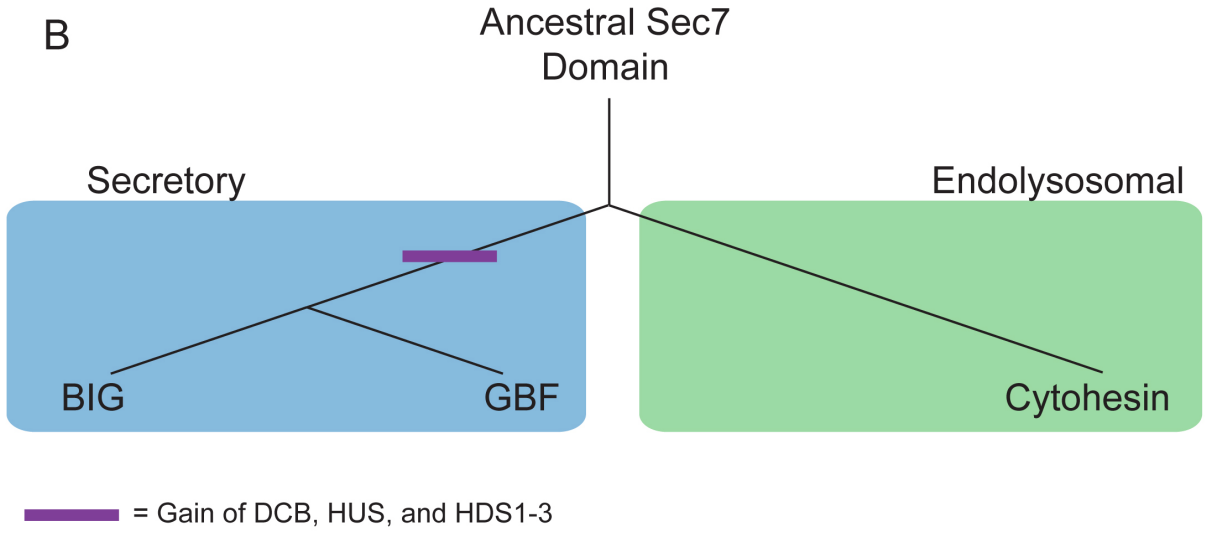
Although no protein family has yet been able to provide a complete accounting for the order in which the organelles of the membrane trafficking system evolved, analyses carried out thus far are consistent in suggesting what may have been the earliest events in the evolution of this system. Scrollsaw analysis of Rab proteins resolved an early split into endocytic and exocytic Rabs (Elias et al., 2012). The ArfGAP proteins present in the LECA suggest a similar event. ArfGAP1 and ArfGAP2 are both active in the early secretory pathway (Bigay et al., 2005; Weimer et al., 2008), and are each other's best reciprocal BLAST hit after themselves, suggesting an ancient pre-LECA duplication that gave rise to these two subfamilies

(Figure 6-1). SMAP, ACAP, and AGFG act in the endocytic system (Li et al., 2007; Natsume et al., 2006; Pryor et al., 2008). BLAST experiments against the human genome for these latter four subfamilies retrieve each other before retrieving either ArfGAP1 or ArfGAP2, suggesting a more recent common ancestor. BLAST results suggest that ArfGAPC2 should also be placed in this clade although its function is presently unknown. Taken together, this suggests an ancient duplication of an ancestral ArfGAP domain that gave rise to one lineage involved in secretion and the other in endocytosis (Figure 6-1). It should be made clear that is by no means a definitive relationship among ancient ArfGAP subfamilies, but rather a hypothesis to be tested in future analyses.

The ArfGEF proteins suggest a similar duplication (Figure 6-1). The analyses presented in chapter 4 indicate that only three ArfGEFs were present in the LECA: cytohesin, BIG, and GBF. Cytohesin primarily acts at the plasma membrane, regulating endocytic events and membrane remodelling (Geiger et al., 2000). GBF and BIG are primarily localized to the *cis*- and *trans*-Golgi network (Shinotsuka et al., 2002a; Zhao et al., 2006), respectively, indicative of a role in the secretory pathway. It should be noted that BIG2 has been observed to function at endosomal compartments in mammalian cells (Shinotsuka et al., 2002b). Similar reports have been made for some paralogues of GBF in *A. thaliana* (Teh and Moore, 2007). However, given that these represent independent, lineage-specific expansions, these secondary locations of function are likely the result of convergence, rather than a shared ancestral function of both BIG and GBF, similar to the expansions and

Figure 6-1. Hypothesis for the early evolution of ArfGAPs and ArfGEFs. A) Predicted relationships between ancient ArfGAPs results in their clustering into secretory and endolysosomal clades. ArfGAP1 and ArfGAP2 both act in the early secretory system during COPI coat formation, and retrieve each other as their next-best hit when BLASTed against the human genome. SMAP, ACAP, and AGFG act in the endolysosomal system and retrieve each other when BLASTed against the human genome, before hitting ArfGAP1 or ArfGAP2. These observations suggest an ancient duplication of an ancestral ArfGAP domain-containing protein, generating secretory and endolysosomal clades of ArfGAPs. ArfGAPC2 retrieves SMAP as its best hit when BLASTed against the human genome, but is not coloured as the cellular location at which it acts is currently unknown. B) ArfGEFs also segregate into secretory and endolysosomal clades. BIG and GBF are unified by the presence of the DCB, HUS, and HDS domains (purple bar), forming a distinct clade from cytohesin. These two lineages are likely the products of an ancient duplication of the ancestral Sec7 domain-containing protein.


A
Ancestral Arf GAP Domain
Secretory
Endolysosomal
Endolysosomal
ArfGAP1
ArfGAP2
ACAP
SMAP
ArfGAPC2
AGFG


B
Ancestral Sec7 Domain
Secretory
Endolysosomal
BIG
GBF
Cytohesin
= Gain of DCB, HUS, and HDS1-3

functional convergence of endosomal syntaxins in animals and plants (Dacks et al., 2008). This is especially likely in *A. thaliana*, as only BIG and GBF are present, requiring paralogues of these proteins to fill functional roles carried out by other subfamilies in mammalian cells.

Previous reports have suggested a common origin for BIG and GBF as the result of a pre-LECA gene duplication. This was based on the extreme overlap in domain conservation along the length of their sequences (Mouratou et al., 2005). Cytohesin would then have either emerged prior to the duplication giving rise to BIG and GBF, or would have arisen from BIG or from GBF. I propose that cytohesin arose from an earlier gene duplication that also gave rise to the ancestor of BIG and GBF. The rationale is as follows: first, the two scenarios presented above are equally parsimonious, indicating that either scenario is equally probable. Second, if cytohesin diverged from either BIG or GBF, one would expect conservation of at least a portion the HUS or HDS domains to be present in cytohesin as well. Therefore, it is most likely that the first duplication of the Sec7 domain gave rise to one lineage involved in secretion and one lineage involved in endocytosis. Nonetheless, early duplications of ArfGAPs and ArfGEFs into secretory and endosomal clades is consistent with the model proposed in Figure 5-13, where the duplication of the ancestral heterotetramer gave rise to secretory and endocytic forms. This, along with the early duplication of Rabs into largely endo- and exocytic clades, helps to solidify one of the earliest steps in the evolution of the membrane trafficking system.

6.5.2 Placing TSET and COPII on the protocoatomer tree

The analysis of the COPII complex, and the discovery and characterization of TSET, requires an assessment of their relationship with the other protocoatomer domain-containing complexes. The protocoatomer hypothesis postulates a common origin for cellular complexes such as the NPC, COPI, COPII, and AP complexes, based on a shared domain composition consisting of a β-propeller/α-solenoid domain composition, combined with a conserved role in membrane deformation (Devos et al., 2004).

Since the original proposal of the hypothesis, new complexes possessing this domain architecture have been identified. One such complex, the IFT, is responsible for bidirectional transport of cargo along the length of the cilium (Kee and Verhey, 2013). The eukaryotic cilium was originally thought to be derived from the symbiosis of eukaryotic ancestor with a spirochete bacterium (Margulis, 1981), but it has since been shown that no homologues of the IFT can be uniquely identified as spirochete derived, and neither does it posses a double membrane structure as do mitochondria and chloroplasts, nor does it possess an organellar genome, suggesting an autogenous origin. Moreover, structural analyses have identified strong similarity of IFT subunits to α/β'-COPI, clathrin and, more recently, ε-COPI (Avidor-Reiss et al., 2004; Jékely and Arendt, 2006; van Dam et al., 2013). There is also evidence for a role of some nucleoporins in the ciliary pore complex, selectively controlling the entry and exit, like at the nuclear pore. Additionally, a Ran-GTP gradient is also though to mediate dissociation of cargo from importins as they cross the ciliary pore, similar to the mechanism used during nucleocytoplasmic transport

(Kee and Verhey, 2013). Overall, the cilium clearly shares an ancestor with the nucleus, as they share similar pore structures, and subunits thereof, and are also linked through the presence of protocoatomer derived complexes, requiring the incorporation of this organelle typically involved in cellular motility, into models of the origin of the endomembrane system.

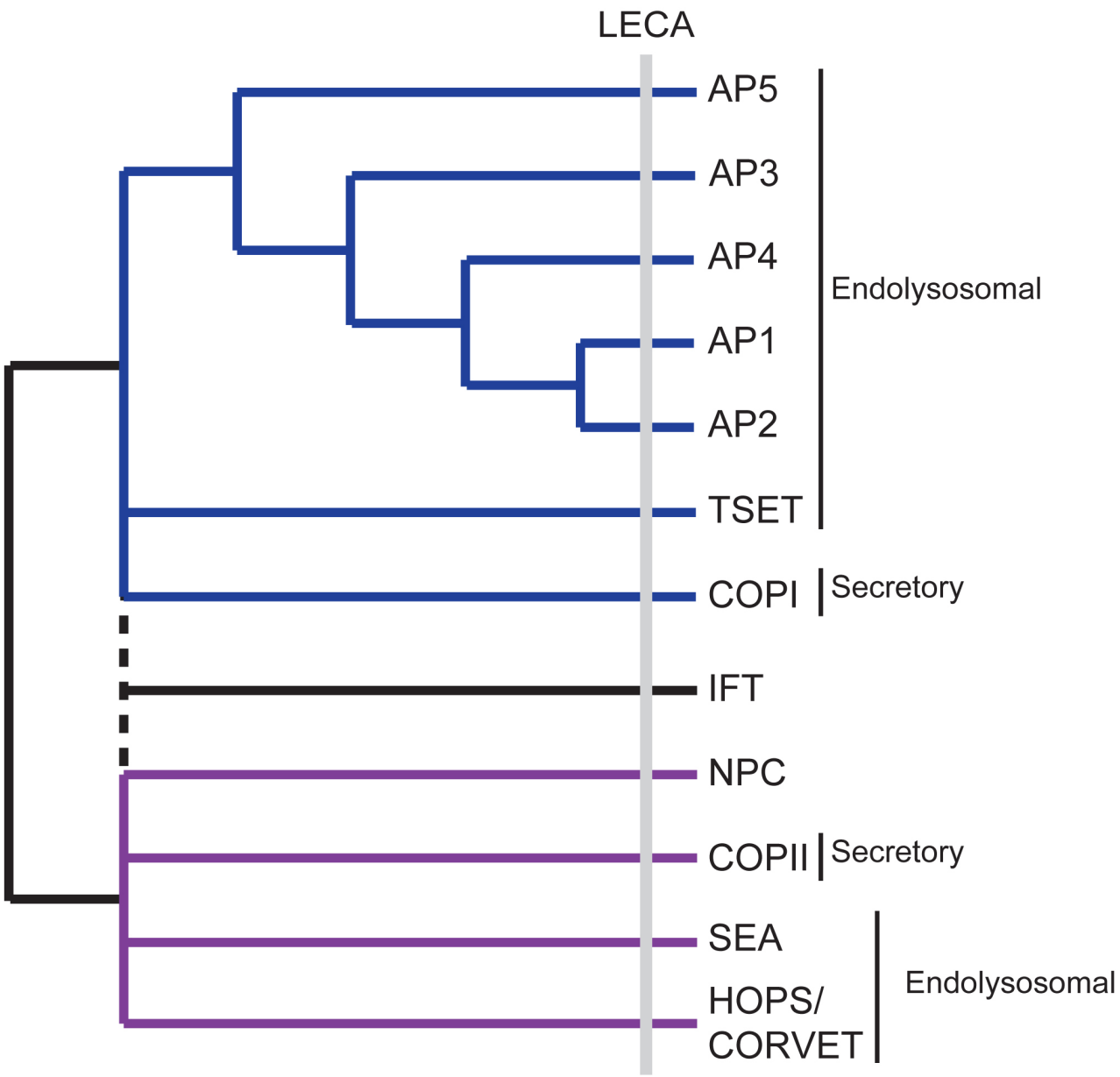
More recently, the SEA complex has been identified as a novel protocoatomer fold-containing complex (Dokudovskaya et al., 2011). The SEA complex is localized to the vacuole in *S. cerevisiae* and is composed of Sea1-4, Seh1, Sec13, Npr2, and Npr3, and have been shown to play a role in intracellular trafficking, amino acid biogenesis, and regulation of TORC1 (Algret et al., 2014). Incorporation of Sec13 and Seh1, in addition to the structural similarity of Sea4 (β-propeller-α-solenoid) to Sec31 and members of the Vps-C core complex of the HOPS and CORVET tethering complexes (notably Vps39), connects the SEA complex to the protocoatomer lineage (Dokudovskaya et al., 2011; Nickerson et al., 2009).

Hypothetically, deducing the relationships between these complexes would uncover the relationships between the organelles of the endomembrane system itself. Therefore, determining the relationships between each of these complexes is integral to a complete understanding of the evolution of the membrane trafficking system. It had previously been suggested that the NPC and COPII are derived from a recent common ancestor based on the shared Sec13 subunit (Devos et al., 2004; Field and Dacks, 2009). More recently, it has been discovered that the yeast nucleoporin Nup145C is similar structurally to Sec31 and in its mechanism of binding to Sec13 (Brohawn and Schwartz, 2009). This finding tightens the

connection between these two membrane deformation complexes. The observation that the NPC, COPII, and the SEA complex share Sec13, and that the NPC and the SEA complex share Seh1, combined with the structural similarity of Nup145C, Sec31, Sea4, and Vps39 unifies these complexes, along with HOPS/CORVET, as a single group.

Similarly, TSET is clearly related to the COPI-AP complexes, forming another group based on the structural and phylogenetic analyses discussed in chapter 5. However, the placement of the last complex, the IFT, is less clear. Comparative genomic and phylogenetic analysis suggested that some members of the IFT complex are distant relatives of the α-, β'- and ε-COPI subunits (van Dam et al., 2013), whereas physical and structural similarities suggest that it is more closely related to the NPC (Kee and Verhey, 2013). However, further analysis will be required to determine the precise phylogenetic position of this important cellular complex. Nonetheless, this analysis suggests the presence of two distinct clades that make up the protocoatomer tree (Figure 6-2), one composed of TSET, COPI, and the AP complexes, unified by the heterotetrameric core. The second clade comprises the NPC, COPII, the SEA complex, and HOPS/CORVET, all unified by the presence of Sec13, Seh1, and an interacting protein structurally similar to Sec31. The IFT complex could be a part of either group. This speculation by no means represents a definitive analysis of the protocoatomer domain-containing proteins; a thorough phylogenetic analysis of these proteins remains to be completed. However, this does provide a set of hypotheses to be tested once the phylogenetic tools and datasets become available.

Figure 6-2. Hypothesized relationships of known protocoatomer domain-containing complexes. COPII, the NPC, the SEA complex, and HOPS/CORVET likely form a single group based on both the shared presence of Sec13 and the presence of subunits that share similar structures to Sec31. COPI, TSET, and the APs form a distinct group based on their shared tetrameric structure. It is currently unclear where the IFT complex is placed, with some suggesting that it is sister to COPI, whereas others suggest that it may share a more recent common ancestor with the COPII-containing clade. Clades can be subdivided into complexes acting in endolysosomal and secretory transport steps. The relationship between the secretory complexes and endolysosomal complexes is suggestive that coat complexes involved in these processes evolved convergently.


LECA
AP5
AP3
AP4
AP1
AP2
TSET
Endolysosomal
COPI
Secretory
IFT
NPC
COPII
Secretory
SEA
HOPS/
CORVET
Endolysosomal

6.5.3 Different steps in the secretory system arose independently from different endolysosomal transport pathways

Segregating the protocoatomer-derived lineages into two groups has interesting implications for the evolution of the membrane trafficking system. Most importantly, secretory and endolysosomal trafficking appears to have evolved independently multiple times. Roles in endolysosomal trafficking have been shown for the AP complexes and for TSET (Gadeyne et al., 2014; Hirst et al., 2014). Functions in secretion have been shown for AP-1 and AP-4; however, given their positions in the AP phylogeny (Figure 6-2), and their role in endosomal transport, these functions are likely to be secondarily derived. By contrast, COPI is the only complex in the heterotetrameric clade (blue, Figure 6-2) that is only involved in a secretory pathway. Regardless of the branching order between the APs, TSET, and COPI, this represents at least one independent acquisition of function in secretion, if not several.

Similarly, the function of the COPII complex also likely represents an independent origin of secretion. In this clade (purple, Figure 6-2), both the SEA complex and HOPS/CORVET complex are both involved in the endolysosomal system, whereas the NPC plays a structural role in maintaining membrane curvature at the NPC. A parsimony based analysis would reason that the ancestor of this clade would also likely have been involved in the endolysosomal system and/or play a structural role in the early membrane trafficking system, indicating that the secretory function of COPII is the result of an acquisition of function, rather than descent from an ancestral state.

It is tempting to speculate 'why' or 'how' this convergence to secretion occurred. One explanation could be the necessity to connect the early secretory system with the endolysosomal system, and could have centered on translation. In prokaryotes, the signal recognition particle (SRP) directs nascent proteins destined for the extracellular space, to the SecYEG translocon. SecYEG is the prokaryote homologue of the eukaryote Sec61 translocon found at the ER (Gorlich et al., 1992; Stirling et al., 1992). In early eukaryotes, it is possible that the Sec61 translocon was still located at the plasma membrane. This would have allowed the evolution of phagocytosis without a developed membrane trafficking system, as Sec61 translocons would have been incorporated into phagosomal membranes allowing the translation of hydrolytic enzymes directly into the phagosome producing a primordial phagosome/lysosome hybrid (Figure 6-3i). At this stage, an early protocoatomer domain-containing protein was likely associated with the primordial phagosome/lysosome, contributing either to membrane deformation during phagocytosis or to tethering of phagosome/lysosome membranes during exocytosis.

Duplication of this complex would have produced a second protocoatomer complex able to induce membrane curvature, resulting in the production of membrane invaginations (Figure 6-3ii). Eventually, these invaginations would have given rise to a stable organelle (Figure 6-3iii). In addition to possessing Sec61-like translocons, this compartment may also have associated with the cell's DNA, eventually giving rise to the nuclear envelope and the ER (Figure 6-3iii). Additional

Figure 6-3. Integration of the evolution of the early secretory system and endolysosomal system. i) Early eukaryotic cell able to undergo phagocytosis, incorporating the Sec61 translocon into the phagosomal membrane (grey rectangle). An early protocoatomer element (dark blue) associates with the phagosome/lysosome, either as a membrane-deforming complex during phagocytosis or as a tethering complex facilitating membrane fusion. ii) Duplication of the protocoatomer ancestor gives rise to a lineage able to generate membrane invaginations. iii) Membrane invaginations give rise to a stable organelle containing the Sec61 translocon and formed an early association with the cell's DNA. iv) The early protocoatomer element gives rise to multiple coats and a novel organelle, forming an early endolysosomal system. Additional duplications of the phagosomal/lysosomal protocoatomer complex gives rise to tethering complexes (*i.e.,* HOPS/CORVET) and the SEA complex. v) Loss of the Sec61 translocon from the plasma membrane, resulting in its relocation to the ER. vi) Additional gene duplications give rise to COPI, COPII, and additional AP complexes, bridging the gap between the ER, the endolysosomal system, and the phagosome. vii) Further diversification of the endolysosomal system. Double grey arrow represents multiple steps, the order of which remains uncertain. Coloured, rounded edges represent distinct protocoatomer complexes. ER = endoplasmic reticulum, NE = nuclear envelope, G = Golgi, Endo = endosomes, TGN = *trans*-Golgi Network.

i
ii
iii
ER/NE
iv
G/Endo/TGN
ER/NE
v
G/Endo./TGN
ER/NE
vi
Endo./TGN
Golgi
ER/NE
vii
Endo.
TGN
Golgi
ER/NE

gene duplications would allow the endolysosomal system to evolve, mediating trafficking to and from the plasma membrane (Figure 6-3iv).

These independent pathways would be able to continue so long as the Sec61 translocon remains in the plasma membrane. Loss of plasma membrane localized translocons, perhaps the result of selection for individuals that utilize translocons that are closer in proximity to the cell's DNA, would result in the cell's reliance on ER localized translocons (Figure 6-3v), selecting for cells able to undergo anterograde transport from the ER to the Golgi, and subsequent trafficking to the phagosome (Figure 6-3vi). In this scenario, COPII would have evolved as a mechanism to transport lytic enzymes from the ER to phagosomes, whereas COPI would have evolved as a mechanism to return lipids and other proteins necessary for transport, back to the ER. From here, expansion of the endolysosomal system would continue, concurrent with the evolution of additional AP complexes (Figure 6-3vii).

Although this hypothesis is highly speculative, it is consistent with the data at hand, and represents an attempt to retrace the steps taken by our eukaryotic forebears during the construction of an important cellular system.

6.6 Conclusion

It is clear from the analyses presented here that the machinery involved in vesicle formation is also very well conserved much like the machinery involved in vesicle fusion. These analyses have permitted speculation on events that occurred in the early evolution of the membrane trafficking system, which can be further tested

through analysis of additional protein families involved in membrane trafficking. This may have been expected, as the formation of vesicles is a requisite step for their fusion. The use of ever increasingly powerful comparative genomic and phylogenetic tools to study the membrane trafficking system in diverse eukaryotes will shed further light onto the origin and evolution of this network of organelles and the eukaryotic cell itself. We are beginning to understand how such complexities arose and that there is an answer to these questions lying in wait. As frequently reminded by Fox Mulder: “The truth is out there” (Carter).

Bibliography

Abascal, F., Zardoya, R. and Posada, D. (2005). ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**, 2104–5.

Abascal, F., Posada, D. and Zardoya, R. (2007). MtArt: A new model of amino acid replacement for Arthropoda. *Mol. Biol. Evol.* **24**, 1–5.

Adachi, J. and Hasegawa, M. (1996). Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J. Mol. Evol.* **42**, 459–468.

Adachi, J., Waddell, P. J., Martin, W. and Hasegawa, M. (2000). Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. *J. Mol. Evol.* **50**, 348–358.

Adl, S. M., Simpson, A. G. B., Farmer, M. a, Andersen, R. a, Anderson, O. R., Barta, J. R., Bowser, S. S., Brugerolle, G., Fensome, R. a, Fredericq, S., et al. (2005). The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *J. Eukaryot. Microbiol.* **52**, 399–451.

Adl, S. M., Simpson, A. G. B., Lane, C. E., Lukeš, J., Bass, D., Bowser, S. S., Brown, M. W., Burki, F., Dunthorn, M., Hampl, V., et al. (2012). The revised classification of eukaryotes. *J. Eukaryot. Microbiol.* **59**, 429–93.

Adung'a, V. O., Gadelha, C. and Field, M. C. (2013). Proteomic analysis of clathrin interactions in trypanosomes reveals dynamic evolution of endocytosis. *Traffic* **14**, 440–57.

Algret, R., Fernandez-Martinez, J., Shi, Y., Kim, S., Pellarin, R., Cimermancic, P., Cochet, E., Sali, A., Chait, B. T., Rout, M. P., et al. (2014). Molecular Architecture and Function of the SEA Complex, a Modulator of the TORC1 Pathway. *Mol. Cell. Proteomics* **13**, 2855–70.

Allen, C. L., Goulding, D. and Field, M. C. (2003). Clathrin-mediated endocytosis is essential in Trypanosoma brucei. *EMBO J.* **22**, 4991–5002.

Altschul, S. F., Gish, W., Miller, W., Meyers, E. W. and Lipman, D. J. (1990). Basic Local Alignment Search Tool. *J. Mol. Biol.* **215**, 403–410.

Altschul, S. F., Madden, T. L., Schäffer, a a, Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–402.

Andag, U. and Schmitt, H. D. (2003). Dsl1p, an Essential Component of the Golgi-Endoplasmic Reticulum Retrieval System in Yeast, Uses the Same Sequence Motif to Interact with Different Subunits of the COPI Vesicle Coat. *J. Biol. Chem.*

278, 51722–51734.

Antonny, B., Madden, D., Hamamoto, S., Orci, L. and Schekman, R. (2001). Dynamics of the COPII coat with GTP and stable analogues. *Nat. Cell Biol.* **3**, 531–537.

Archibald, J. M., Rogers, M. B., Toop, M., Ishida, K.-I. and Keeling, P. J. (2003). Lateral gene transfer and the evolution of plastid-targeted proteins in the secondary plastid-containing alga Bigelowiella natans. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 7678–7683.

Arias-Salgado, E. G., Lizano, S., Sarkar, S., Brugge, J. S., Ginsberg, M. H. and Shattil, S. J. (2003). Src kinase activation by direct interaction with the integrin beta cytoplasmic domain. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 13298–13302.

Arisue, N., Maki, Y., Yoshida, H., Wada, A., Sánchez, L. B., Müller, M. and Hashimoto, T. (2004). Comparative analysis of the ribosomal components of the hydrogenosome-containing protist, Trichomonas vaginalis. *J. Mol. Evol.* **59**, 59–71.

Asensio, C. S., Sirkis, D. W., Maas, J. W., Egami, K., To, T. L., Brodsky, F. M., Shu, X., Cheng, Y. and Edwards, R. H. (2013). Self-Assembly of VPS41 Promotes Sorting Required for Biogenesis of the Regulated Secretory Pathway. *Dev. Cell* **27**, 425–437.

Avidor-Reiss, T., Maer, A. M., Koundakjian, E., Polyanovsky, A., Keil, T., Subramaniam, S. and Zuker, C. S. (2004). Decoding Cilia Function: Defining Specialized Genes Required for Compartmentalized Cilia Biogenesis. *Cell* **117**, 527–539.

Baldauf, S. L. (2000). A Kingdom-Level Phylogeny of Eukaryotes Based on Combined Protein Data. *Science (80-.).* **290**, 972–977.

Baldauf, S. L. and Palmer, J. D. (1993). Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. *Proc. Natl. Acad. Sci. U. S. A.* **90**, 11558–11562.

Bannykh, S. I., Rowe, T. and Balch, W. E. (1996). The organization of endoplasmic reticulum export complexes. *J. Cell Biol.* **135**, 19–35.

Bapteste, E., Brinkmann, H., Lee, J. a, Moore, D. V, Sensen, C. W., Gordon, P., Duruflé, L., Gaasterland, T., Lopez, P., Müller, M., et al. (2002). The analysis of 100 genes supports the grouping of three highly divergent amoebae: Dictyostelium, Entamoeba, and Mastigamoeba. *Proc. Natl. Acad. Sci. U. S. A.* **99**,

1414–1419.

Barlow, L. D., Dacks, J. B. and Wideman, J. G. (2014). From all to (nearly) none: Tracing adaptin evolution in Fungi. *Cell. Logist.* **4**, e28114.

Barlowe, C. and Schekman, R. (1993). SEC12 encodes a guanine-nucleotide-exchange factor essential for transport vesicle budding from the ER. *Nature* **365**, 347–349.

Barlowe, C., Orci, L., Yeung, T., Hosobuchi, M., Hamamoto, S., Salama, N., Rexach, M. F., Ravazzola, M., Amherdt, M. and Schekman, R. (1994). COPII: a membrane coat formed by Sec proteins that drive vesicle budding from the endoplasmic reticulum. *Cell* **77**, 895–907.

Baum, D. and Baum, B. (2014). An inside-out origin for the eukaryotic cell. *BMC Biol.* **12**, 76.

Bell, R. M., Ballas, L. M. and Coleman, R. A. (1981). Lipid Topogenesis. *J. Lipid Res.* **22**, 391–403.

Bergthorsson, U., Adams, K. L., Thomason, B. and Palmer, J. D. (2003). Widespread horizontal transfer of mitochondrial genes in flowering plants. *Nature* **424**, 197–201.

Berke, I. C., Boehmer, T., Blobel, G. and Schwartz, T. U. (2004). Structural and functional analysis of Nup133 domains reveals modular building blocks of the nuclear pore complex. *J. Cell Biol.* **167**, 591–597.

Berriman, M., Ghedin, E., Hertz-Fowler, C., Blandin, G., Renauld, H., Bartholomeu, D. C., Lennard, N. J., Caler, E., Hamlin, N. E., Haas, B., et al. (2005). The genome of the African trypanosome Trypanosoma brucei. *Science* **309**, 416–22.

Bi, X., Corpina, R. A. and Goldberg, J. (2002). Structure of the Sec23/24-Sar1 pre-budding complex of the COPII vesicle coat. *Nature* **419**, 271–7.

Bigay, J., Casella, J.-F., Drin, G., Mesmin, B. and Antonny, B. (2005). ArfGAP1 responds to membrane curvature through the folding of a lipid packing sensor motif. *EMBO J.* **24**, 2244–53.

Blanc, C., Charette, S. J., Mattei, S., Aubry, L., Smith, E. W., Cosson, P. and Letourneur, F. (2009). Dictyostelium Tom1 participates to an ancestral ESCRT-0 complex. *Traffic* **10**, 161–71.

Blobel, G. and Dobberstein, B. (1975). Transfer of Proteins Across Membranes. I.

Presence of Proteolytically Processed and Unprocessed Nascent Immunoglobulin Light Chains on Membrane-Bound Ribosomes of Murine Myeloma. *J. Cell Biol.* **67**, 835–851.

Boehm, M. and Bonifacino, J. S. (2001). Adaptins: The Final Recount. *Mol. Biol. Cell* **12**, 2907–2920.

Boehmer, T., Jeudy, S., Berke, I. C. and Schwartz, T. U. (2008). Structural and Functional Studies of Nup107/Nup133 Interaction and Its Implications for the Architecture of the Nuclear Pore Complex. *Mol. Cell* **30**, 721–731.

Bonen, L., Cunningham, R. S., Gray, M. W. and Doolittle, W. F. (1977). Wheat embryo mitochodrial 18S ribosomal RNA: evidence for its prokaryotic nature. *Nucleic Acids Res.* **4**, 663–671.

Bonfanti, L., Mironov, A. a., Martínez-Menárguez, J. a., Martella, O., Fusella, A., Baldassarre, M., Buccione, R., Geuze, H. J., Mironov, A. a. and Luini, A. (1998). Procollagen traverses the Golgi stack without leaving the lumen of cisternae: Evidence for cisternal maturation. *Cell* **95**, 993–1003.

Bonifacino, J. S. and Glick, B. S. (2004). The mechanisms of vesicle budding and fusion. *Cell* **116**, 153–66.

Bork, P. (1993). Hundreds of ankyrin-like repeats in functionally diverse proteins: Mobile modules that cross phyla horizontally? *Proteins Struct. Funct. Genet.* **17**, 363–374.

Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M. and Yuan, Y. (1998). Predicting function: from genes to genomes and back. *J. Mol. Biol.* **283**, 707–25.

Bouvet, S., Golinelli-Cohen, M.-P., Contremoulins, V. and Jackson, C. L. (2013). Targeting of the Arf-GEF GBF1 to lipid droplets and Golgi membranes. *J. Cell Sci.* **126**, 4794–805.

Brady, R. J., Wen, Y. and O'Halloran, T. J. (2008). The ENTH and C-terminal domains of Dictyostelium epsin cooperate to regulate the dynamic interaction with clathrin-coated pits. *J. Cell Sci.* **121**, 3433–3444.

Briguglio, J. S., Kumar, S. and Turkewitz, A. P. (2013). Lysosomal sorting receptors are essential for secretory granule biogenesis in Tetrahymena. *J. Cell Biol.* **203**, 537–550.

Brohawn, S. G. and Schwartz, T. U. (2009). Molecular architecture of the Nup84-

Nup145C-Sec13 edge element in the nuclear pore complex lattice. *Nat. Struct. Mol. Biol.* **16**, 1173–1177.

Brohawn, S. G., Leksa, N. C., Spear, E. D., Rajashankar, K. R. and Schwartz, T. U. (2008). Structural evidence for common ancestry of the nuclear pore complex and vesicle coats. *Science* **322**, 1369–1373.

Brown, H. a, Gutowski, S., Moomaw, C. R., Slaughter, C. and Sternweis, P. C. (1993). ADP-ribosylation factor, a small GTP-dependent regulatory protein, stimulates phospholipase D activity. *Cell* **75**, 1137–1144.

Brown, M. T., Andrade, J., Radhakrishna, H., Donaldson, J. G., Cooper, J. a and Randazzo, P. a (1998). ASAP1, a phospholipid-dependent arf GTPase-activating protein that associates with and is phosphorylated by Src. *Mol. Cell. Biol.* **18**, 7038–7051.

Brugerolle, G., Bricheux, G., Philippe, H., Coffea, G. and Coffe, G. (2002). Collodictyon triciliatum and Diphylleia rotans (=Aulacomonas submarina) form a new family of flagellates (Collodictyonidae) with tubular mitochondrial cristae that is phylogenetically distant from other flagellate groups. *Protist* **153**, 59–70.

Bui, Q., Golinelli-Cohen, M. and Jackson, C. L. (2009). Large Arf1 guanine nucleotide exchange factors: evolution, domain structure, and roles in membrane trafficking and human disease. *Mol Genet Genomics* **282**, 329–350.

Bultema, J. J., Ambrosio, A. L., Burek, C. L. and Di Pietro, S. M. (2012). BLOC-2, AP-3, and AP-1 proteins function in concert with Rab38 and Rab32 proteins to mediate protein trafficking to lysosome-related organelles. *J. Biol. Chem.* **287**, 19550–19563.

Burgos, P. V, Mardones, G. a, Rojas, A. L., daSilva, L. L. P., Prabhu, Y., Hurley, J. H. and Bonifacino, J. S. (2010). Sorting of the Alzheimer's disease amyloid precursor protein mediated by the AP-4 complex. *Dev. Cell* **18**, 425–436.

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

Burki, F., Shalchian-Tabrizi, K. and Pawlowski, J. (2008). Phylogenomics reveals a new "megagroup" including most photosynthetic eukaryotes. *Biol. Lett.* **4**, 366–9.

Burki, F., Inagaki, Y., Bråte, J., Archibald, J. M., Keeling, P. J., Cavalier-Smith, T.,

Sakaguchi, M., Hashimoto, T., Horak, A., Kumar, S., et al. (2009). Large-scale phylogenomic analyses reveal that two enigmatic protist lineages, telonemia and centroheliozoa, are related to photosynthetic chromalveolates. *Genome Biol. Evol.* **1**, 231–238.

Burki, F., Okamoto, N., Pombert, J.-F. and Keeling, P. J. (2012). The evolutionary history of haptophytes and cryptophytes: phylogenomic evidence for separate origins. *Proc. R. Soc. B.* **279**, 2246–54.

Cai, H., Reinisch, K. and Ferro-Novick, S. (2007). Coats, tethers, Rabs, and SNAREs work together to mediate the intracellular destination of a transport vesicle. *Dev. Cell* **12**, 671–82.

Cao, Y., Adachi, J., Janke, a, Pääbo, S. and Hasegawa, M. (1994). Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: instability of a tree based on a single gene. *J. Mol. Evol.* **39**, 519–527.

Cao, Y., Janke, A., Waddell, P. J., Westerman, M., Takenaka, O., Murata, S., Okada, N., Pääbo, S. and Hasegawa, M. (1998). Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders. *J. Mol. Evol.* **47**, 307–322.

Caporaso, G., Takei, K., Gandy, S., Matteoli, M., Mundigl, O., Greengard, P. and de Camilli, P. (1994). Morphologic and biochemical analysis of the intracellular trafficking of the Alzheimer beta/A4 amyloid precursor protein. *J. Neurosci.* **14**, 3122–38.

Carter, C. X-Files.

Casanova, J. E. (2007). Regulation of Arf activation: the Sec7 family of guanine nucleotide exchange factors. *Traffic* **8**, 1476–85.

Cavalier-Smith, T. (1975). The origin of nuclei and of eukaryotic cells. *Nature* **256**, 463–468.

Cavalier-Smith, T. (1987a). *Evolutionary Biology of Fungi*. (ed. Rayner, A. D. M.), Brasier, C. M.), and Moore, D. M.) Cambridge: Cambridge University Press.

Cavalier-Smith, T. (1987b). Eukaryotes with no mitochondria. *Nature* **326**, 332–333.

Cavalier-Smith, T. (1998). A revised six-kingdom system of life. *Biol. Rev. Camb. Philos. Soc.* **73**, 203–266.

Cavalier-Smith, T. (2002). The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa. *Int. J. Syst. Evol. Microbiol.* **52**, 297–354.

Cavalier-Smith, T. (2003). The excavate protozoan phyla Metamonada Grasse 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.

Cavalier-Smith, T. (2010). Kingdoms Protozoa and Chromista and the eozoan root of the eukaryotic tree. *Biol. Lett.* **6**, 342–5.

Cavalier-Smith, T., Chao, E. E. Y. and Oates, B. (2004). Molecular phylogeny of Amoebozoa and the evolutionary significance of the unikont Phalansterium. *Eur. J. Protistol.* **40**, 21–48.

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

Chaineau, M., Danglot, L., Proux-Gillardeaux, V. and Galli, T. (2008). Role of HRB in clathrin-dependent endocytosis. *J. Biol. Chem.* **283**, 34365–34373.

Chardin, P. and Callebaut, I. (2002). The yeast Sar exchange factor Sec12, and its higher organism orthologs, fold as beta-propellers. *FEBS Lett.* **525**, 171–173.

Chatton, E. (1938). *Titre et travaux scientifique (1906-1937) de Edouard Chatton*. Sette, Sottano, Italy.

Chen, Y. J. and Stevens, T. H. (1996). The VPS8 gene is required for localization and trafficking of the CPY sorting recpetor in Sacchraomyces cerevisiae. *Eur. J. Cell Biol.* **70**, 289–297.

Chen, K.-Y., Tsai, P.-C., Hsu, J.-W., Hsu, H.-C., Fang, C.-Y., Chang, L.-C., Tsai, Y.-T., Yu, C.-J. and Lee, F.-J. S. (2010). Syt1p promotes activation of Arl1p at the late Golgi to recruit Imh1p. *J. Cell Sci.* **123**, 3478–89.

Chong, Y. T., Gidda, S. K., Sanford, C., Parkinson, J., Mullen, R. T. and Goring, D. R. (2010). Characterization of the Arabidopsis thaliana exocyst complex gene families by phylogenetic, expression profiling, and subcellular localization studies. *New Phytol.* **185**, 401–419.

Chun, J., Shapovalova, Z., Dejgaard, S. Y., Presley, J. F. and Melancon, P. (2008). Characterization of Class I and II ADP-Ribosylation Factors (Arfs) in Live Cells:

GDP-bound Class II Arfs Associate with the ER-Golgi Intermediate Compartment Independently of GBF1. *Mol. Biol. Evol.* **19**, 3488–3500.

Claing, a, Chen, W., Miller, W. E., Vitale, N., Moss, J., Premont, R. T. and Lefkowitz, R. J. (2001). beta-Arrestin-mediated ADP-ribosylation factor 6 activation and beta 2-adrenergic receptor endocytosis. *J. Biol. Chem.* **276**, 42509–42513.

Clarke, M., Lohan, A. J., Liu, B., Lagkouvardos, I., Roy, S., Zafar, N., Bertelli, C., Schilde, C., Kianianmomeni, A., Bürglin, T. R., et al. (2013). Genome of Acanthamoeba castellanii highlights extensive lateral gene transfer and early evolution of tyrosine kinase signaling. *Genome Biol.* **14**, R11.

Cockcroft, S., Thomas, G. M., Fensome, A., Geny, B., Cunningham, E., Gout, I., Hiles, I., Totty, N. F., Truong, O. and Husan, J. J. (1994). Phospholipase D: a downstream effector of ARF in granulocytes. *Science* **263**, 523–526.

Cocucci, E., Aguet, F., Boulant, S. and Kirchhausen, T. (2012). The First Five Seconds in the Life of a Clathrin-Coated Pit. *Cell* **150**, 495–507.

Conibear, E. and Stevens, T. H. (2000). Vps52p, Vps53p, and Vps54p form a novel multisubunit complex required for protein sorting at the yeast late Golgi. *Mol. Biol. Cell* **11**, 305–323.

Cosson, P., Démollière, C., Hennecke, S., Duden, R. and Letourneur, F. (1996). Delta- and zeta-COP, two coatomer subunits homologous to clathrin-associated proteins, are involved in ER retrieval. *EMBO J.* **15**, 1792–1798.

Cox, R., Mason-gamer, R. J., Jackson, C. L. and Segev, N. (2004). Phylogenetic Analysis of Sec7-Domain – containing Arf Nucleotide Exchangers. *Mol. Biol. Cell* **15**, 1487–1505.

Cubonova, L., Sandman, K., Hallam, S. J., DeLong, E. F. and Reeve, J. N. (2005). Histones in Cenarchaea. *J. Bacteriol.* **187**, 5482–5485.

Cukierman, E., Huber, I., Rotman, M. and Cassel, D. (1995). The ARF1 GTPase-Activating Protein: Zinc Finger Motif and Golgi Complex Localization. *Science (80-.).* **270**, 1999–2002.

Curtis, B. a, Tanifuji, G., Burki, F., Gruber, A., Irimia, M., Maruyama, S., Arias, M. C., Ball, S. G., Gile, G. H., Hirakawa, Y., et al. (2012). Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. *Nature* **492**, 59–65.

D'Souza-Schorey, C. and Chavrier, P. (2006). ARF proteins: roles in membrane

traffic and beyond. *Nat. Rev. Mol. Cell Biol.* **7**, 347–58.

D'Souza-Schorey, C. and Stahl, P. D. (1995). Myristolylation Is Required for the Intracellular Localization and Endocytic Function of ARF6. *Exp. Cell Res.* **221**, 153–159.

Dacks, J. B. and Doolittle, W. F. (2001). Reconstructing/deconstructing the earliest eukaryotes: how comparative genomics can help. *Cell* **107**, 419–25.

Dacks, J. B. and Doolittle, W. F. (2002). Novel syntaxin gene sequences from Giardia, Trypanosoma and algae: implications for the ancient evolution of the eukaryotic endomembrane system. *J. Cell Sci.* **115**, 1635–42.

Dacks, J. B. and Doolittle, W. F. (2004). Molecular and phylogenetic characterization of syntaxin genes from parasitic protozoa. *Mol. Biochem. Parasitol.* **136**, 123–136.

Dacks, J. B. and Field, M. C. (2004). Eukaryotic Cell Evolution from a Comparative Genomic Perspective : The Endomembrane System. In *In organelles, genomes and eukaryote phylogeny: An evolutionary synthesis in the age of genomics* (ed. Hirt, R.) and Horner, D.), pp. 309–34. London: CRC Press.

Dacks, J. B. and Field, M. C. (2007). Evolution of the eukaryotic membrane-trafficking system: origin, tempo and mode. *J. Cell Sci.* **120**, 2977–85.

Dacks, J. B., Marinets, A., Ford Doolittle, W., Cavalier-Smith, T. and Logsdon, J. M. (2002). Analyses of RNA Polymerase II genes from free-living protists: phylogeny, long branch attraction, and the eukaryotic big bang. *Mol. Biol. Evol.* **19**, 830–40.

Dacks, J. B., Poon, P. P. and Field, M. C. (2008). Phylogeny of endocytic components yields insight into the process of nonendosymbiotic organelle evolution. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 588–93.

Dang, C. C., Le, Q. S., Gascuel, O. and Le, V. S. (2010). FLU, an amino acid substitution model for influenza proteins. *BMC Evol. Biol.* **10**, 99.

Dart, J. K. G., Saw, V. P. J. and Kilvington, S. (2009). Acanthamoeba Keratitis: Diagnosis and Treatment Update 2009. *Am. J. Ophthalmol.* **148**, 487–499.e2.

Davletov, B. and Sudhof, T. (1993). A Single C2 Domain & from Synaptotagmin I Is Sufficient for High Affinity Ca2+/Phospholipid Binding. *J. Biol. Chem.* **268**, 26386–26390.

Dayhoff, M. O., Schwartz, R. M. and Orcutt, B. C. (1978). A model of evolutionary

change in proteins. In *Atlas of protein sequence and structure* (ed. Dayhoff, M. O.), pp. 345–352. Washington, D.C.: National Biomedica Research Foundation.

De Craene, J.-O., Courte, F., Rinaldi, B., Fitterer, C., Herranz, M. C., Schmitt-Keichinger, C., Ritzenthaler, C. and Friant, S. (2014). Study of the plant COPII vesicle coat subunits by functional complementation of yeast Saccharomyces cerevisiae mutants. *PLoS One* **9**, e90072.

de Duve, C. (2007). The origin of eukaryotes: a reappraisal. *Nat. Rev. Genet.* **8**, 395–403.

De Duve, C. and Wattiaux, R. (1966). Functions of Lysosomes. *Annu. Rev. Physiol.* **28**, 435–492.

De Franceschi, N., Wild, K., Schlacht, A., Dacks, J. B., Sinning, I. and Filippini, F. (2014). Longin and GAF Domains: Structural Evolution and Adaptation to the Subcellular Trafficking Machinery. *Traffic* **15**, 104–21.

Deakin, N. O. and Turner, C. E. (2008). Paxillin comes of age. *J. Cell Sci.* **121**, 2435–2444.

Dell'Angelica, E. C., Klumperman, J., Stoorvogel, W. and Bonifacino, J. S. (1998). Association of the AP-3 adaptor complex with clathrin. *Science* **280**, 431–434.

Dell'Angelica, E. C., Mullins, C. and Bonifacino, J. S. (1999a). AP-4, a novel protein complex related to clathrin adaptors. *J. Biol. Chem.* **274**, 7278–7285.

Dell'Angelica, E. C., Shotelersuk, V., Aguilar, R. C., Gahl, W. a. and Bonifacino, J. S. (1999b). Altered trafficking of lysosomal proteins in Hermansky-Pudlak syndrome due to mutations in the ??3A subunit of the AP-3 adaptor. *Mol. Cell* **3**, 11–21.

Delwiche, C. F., Kuhsel, M. and Palmer, J. D. (1995). Phylogenetic Analysis of tufA sequences indicates a Cyanobacterial origin of all plastids. *Mol. Phylogenet. Evol.* **4**, 110–128.

Demmel, L., Melak, M., Kotisch, H., Fendos, J., Reipert, S. and Warren, G. (2011). Differential selection of Golgi proteins by COPII Sec24 isoforms in procyclic Trypanosoma brucei. *Traffic* **12**, 1575–91.

Derelle, R. and Lang, B. F. (2012). Rooting the eukaryotic tree with mitochondrial and bacterial proteins. *Mol. Biol. Evol.* **29**, 1277–89.

Derelle, R., Torruella, G. and Klime, V. (2015). Bacterial proteins pinpoint a single eukaryotic root. 1–7.

Devos, D., Dokudovskaya, S., Alber, F., Williams, R., Chait, B. T., Sali, A. and Rout, M. P. (2004). Components of coated vesicles and nuclear pore complexes share a common molecular architecture. *PLoS Biol.* **2**, e380.

Dias, M., Blanc, C., Thazar-Poulot, N., Ben Larbi, S., Cosson, P. and Letourneur, F. (2013). Dictyostelium ACAP-A is an ArfGAP involved in cytokinesis, cell migration and actin cytoskeleton dynamics. *J Cell Sci* **126**, 756–766.

Diekmann, Y., Seixas, E., Gouw, M., Tavares-Cadete, F., Seabra, M. C. and Pereira-Leal, J. B. (2011). Thousands of rab GTPases for the cell biologist. *PLoS Comput. Biol.* **7**, e1002217.

Dimmic, M. W., Rest, J. S., Mindell, D. P. and Goldstein, R. a. (2002). rtREV: An amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. *J. Mol. Evol.* **55**, 65–73.

Diril, M. K., Wienisch, M., Jung, N., Klingauf, J. and Haucke, V. (2006). Stonin 2 is an AP-2-dependent endocytic sorting adaptor for synaptotagmin internalization and recycling. *Dev. Cell* **10**, 233–244.

Dokudovskaya, S., Waharte, F., Schlessinger, A., Pieper, U., Devos, D. P., Cristea, I. M., Williams, R., Salamero, J., Chait, B. T., Sali, A., et al. (2011). A conserved coatomer-related complex containing Sec13 and Seh1 dynamically associates with the vacuole in Saccharomyces cerevisiae. *Mol. Cell. Proteomics* **10**, M110.006478.

Donaldson, J. G. and Jackson, C. L. (2011). ARF family G proteins and their regulators: roles in membrane transport, development and disease. *Nat. Rev. Mol. Cell Biol.* **12**, 362–375.

Dong, J.-H., Wen, J.-F. and Tian, H.-F. (2007). Homologs of eukaryotic Ras superfamily proteins in prokaryotes and their novel phylogenetic correlation with their eukaryotic analogs. *Gene* **396**, 116–24.

Doolittle, W. F. and Bonen, L. (1981). Molecular sequence data indicating an endosymbiotic origin for plastids. *Ann. N. Y. Acad. Sci.* **361**, 248–259.

Duden, R., Griffiths, G., Frank, R., Argos, P. and Kreis, T. E. (1991). Beta-COP, a 110 kd protein associated with non-clathrin-coated vesicles and the Golgi complex, shows homology to beta-adaptin. *Cell* **64**, 649–65.

Dunphy, J. L., Moravec, R., Ly, K., Lasell, T. K., Melancon, P. and Casanova, J. E. (2006). The Arf6 GEF GEP100/BRAG2 regulates cell adhesion by controlling endocytosis of beta-1 integrins. *Curr. Biol.* **16**, 315–320.

East, M. P. and Kahn, R. a (2011). Models for the functions of Arf GAPs. *Semin. Cell Dev. Biol.* **22**, 3–9.

East, M. P., Bowzard, J. B., Dacks, J. B. and Kahn, R. a (2012). ELMO domains, evolutionary and functional characterization of a novel GTPase-activating protein (GAP) domain for Arf protein family GTPases. *J. Biol. Chem.* **287**, 39538–53.

Eddy, S. R. (1996). Hidden Markov models. *Curr. Opin. Struct. Biol.* **6**, 361–365.

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics* **14**, 755–763.

Eddy, S. R. (2008). A probabilistic model of local sequence alignment that simplifies statistical significance estimation. *PLoS Comput. Biol.* **4**,.

Eddy, S. R. (2009). A new generation of homology search tools based on probabilistic inference. *Genome Inf.* **23**, 205–211.

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195.

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113.

Eggert, U. S., Mitchison, T. J. and Field, C. M. (2006). Animal cytokinesis: from parts list to mechanisms. *Annu. Rev. Biochem.* **75**, 543–566.

Eichinger, L., Pachebat, J. a, Glöckner, G., Rajandream, M., Sucgang, R., Berriman, M., Song, J., Olsen, R., Szafranski, K., Xu, Q., et al. (2005). The genome of the social amoeba Dictyostelium discoideum. *Nature* **435**, 43–57.

El-Kasmi, F., Pacher, T., Strompen, G., Stierhof, Y. D., Müller, L. M., Koncz, C., Mayer, U. and Jürgens, G. (2011). Arabidopsis SNARE protein SEC22 is essential for gametophyte development and maintenance of Golgi-stack integrity. *Plant J.* **66**, 268–279.

Elde, N. C., Long, M. and Turkewitz, A. P. (2007). A role for convergent evolution in the secretory life of cells. *Trends Cell Biol.* **17**, 157–164.

Elias, M., Brighouse, A., Gabernet-Castello, C., Field, M. C. and Dacks, J. B. (2012). Sculpting the endomembrane system in deep time: high resolution phylogenetics of Rab GTPases. *J. Cell Sci.* **125**, 2500–8.

Eme, L., Moreira, D., Talla, E. and Brochier-Armanet, C. (2009). A complex cell division machinery was present in the last common ancestor of eukaryotes. *PLoS One* **4**,.

Espenshade, P., Gimeno, R. E., Holzmacher, E., Teung, P. and Kaiser, C. A. (1995). Yeast Sec16 Gene Encodes a Multidomain Vesicle Coat Protein that Interacts with Sec23p. **131**, 311–324.

Eugster, a, Frigerio, G., Dale, M. and Duden, R. (2000). COP I domains required for coatomer integrity, and novel interactions with ARF and ARF-GAP. *EMBO J.* **19**, 3905–3917.

Fath, S., Mancias, J. D., Bi, X. and Goldberg, J. (2007). Structure and Organization of Coat Proteins in the COPII Cage. *Cell* **129**, 1325–1336.

Felsenstein, J. (1985). Confidence Limits on Phylogenies : An Approach Using the Bootstrap. *Evolution (N. Y).* **39**, 783–791.

Feng, Q., Albeck, J. G., Cerione, R. a. and Yang, W. (2002). Regulation of the Cool/Pix proteins. Key binding partners of the Cdc42/Rac targets, the p21-activated kinases. *J. Biol. Chem.* **277**, 5644–5650.

Field, M. C. and Dacks, J. B. (2009). First and last ancestors: reconstructing evolution of the endomembrane system with ESCRTs, vesicle coat proteins, and nuclear pore complexes. *Curr. Opin. Cell Biol.* **21**, 4–13.

Field, M. C., Gabernet-Castello, C. and Dacks, J. B. (2007a). Reconstructing the evolution of the endocytic system: insights from genomics and molecular cell biology. *Adv. Exp. Med. Biol.* **607**, 84–96.

Field, M. C., Gabernet-Castello, C. and Dacks, J. B. (2007b). Reconstructing the evolution of the endocytic system: insights from genomics and molecular cell biology. In *Eukaryotic Membranes and Cytoskeleton: Origins and Evolution* (ed. Jekely, G.), pp. 84–96.

Field, M. C., Sali, A. and Rout, M. P. (2011). Evolution: On a bender--BARs, ESCRTs, COPs, and finally getting your coat. *J. Cell Biol.* **193**, 963–72.

Field, M. C., Koreny, L. and Rout, M. P. (2014). Enriching the pore: splendid complexity from humble origins. *Traffic* **15**, 141–56.

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Syst. Zool.* **19**, 99–113.

Fölsch, H., Ohno, H., Bonifacino, J. S. and Mellman, I. (1999). A novel clathrin adaptor complex mediates basolateral targeting in polarized epithelial cells. *Cell* **99**, 189–198.

Forterre, P. (2011). A new fusion hypothesis for the origin of Eukarya: better than

previous ones, but probably also wrong. *Res. Microbiol.* **162**, 77–91.

Franco, M., Chardin, P., Chabre, M. and Paris, S. (1996). Myristoylation-facilitated Binding of the G Protein ARF1GDP to Membrane Phospholipids Is Required for Its Activation by a Soluble Nucleotide Exchange Factor. *J. Biol. Chem.* **271**, 1573–1578.

Franco, M., Peters, P. J., Boretto, J., van Donselaar, E., Neri, a, D'Souza-Schorey, C. and Chavrier, P. (1999). EFA6, a sec7 domain-containing exchange factor for ARF6, coordinates membrane recycling and actin cytoskeleton organization. *EMBO J.* **18**, 1480–1491.

Frank, S., Upender, S., Hansen, S. H. and Casanova, J. E. (1998). ARNO is a guanine nucleotide exchange factor for ADP-ribosylation factor 6. *J. Biol. Chem.* **273**, 23–27.

Franzusoff, A., Redding, K., Crosby, J., Fuller, R. S. and Schekman, R. (1991). Localization of components involved in protein transport and processing through the yeast Golgi apparatus. *J. Cell Biol.* **112**, 27–37.

Friedman, J. R., Lackner, L. L., West, M., DiBenedetto, J. R., Nunnari, J. and Voeltz, G. K. (2011). ER Tubules Mark Sites of Mitochondrial Division. *Science (80-.).* **334**, 358–362.

Frigerio, G., Grimsey, N., Dale, M., Majoul, I. and Duden, R. (2007). Two human ARFGAPs associated with COP-I-coated vesicles. *Traffic* **8**, 1644–1655.

Fritz-Laylin, L. K., Prochnik, S. E., Ginger, M. L., Dacks, J. B., Carpenter, M. L., Field, M. C., Kuo, A., Paredez, A., Chapman, J., Pham, J., et al. (2010). The genome of Naegleria gruberi illuminates early eukaryotic versatility. *Cell* **140**, 631–42.

Fromme, J. C., Ravazzola, M., Hamamoto, S., Al-Balwi, M., Eyaid, W., Boyadjiev, S. a., Cosson, P., Schekman, R. and Orci, L. (2007). The Genetic Basis of a Craniofacial Disease Provides Insight into COPII Coat Assembly. *Dev. Cell* **13**, 623–634.

Fuerst, J. a and Webb, R. I. (1991). Membrane-bounded nucleoid in the eubacterium Gemmata obscuriglobus. *Proc. Natl. Acad. Sci. U. S. A.* **88**, 8184–8188.

Gabernet-Castello, C., O'Reilly, A. J., Dacks, J. B. and Field, M. C. (2013). Evolution of Tre-2/Bub2/Cdc16 (TBC) Rab GTPase-activating proteins. *Mol. Biol. Cell* **24**, 1574–83.

Gadeyne, A., Sánchez-Rodríguez, C., Vanneste, S., Di Rubbo, S., Zauber, H., Vanneste, K., Van Leene, J., De Winne, N., Eeckhout, D., Persiau, G., et al. (2014). The TPLATE adaptor complex drives clathrin-mediated endocytosis in plants. *Cell* **156**, 691–704.

Gaidarov, I. and Keen, J. H. (1999). Phosphoinositide-AP-2 interactions required for targeting to plasma membrane clathrin-coated pits. *J. Cell Biol.* **146**, 755–764.

García-Mata, R. and Sztul, E. (2003). The membrane-tethering protein p115 interacts with GBF1, an ARF guanine-nucleotide-exchange factor. *EMBO Rep.* **4**, 320–325.

Geiger, C., Nagel, W., Boehm, T., van Kooyk, Y., Figdor, C. G., Kremmer, E., Hogg, N., Zeitlmann, L., Dierks, H., Weber, K. S., et al. (2000). Cytohesin-1 regulates beta-2 integrin-mediated adhesion through both ARF-GEF function and interaction with LFA-1. *EMBO J.* **19**, 2525–2536.

Gershlick, D. C., de Marcos Lousa, C., Foresti, O., Lee, A. J., Pereira, E. A., daSilva, L. L. P., Bottanelli, F. and Denecke, J. (2014). Golgi-Dependent Transport of Vacuolar Sorting Receptors is Regulated by COPII, AP1, and AP4 Protein Complexes in Tabacco. *Plant Cell* **26**, 1308–1329.

Gillingham, A. K. and Munro, S. (2007a). Identification of a guanine nucleotide exchange factor for Arf3, the yeast orthologue of mammalian Arf6. *PLoS One* **2**, e842.

Gillingham, A. K. and Munro, S. (2007b). The small G proteins of the Arf family and their regulators. *Annu. Rev. Cell Dev. Biol.* **23**, 579–611.

Gimeno, R. E., Espenshade, P. and Kaiser, C. A. (1995). SED4 encodes a yeast endoplasmic reticulum protein that binds Sec16p and participates in vesicle formation. *J. Cell Biol.* **131**, 325–38.

Gimeno, R. E., Espenshade, P. and Kaiser, C. a (1996). COPII coat subunit interactions: Sec24p and Sec23p bind to adjacent regions of Sec16p. *Mol. Biol. Cell* **7**, 1815–23.

Gorlich, D., Prehn, S., Hartmann, E., Kalies, K. U. and Rapoport, T. a. (1992). A mammalian homolog of SEC61p and SECYp is associated with ribosomes and nascent polypeptides during translocation. *Cell* **71**, 489–503.

Gray, M. W. and Doolitle, W. F. (1982). Has the endosymbiont hypothesis been proven ? *Microbiol. Rev.* **46**, 1–42.

Grebe, M., Gadea, J., Steinmann, T., Kientz, M., Rahfeld, J. U., Salchert, K., Koncz, C. and Jürgens, G. (2000). A conserved domain of the arabidopsis GNOM protein mediates subunit interaction and cyclophilin 5 binding. *Plant Cell* **12**, 343–356.

Gribaldo, S., Poole, A. M., Daubin, V., Forterre, P. and Brochier-Armanet, C. (2010). The origin of eukaryotes and their relationship with the Archaea: are we at a phylogenomic impasse? *Nat. Rev. Microbiol.* **8**, 743–752.

Guindon, S. and Gascuel, O. (2003). A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Syst. Biol.* **52**, 696–704.

Gupta, R. S. and Golding, G. B. (1996). The origin of the eukaryotic cell. *Trends Biochem. Sci.* **21**, 166–171.

Haass, C., Koo, E. H., Mellon, A., Hung, A. Y. and Selkoe, D. J. (1992). Targeting of cell-surface beta-amyloid precursor protein to lysosomes: alternative processing into amyloid-bearing fragments. *Nature* **357**, 500–503.

Hall, B., Allen, C. L., Goulding, D. and Field, M. C. (2004). Both of the Rab5 subfamily small GTPases of Trypanosoma brucei are essential and required for endocytosis. *Mol. Biochem. Parasitol.* **138**, 67–77.

Hampl, V., Hug, L., Leigh, J. W., Dacks, J. B., Lang, B. F., Simpson, A. G. B. and Roger, A. J. (2009). Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". *Proc. Natl. Acad. Sci. U. S. A.* **106**, 3859–64.

Hardwick, K. G., Boothroyd, J. C., Rudner, A. D. and Pelham, H. R. B. (1992). Genes that allow yeast cells to the HDEL receptor in the absence of. *EMBO J.* **1**, 4187–4195.

Hasegawa, M., Kishino, H. and Yano, T. A. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**, 160–174.

He, D., Fiz-Palacios, O., Fu, C.-J., Fehling, J., Tsai, C.-C. and Baldauf, S. L. (2014). An Alternative Root for the Eukaryote Tree of Life. *Curr. Biol.* **24**, 465–470.

Heidel, A. J., Lawal, H. M., Felder, M., Schilde, C., Helps, N. R., Tunggal, B., Rivero, F., John, U., Schleicher, M., Eichinger, L., et al. (2011). Phylogeny-wide analysis of social amoeba genomes highlights ancient origins for complex intercellular communication. *Genome Res.* **21**, 1882–1891.

Heldwein, E. E., Macia, E., Wang, J., Yin, H. L., Kirchhausen, T. and Harrison, S. C.

(2004). Crystal structure of the clathrin adaptor protein 1 core. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 14108–14113.

Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 10915–10919.

Henne, W. M., Boucrot, E., Meinecke, M., Evergren, E., Vallis, Y., Mittal, R. and McMahon, H. T. (2010). FCHo proteins are nucleators of clathrin-mediated endocytosis. *Science* **328**, 1281–4.

Henne, W. M., Buchkovich, N. J. and Emr, S. D. (2011). The ESCRT Pathway. *Dev. Cell* **21**, 77–91.

Herman, E. K., Walker, G., van der Giezen, M. and Dacks, J. B. (2011). Multivesicular bodies in the enigmatic amoeboflagellate Breviata anathema and the evolution of ESCRT 0. *J. Cell Sci.* **124**, 613–621.

Hirst, J., Bright, N. a, Rous, B. and Robinson, M. S. (1999). Characterization of a fourth adaptor-related protein complex. *Mol. Biol. Cell* **10**, 2787–2802.

Hirst, J., Barlow, L. D., Francisco, G. C., Sahlender, D. A., Seaman, M. N. J., Dacks, J. B. and Robinson, M. S. (2011). The fifth adaptor protein complex. *PLoS Biol.* **9**, e1001170.

Hirst, J., Irving, C. and Borner, G. H. H. (2013a). Adaptor Protein Complexes AP-4 and AP-5: New Players in Endosomal Trafficking and Progressive Spastic Paraplegia. *Traffic* **14**, 153–164.

Hirst, J., Borner, G. H. H., Edgar, J., Hein, M. Y., Mann, M., Buchholz, F., Antrobus, R. and Robinson, M. S. (2013b). Interaction between AP-5 and the hereditary spastic paraplegia proteins SPG11 and SPG15. *Mol. Biol. Cell* **24**, 2558–69.

Hirst, J., Schlacht, A., Norcott, J. P., Traynor, D., Bloomfield, G., Antrobus, R., Kay, R. R., Dacks, J. B. and Robinson, M. S. (2014). Characterization of TSET, an ancient and widespread membrane trafficking complex. *Elife* **3**, e02866.

Hjort, K., Goldberg, A. V, Tsaousis, A. D., Hirt, R. P. and Embley, T. M. (2010). Diversity and reductive evolution of mitochondria among microbial eukaryotes. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **365**, 713–727.

Hoefen, R. J. and Berk, B. C. (2006). The multifunctional GIT family of proteins. *J. Cell Sci.* **119**, 1469–75.

Hoffman, G. R., Rahl, P. B., Collins, R. N. and Cerione, R. a (2003). Conserved Structural Motifs in Intracellular Trafficking Pathways. *Mol. Cell* **12**, 615–625.

Hollopeter, G., Lange, J. J., Zhang, Y., Vu, T. N., Gu, M., Ailion, M., Lambie, E. J., Slaughter, B. D., Unruh, J. R., Florens, L., et al. (2014). The membrane-associated proteins FCHo and SGIP are allosteric activators of the AP2 clathrin adaptor complex. *Elife* **3**, 1–23.

Honda, a, Nogami, M., Yokozeki, T., Yamazaki, M., Nakamura, H., Watanabe, H., Kawamoto, K., Nakayama, K., Morris, a J., Frohman, M. a, et al. (1999). Phosphatidylinositol 4-phosphate 5-kinase alpha is a downstream effector of the small G protein ARF6 in membrane ruffle formation. *Cell* **99**, 521–532.

Hong, W. and Lev, S. (2014). Tethering the assembly of SNARE complexes. *Trends Cell Biol.* **24**, 35–43.

Hsia, K. C., Stavropoulos, P., Blobel, G. and Hoelz, A. (2007). Architecture of a Coat for the Nuclear Pore Membrane. *Cell* **131**, 1313–1326.

Hughson, F. M. and Reinisch, K. M. (2010). Structure and mechanism in membrane trafficking. *Curr. Opin. Cell Biol.* **22**, 454–460.

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., et al. (2009). InterPro: the integrative protein signature database. *Nucleic Acids Res.* **37**, D211–5.

Hynes, R. O. (2002). Integrins: Bidirectional, allosteric signaling machines. *Cell* **110**, 673–687.

Iinuma, T., Shiga, A., Nakamoto, K., O'Brien, M. B., Aridor, M., Arimitsu, N., Tagaya, M. and Tani, K. (2007). Mammalian Sec16/p250 plays a role in membrane traffic from the endoplasmic reticulum. *J. Biol. Chem.* **282**, 17632–9.

Inoue, H. and Randazzo, P. a (2007). Arf GAPs and their interacting proteins. *Traffic* **8**, 1465–75.

Ismail, S. a, Vetter, I. R., Sot, B. and Wittinghofer, A. (2010). The structure of an Arf-ArfGAP complex reveals a Ca2+ regulatory mechanism. *Cell* **141**, 812–21.

Ivan, V., de Voer, G., Xanthakis, D., Spoorendonk, K. M., Kondylis, V. and Rabouille, C. (2008). Drosophila Sec16 mediates the biogenesis of tER sites upstream of Sar1 through an arginine-rich motif. *Mol. Biol. Cell* **19**, 4352–65.

Jackson, C. L. (2009). Mechanisms of transport through the Golgi complex. *J. Cell Sci.* **122**, 443–452.

Jackson, C. L. (2014). GEF-effector interactions. *Cell. Logist.* **4**, e943616.

Jackson, T. R., Brown, F. D., Nie, Z., Miura, K., Foroni, L., Sun, J., Hsu, V. W.,

Donaldson, J. G. and Randazzo, P. a (2000). ACAPs are arf6 GTPase-activating proteins that function in the cell periphery. *J. Cell Biol.* **151**, 627–38.

Jackson, L. P., Kelly, B. T., McCoy, A. J., Gaffry, T., James, L. C., Collins, B. M., Höning, S., Evans, P. R. and Owen, D. J. (2010). A large-scale conformational change couples membrane recruitment to cargo binding in the AP2 clathrin adaptor complex. *Cell* **141**, 1220–1229.

Jackson, L. P., Kümmel, D., Reinisch, K. M. and Owen, D. J. (2012). Structures and mechanisms of vesicle coat components and multisubunit tethering complexes. *Curr. Opin. Cell Biol.* **24**, 475–483.

Jékely, G. (2003). Small GTPases and the evolution of the eukaryotic cell. *Bioessays* **25**, 1129–38.

Jékely, G. and Arendt, D. (2006). Evolution of intraflagellar transport from coated vesicles and autogenous origin of the eukaryotic cilium. *BioEssays* **28**, 191–198.

Jian, X., Brown, P., Schuck, P., Gruschus, J. M., Balbo, A., Hinshaw, J. E. and Randazzo, P. a (2009). Autoinhibition of Arf GTPase-activating protein activity by the BAR domain in ASAP1. *J. Biol. Chem.* **284**, 1652–63.

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202.

Jones, D. T., Taylor, W. R. and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **8**, 275–282.

Jones, B., Jones, E. L., Bonney, S. a, Patel, H. N., Mensenkamp, A. R., Eichenbaum-Voline, S., Rudling, M., Myrdal, U., Annesi, G., Naik, S., et al. (2003). Mutations in a Sar1 GTPase of COPII vesicles are associated with lipid absorption disorders. *Nat. Genet.* **34**, 29–31.

Jung, N., Wienisch, M., Gu, M., Rand, J. B., Müller, S. L., Krause, G., Jorgensen, E. M., Klingauf, J. and Haucke, V. (2007). Molecular basis of synaptic vesicle cargo recognition by the endocytic sorting adaptor stonin 2. *J. Cell Biol.* **179**, 1497–1510.

Kahn, R. A. and Gilman, A. G. (1984). Purification of a Protein CofactorRequired for ADP-ribosylation of the Stimulatory Regulatory Component of Adenylate Cyclase by Cholera Toxin. *J. Biol. Chem.* **259**, 6228–6234.

Kahn, R. A. and Gilman, A. G. (1986). The Protin Cofactor Necessary for ADP-

ribosylation of Gs by Cholera Toxin is Itself a GTP Binding Protein. *J. Biol. Chem.* **261**, 7906–7911.

Kahn, R. a, Bruford, E., Inoue, H., Logsdon, J. M., Nie, Z., Premont, R. T., Randazzo, P. a, Satake, M., Theibert, A. B., Zapp, M. L., et al. (2008). Consensus nomenclature for the human ArfGAP domain-containing proteins. *J. Cell Biol.* **182**, 1039–44.

Kaiser, C. a. and Schekman, R. (1990). Distinct sets of SEC genes govern transport vesicle formation and fusion early in the secretory pathway. *Cell* **61**, 723–733.

Kam, J. L., Miura, K., Jackson, T. R., Gruschus, J., Roller, P., Stauffer, S., Clark, J., Aneja, R. and Randazzo, P. A. (2000). Phosphoinositide-dependent Activation of the ADP-ribosylation Factor GTPase-activating Protein ASAP1. *J. Biol. Chem.* **275**, 9653–9663.

Kamiguchi, H., Long, K. E., Pendergast, M., Schaefer, a W., Rapoport, I., Kirchhausen, T. and Lemmon, V. (1998). The neural cell adhesion molecule L1 interacts with the AP-2 adaptor and is endocytosed via the clathrin-mediated pathway. *J. Neurosci.* **18**, 5311–5321.

Kanamarlapudi, V. (2014). Exchange Factor EFA6R Requires C-terminal Targeting to the Plasma Membrane to Promote Cytoskeletal Rearrangement through the Activation of ADP-ribosylation Factor 6 (ARF6). *J. Biol. Chem.* **289**, 33378–33390.

Karlin, S. and Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. U. S. A.* **87**, 2264–2268.

Karlin, S. and Altschul, S. F. (1993). Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl. Acad. Sci. U. S. A.* **90**, 5873–5877.

Karpoff, S. A. and Zhukov, B. F. (1986). Ultrastructure and taxonomic position of Apusomonas proboscidea Alexeieff. *Arch Protistenkd* **131**, 13–26.

Kartberg, F., Asp, L., Dejgaard, S. Y., Smedh, M., Fernandez-Rodriguez, J., Nilsson, T. and Presley, J. F. (2010). ARFGAP2 and ARFGAP3 are essential for COPI coat assembly on the Golgi membrane of living cells. *J. Biol. Chem.* **285**, 36709–20.

Katz, L. A. and Grant, J. R. (2015). Taxon-Rich Phylogenomic Analyses Resolve the Eukaryotic Tree of Life and Reveal the Power of Subsampling by Sites. *Syst. Biol.*

64, 406–415.

Katz, L. a., Grant, J. R., Parfrey, L. W. and Burleigh, J. G. (2012). Turning the crown upside down: Gene tree parsimony roots the eukaryotic tree of life. *Syst. Biol.* **61**, 653–660.

Kee, H. L. and Verhey, K. J. (2013). Molecular connections between nuclear and ciliary import processes. *Cilia* **2**, 11.

Keeling, P. J. and Palmer, J. D. (2008). Horizontal gene transfer in eukaryotic evolution. *Nat. Rev. Genet.* **9**, 605–618.

Kelley, L. A. and Sternberg, M. J. E. (2009). Protein structure prediction on the Web: a case study using the Phyre server. *Nat. Protoc.* **4**, 363–71.

Kelley, L. a, Mezulis, S., Yates, C. M., Wass, M. N. and Sternberg, M. J. E. (2015). The Phyre2 web portal for protein modeling , prediction and analysis. *Nat. Protoc.* **10**, 845–858.

Kim, E., Simpson, A. G. B. and Graham, L. E. (2006). Evolutionary relationships of apusomonads inferred from taxon-rich analyses of 6 nuclear encoded genes. *Mol. Biol. Evol.* **23**, 2455–66.

King, N., Westbrook, M. J., Young, S. L., Kuo, A., Abedin, M., Chapman, J., Fairclough, S., Hellsten, U., Isogai, Y., Letunic, I., et al. (2008). The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. *Nature* **451**, 783–8.

Kipreos, E. T. and Pagano, M. (2000). The F-box protein family. *Genome Biol.* **1**, REVIEWS3002.

Klarlund, J. K., Guilherme, a, Holik, J. J., Virbasius, J. V, Chawla, a and Czech, M. P. (1997). Signaling by phosphoinositide-3,4,5-trisphosphate through proteins containing pleckstrin and Sec7 homology domains. *Science* **275**, 1927–1930.

Klein, S., Partisani, M., Franco, M. and Luton, F. (2008). EFA6 facilitates the assembly of the tight junction by coordinating an Arf6-dependent and -independent pathway. *J. Biol. Chem.* **283**, 30129–30138.

Klinger, C. M., Klute, M. J. and Dacks, J. B. (2013). Comparative Genomic Analysis of Multi-Subunit Tethering Complexes Demonstrates an Ancient Pan-Eukaryotic Complement and Sculpting in Apicomplexa. *PLoS One* **8**, 1–15.

Kodera, C., Yorimitsu, T., Nakano, A. and Sato, K. (2011). Sed4p stimulates Sar1p GTP hydrolysis and promotes limited coat disassembly. *Traffic* **12**, 591–9.

Kondo, A., Hashimoto, S., Yano, H., Nagayama, K., Mazaki, Y. and Sabe, H. (2000). A new paxillin-binding protein, PAG3/Papalpha/KIAA0400, bearing an ADP-ribosylation factor GTPase-activating protein activity, is involved in paxillin recruitment to focal adhesions and cell migration. *Mol. Biol. Cell* **11**, 1315–1327.

Kosiol, C. and Goldman, N. (2005). Different versions of the dayhoff rate matrix. *Mol. Biol. Evol.* **22**, 193–199.

Koumandou, V. L., Dacks, J. B., Coulson, R. M. R. and Field, M. C. (2007). Control systems for membrane fusion in the ancestral eukaryote; evolution of tethering complexes and SM proteins. *BMC Evol. Biol.* **7**, 29.

Koumandou, V. L., Klute, M. J., Herman, E. K., Nunez-Miguel, R., Dacks, J. B. and Field, M. C. (2011). Evolutionary reconstruction of the retromer complex and its function in Trypanosoma brucei. *J. Cell Sci.* **124**, 1496–509.

Koumandou, V. L., Wickstead, B., Ginger, M. L., van der Giezen, M., Dacks, J. B. and Field, M. C. (2013). Molecular paleontology and complexity in the last eukaryotic common ancestor. *Crit. Rev. Biochem. Mol. Biol.* **48**, 373–96.

Krogh, A., Brown, M., Mian, I. S., Sjolander, K. and Haussler, D. (1994). Hidden Markov Models in Computational Biology: Applications to Protein Modeling. *J. Mol. Biol.* **235**, 1501–1531.

Krugmann, S., Anderson, K. E., Ridley, S. H., Risso, N., McGregor, a, Coadwell, J., Davidson, K., Eguinoa, a, Ellson, C. D., Lipp, P., et al. (2002). Identification of ARAP3, a novel PI3K effector regulating both Arf and Rho GTPases, by selective capture on phosphoinositide affinity matrices. *Mol. Cell* **9**, 95–108.

Krugmann, S., Williams, R., Stephens, L. and Hawkins, P. T. (2004). ARAP3 Is a PI3K- and Rap-Regulated GAP for RhoA. *Curr. Biol.* **14**, 1380–1384.

Krugmann, S., Andrews, S., Stephens, L. and Hawkins, P. T. (2006). ARAP3 is essential for formation of lamellipodia after growth factor stimulation. *J. Cell Sci.* **119**, 425–32.

Ku, C., Nelson-sathi, S., Roettger, M., Sousa, F. L., Lockhart, P. J., Bryant, D., Hazkani-covo, E., Mcinerney, J. O., Landan, G. and Martin, W. F. (2015). Endosymbiotic origin and differential loss of eukaryotic genes. *Nature* **524**, 427–432.

Kung, L. F., Pagant, S., Futai, E., D'Arcangelo, J. G., Buchanan, R., Dittmar, J. C., Reid, R. J. D., Rothstein, R., Hamamoto, S., Snapp, E. L., et al. (2012). Sec24p and Sec16p cooperate to regulate the GTP cycle of the COPII coat. *EMBO J.* **31**,

1014–27.

Kurz, S. and Tiedtke, A. (1993). The Golgi Apparatus of Tetrahymena thermophila. *J. Eukaryot. Microbiol.* **40**, 10–13.

Ladinsky, M. S., Mastronarde, D. N., McIntosh, J. R., Howell, K. E. and Staehlin, L. a (1999). Golgi structure in three dimensions: functional insights from the {NRK} cell. *J. Cell Biol.* **144**, 1135–1149.

Langhans, M., Marcote, M. J., Pimpl, P., Virgili-López, G., Robinson, D. G. and Aniento, F. (2008). In vivo trafficking and localization of p24 proteins in plant cells. *Traffic* **9**, 770–785.

Langworthy, T. A. and Pond, J. L. (1986). Archaebacterial ether lipids and chemotaxonomy. *Syst. Appl. Microbiol.* **7**, 253–257.

Lartillot, N. and Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* **21**, 1095–1109.

Lartillot, N., Lepage, T. and Blanquart, S. (2009). PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25**, 2286–8.

Le, S. Q. and Gascuel, O. (2008). An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**, 1307–1320.

Le, S. Q., Lartillot, N. and Gascuel, O. (2008). Phylogenetic mixture models for proteins. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*

Lee, C. and Goldberg, J. (2010). Structure of coatomer cage proteins and the relationship among COPI, COPII, and clathrin vesicle coats. *Cell* **142**, 123–32.

Lee, M. C. S., Orci, L., Hamamoto, S., Futai, E., Ravazzola, M. and Schekman, R. (2005). Sar1p N-terminal helix initiates membrane curvature and completes the fission of a COPII vesicle. *Cell* **122**, 605–617.

Legate, K. R., Montañez, E., Kudlacek, O. and Fässler, R. (2006). ILK, PINCH and parvin: the tIPP of integrin signalling. *Nat. Rev. Mol. Cell Biol.* **7**, 20–31.

Leksa, N. C., Brohawn, S. G. and Schwartz, T. U. (2009). The Structure of the Scaffold Nucleoporin Nup120 Reveals a New and Unexpected Domain Architecture. *Structure* **17**, 1082–1091.

Leung, K. F., Dacks, J. B. and Field, M. C. (2008). Evolution of the multivesicular body ESCRT machinery; retention across the eukaryotic lineage. *Traffic* **9**,

1698–716.

Li, Y., Kelly, W. G., Logsdon, J. M., Schurko, A. M., Harfe, B. D., Hill-Harfe, K. L. and Kahn, R. a (2004). Functional genomic analysis of the ADP-ribosylation factor family of GTPases: phylogeny among diverse eukaryotes and function in C. elegans. *FASEB J.* **18**, 1834–50.

Li, J., Peters, P. J., Bai, M., Dai, J., Bos, E., Kirchhausen, T., Kandror, K. V. and Hsu, V. W. (2007). An ACAP1-containing clathrin coat complex for endocytic recycling. *J. Cell Biol.* **178**, 453–464.

Liu, Y., Loijens, J. C., Martin, K. H., Karginov, A. V and Parsons, J. T. (2002). The Associateion of ASAP1, an ADP Riobosylation Factor-GTPase Activating Protein, with Focal Adhesion Kinase Contributes to the Process of Focal Adhesion Assembly. *Mol. Biol. Cell* **13**, 2147–2156.

Liu, L., Liao, H., Castle, A., Zhang, J., Casanova, J., Szabo, G. and Castle, D. (2005a). SCAMP2 Interacts with Arf6 and Phospholipase D1 and Links Their Fucntion to Exocytic Fusion Pore Formaiton in PC12 Cells. *Mol. Biol. Cell* **16**, 4463–4472.

Liu, Y., Huang, C., Huang, K. and Lee, F. S. (2005b). Role for Gcs1p in Regulation of Arl1p at Trans-Golgi Compartments. *Mol. Biol. Cell* **16**, 4024–4033.

Loftus, B., Anderson, I., Davies, R., Alsmark, U. C. M., Samuelson, J., Amedeo, P., Roncaglia, P., Berriman, M., Hirt, R. P., Mann, B. J., et al. (2005). The genome of the protist parasite Entamoeba histolytica. *Nature* **433**, 865–868.

Loo, T., Ng, Y., Lim, L. and Manser, E. (2004). GIT1 Activates p21-Activated Kinase through a Mechanism Independent of p21 Binding GIT1 Activates p21-Activated Kinase through a Mechanism Independent of p21 Binding. **24**, 3849–3859.

Lubián, L. M. (1982). Nannochloropsis gaditana sp . nov., una nueva Eustigmatophyceae marina. **293**, 287–293.

Lynes, E. M. and Simmen, T. (2011). Urban planning of the endoplasmic reticulum (ER): How diverse mechanisms segregate the many functions of the ER. *Biochim. Biophys. Acta - Mol. Cell Res.* **1813**, 1893–1905.

Macia, E., Chabre, M. and Franco, M. (2001). Specificities for the Small G Proteins ARF1 and ARF6 of the Guanine Nucleotide Exchange Factors ARNO and EFA6. *J. Biol. Chem.* **276**, 24925–24930.

Macro, L., Jaiswal, J. K. and Simon, S. M. (2012). Dynamics of clathrin-mediated

endocytosis and its requirement for organelle biogenesis in Dictyostelium. *J. Cell Sci.* **125**, 5721–5732.

Maddison, D. R. and Maddison, W. P. (2005). MacClade 4: Analysis of phylogeny and character evolution.

Maddison, W. P. and Maddison, D. R. (2015). Mesquite: a modular system for evolutionary analysis.

Makarova, K. S., Yutin, N., Bell, S. D. and Koonin, E. V (2010). Evolution of diverse cell division and vesicle formation systems in Archaea. *Nat. Rev. Microbiol.* **8**, 731–741.

Makiuchi, T. and Nozaki, T. (2014). Highly divergent mitochondrion-related organelles in anaerobic parasitic protozoa. *Biochimie* **100**, 3–17.

Manavski, Y., Carmona, G., Bennewitz, K., Tang, Z., Zhang, F., Sakurai, A., Zeiher, A. M., Gutkind, J. S., Li, X., Kroll, J., et al. (2014). Brag2 differentially regulates β1- and β3-integrin-dependent adhesion in endothelial cells and is involved in developmental and pathological angiogenesis. *Basic Res. Cardiol.* **109**, 1–19.

Manna, P. T., Gadelha, C., Puttick, a. E. and Field, M. C. (2015). ENTH and ANTH domain proteins participate in AP2-independent clathrin-mediated endocytosis. *J. Cell Sci.* **128**, 2130–2142.

Manolea, F., Chun, J., Chen, D. W., Clarke, I., Summerfeldt, N., Dacks, J. B. and Melancon, P. (2010). Arf3 Is Activated Uniquely at the trans-Golgi Network by Brefeldin A-inhibited Guanine Nucleotide Exchange Factors. *Mol. Biol. Cell* **21**, 1836–1849.

Manser, E., Loo, T. H., Koh, C. G., Zhao, Z. S., Chen, X. Q., Tan, L., Tan, I., Leung, T. and Lim, L. (1998). PAK kinases are directly coupled to the PIX family of nucleotide exchange factors. *Mol. Cell* **1**, 183–192.

Margulis, L. (1970). *Origin of eukaryotic cells*. New Haven: Yale University Press.

Margulis, L. (1981). *Symbiosis in cell evolution: life and its environment on the early earth*. San Francisco: Freeman, W H.

Martin, W. and Müller, M. (1998). The hydrogen hypothesis for the first eukaryote. *Nature* **392**, 37–41.

Martina, J. A., Bonangelino, C. J., Aguilar, R. C. and Bonifacino, J. S. (2001). Stonin 2: An Adaptor-like Protein that Interacts with Components of the Endocytic Machinery. *J. Cell Biol.* **153**, 1111–1120.

Masuda, M. and Mochizuki, N. (2010). Structural characteristics of BAR domain superfamily to sculpt the membrane. *Semin. Cell Dev. Biol.* **21**, 391–8.

Matsuzaki, M., Misumi, O., Shin-I, T., Maruyama, S., Takahara, M., Miyagishima, S.-Y., Mori, T., Nishida, K., Yagisawa, F., Nishida, K., et al. (2004). Genome sequence of the ultrasmall unicellular red alga Cyanidioschyzon merolae 10D. *Nature* **428**, 653–7.

Mayer, A., Wickner, W. and Haas, A. (1996). Sec18p (NSF)-driven release of Sec17p (alpha-SNAP) can precede docking and fusion of yeast vacuoles. *Cell* **85**, 83–94.

Mazaki, Y., Hashimoto, S., Okawa, K., Tsubouchi, a, Nakamura, K., Yagi, R., Yano, H., Kondo, a, Iwamatsu, a, Mizoguchi, a, et al. (2001). An ADP-ribosylation factor GTPase-activating protein Git2-short/KIAA0148 is involved in subcellular localization of paxillin and actin cytoskeletal organization. *Mol. Biol. Cell* **12**, 645–662.

McDonold, C. M. and Fromme, J. C. (2014). Four GTPases Differentially Regulate the Sec7 Arf-GEF to Direct Traffic at the trans-Golgi Network. *Dev. Cell* **30**, 759–767.

McMahon, H. T. and Boucrot, E. (2011). Molecular mechanism and physiological functions of clathrin-mediated endocytosis. *Nat. Rev. Mol. Cell Biol.* **12**, 517–533.

McMahon, C., Studer, S. M., Clendinen, C., Dann, G. P., Jeffrey, P. D. and Hughson, F. M. (2012). The structure of Sec12 implicates potassium ion coordination in Sar1 activation. *J. Biol. Chem.* **287**, 43599–606.

Meacci, E., Tsai, S. C., Adamik, R., Moss, J. and Vaughan, M. (1997). Cytohesin-1, a cytosolic guanine nucleotide-exchange protein for ADP-ribosylation factor. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 1745–1748.

Mereschkowsky, C. (1905). Uber Natur und Ursprung der Chromatoporen im Pflanzenreiche. *Biol. Cent.* **25**, 593–604.

Mesmin, B., Drin, G., Levi, S., Rawet, M., Cassel, D., Bigay, J. and Antonny, B. (2007). Two lipid-packing sensor motifs contribute to the sensitivity of ArfGAP1 to membrane curvature. *Biochemistry* **46**, 1779–90.

Meyer, C., Zizioli, D., Lausmann, S., Eskelinen, E. L., Hamann, J., Saftig, P., von Figura, K. and Schu, P. (2000). mu1A-adaptin-deficient mice: lethality, loss of AP-1 binding and rerouting of mannose 6-phosphate receptors. *EMBO J.* **19**,

2193–2203.

Miller, E. A., Beilharz, T. H., Malkus, P. N., Lee, M. C. ., Hamamoto, S., Orci, L. and Schekman, R. (2003). Multiple Cargo Binding Sites on the COPII Subunit Sec24p Ensure Capture of Diverse Membrane Proteins into Transport Vesicles. *Cell* **114**, 497–509.

Miller, M. A., Pfeiffer, W. and Schwartz, T. (2010). Creating the CIPRES Science Gateway for inference of large phylogenetic trees. *2010 Gatew. Comput. Environ. Work.* 1–8.

Minge, M. a, Silberman, J. D., Orr, R. J. S., Cavalier-Smith, T., Shalchian-Tabrizi, K., Burki, F., Skjaeveland, A. and Jakobsen, K. S. (2009). Evolutionary position of breviate amoebae and the primary eukaryote divergence. *Proc. Biol. Sci.* **276**, 597–604.

Mitchell, A., Chang, H.-Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., McAnulla, C., McMenamin, C., Nuka, G., Pesseat, S., et al. (2014). The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.* **43**, D213–D221.

Mitsunari, T., Nakatsu, F., Shioda, N., Love, P. E., Grinberg, A., Bonifacino, J. S. and Ohno, H. (2005). Clathrin adaptor AP-2 is essential for early embryonal development. *Mol. Cell. Biol.* **25**, 9318–9323.

Miura, K., Jacques, K. M., Stauffer, S., Kubosaki, A., Zhu, K., Hirsch, D. S., Resau, J., Zheng, Y. and Randazzo, P. a. (2002). ARAP1: A point of convergence for arf and rho signaling. *Mol. Cell* **9**, 109–119.

Molina, F. I. and Nerad, T. A. (1991). Ultrastructure of Amastigomonas bermudensis ATCC 50234 sp. nov.: A new heterotrophic marine flagellate. *Eur. J. Protistol.* **27**, 386–396.

Moravec, R., Conger, K. K., D'Souza, R., Allison, A. B. and Casanova, J. E. (2012). BRAG2/GEP100/IQSec1 interacts with clathrin and regulates alpha5beta1 integrin endocytosis through activation of ADP ribosylation factor 5 (Arf5). *J. Biol. Chem.* **287**, 31138–31147.

Morre, J. D. and Mollenhauer, H. H. (2007). Microscopic Morphology and the Origins of the Membrane Maturation Model of Golgi Apparatus Function. *Int. Rev. Cytol.* **262**, 191–218.

Mossessova, E., Bickford, L. C. and Goldberg, J. (2003). SNARE Selectivity of the COPII Coat. *Cell* **114**, 483–495.

Mouratou, B., Biou, V., Joubert, A., Cohen, J., Shields, D. J., Geldner, N., Jürgens, G., Melançon, P. and Cherfils, J. (2005). The domain architecture of large guanine nucleotide exchange factors for the small GTP-binding protein Arf. *BMC Genomics* **6**, 20.

Mowbrey, K. and Dacks, J. B. (2009). Evolution and diversity of the Golgi body. *FEBS Lett.* **583**, 3738–45.

Müller, T. and Vingron, M. (2000). Modeling amino acid replacement. *J. Comput. Biol.* **7**, 761–776.

Murungi, E., Barlow, L. D., Venkatesh, D., Adung'a, V. O., Dacks, J. B., Field, M. C. and Christoffels, A. (2014). A comparative analysis of trypanosomatid SNARE proteins. *Parasitol. Int.* **63**, 341–348.

Nagy, V., Hsia, K.-C., Debler, E. W., Kampmann, M., Davenport, A. M., Blobel, G. and Hoelz, A. (2009). Structure of a trimeric nucleoporin complex reveals alternate oligomerization states. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 17693–17698.

Nakamura, N., Hirata, A., Ohsumi, Y. and Wada, Y. (1997). Vam2 / Vps41p and Vam6 / Vps39p Are Components of a Protein Complex on the Vacuolar Membranes and Involved in the Vacuolar Assembly in Yeast Saccharomyces cerevisiae. *J. Biol. Chem.* **272**, 11344–11349.

Natsume, W., Tanabe, K., Kon, S., Yoshida, N., Watanabe, T., Torri, T. and Satake, M. (2006). SMAP2, a Novel ARF GTPase-activating Protein, Interacts with Clathrin and Clathrin Assembly Protein and Functions on the AP1-positive Early Endosome/Trans-Golgi Network. *Mol. Biol. Cell* **17**, 2592–2603.

Neumann, N., Lundin, D. and Poole, A. M. (2010). Comparative genomic evidence for a complete nuclear pore complex in the last eukaryotic common ancestor. *PLoS One* **5**, e13241.

Nickel, W., Weber, T., McNew, J. a, Parlati, F., Söllner, T. H. and Rothman, J. E. (1999). Content mixing and membrane integrity during membrane fusion driven by pairing of isolated v-SNAREs and t-SNAREs. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 12571–12576.

Nickerson, D. P., Brett, C. L. and Merz, A. J. (2009). Vps-C complexes: gatekeepers of endolysosomal traffic. *Curr. Opin. Cell Biol.* **21**, 543–551.

Nickle, D. C., Heath, L., Jensen, M. a., Gilbert, P. B., Mullins, J. I. and Kosakovsky Pond, S. L. (2007). HIV-Specific Probabilistic Models of Protein Evolution. *PLoS*

One **2**,.

Nie, Z. and Randazzo, P. a (2006). Arf GAPs and membrane traffic. *J. Cell Sci.* **119**, 1203–11.

Nie, Z., Stanley, K. T., Stauffer, S., Jacques, K. M., Hirsch, D. S., Takei, J. and Randazzo, P. a (2002). AGAP1, an endosome-associated, phosphoinositide-dependent ADP-ribosylation factor GTPase-activating protein that affects actin cytoskeleton. *J. Biol. Chem.* **277**, 48965–75.

Nie, Z., Boehm, M., Boja, E. S., Vass, W. C., Bonifacino, J. S., Fales, H. M. and Randazzo, P. a. (2003). Specific regulation of the adaptor protein complex AP-3 by the Arf GAP AGAP1. *Dev. Cell* **5**, 513–521.

Nie, Z., Fei, J., Premont, R. T. and Randazzo, P. a (2005). The Arf GAPs AGAP1 and AGAP2 distinguish between the adaptor protein complexes AP-1 and AP-3. *J. Cell Sci.* **118**, 3555–3566.

Nikolopoulos, S. N. and Turner, C. E. (2001). Integrin-linked Kinase (ILK) Binding to Paxillin LD1 Motif Regulates ILK Localization to Focal Adhesions. *J. Biol. Chem.* **276**, 23499–23505.

Nishiya, N., Kiosses, W. B., Han, J. and Ginsberg, M. H. (2005). An alpha4 integrin-paxillin-Arf-GAP complex restricts Rac activation to the leading edge of migrating cells. *Nat. Cell Biol.* **7**, 343–352.

Norman, J. C., Jones, D., Barry, S. T., Holt, M. R., Cockcroft, S. and Critchley, D. R. (1998). ARF1 mediates paxillin recruitment to focal adhesions and potentiates rho-stimulated stress fiber formation in intact and permeabilized swiss 3T3 fibroblasts. *J. Cell Biol.* **143**, 1981–1995.

Novoa, R. R., Calderita, G., Arranz, R., Fontana, J., Granzow, H. and Risco, C. (2005). Virus factories: associations of cell organelles for viral replication and morphogenesis. *Biol. Cell* **97**, 147–172.

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

Ogasawara, M., Kim, S. C., Adamik, R., Togawa, A., Ferrans, V. J., Takeda, K., Kirby, M., Moss, J. and Vaughan, M. (2000). Similarities in function and gene structure of cytohesin-4 and cytohesin-1, guanine nucleotide-exchange proteins for ADP-ribosylation factors. *J. Biol. Chem.* **275**, 3221–3230.

Ooi, C. E., Dell'Angelica, E. C. and Bonifacino, J. S. (1998). ADP-ribosylation factor 1 (ARF1) regulates recruitment of the AP-3 adaptor complex to membranes. *J. Cell Biol.* **142**, 391–402.

Orci, L., Ravazzola, M., Meda, P., Holcomb, C., Moore, H. P., Hicke, L. and Schekman, R. (1991). Mammalian Sec23p homologue is restricted to the endoplasmic reticulum transitional cytoplasm. *Proc. Natl. Acad. Sci. U. S. A.* **88**, 8611–8615.

Ostrowicz, C. W., Bröcker, C., Ahnert, F., Nordmann, M., Lachmann, J., Peplowska, K., Perz, A., Auffarth, K., Engelbrecht-Vandré, S. and Ungermann, C. (2010). Defined Subunit Arrangement and Rab Interactions Are Required for Functionality of the HOPS Tethering Complex. *Traffic* **11**, 1334–1346.

Owen, D. J. and Evans, P. R. (1998). A structural explanation for the recognition of tyrosine-based endocytotic signals. *Science* **282**, 1327–1332.

Owen, D. J., Vallis, Y., Noble, M. E. M., Hunter, J. B., Dafforn, T. R., Evans, P. R. and McMahon, H. T. (1999). A structural explanation for the binding of multiple ligands by the alpha- adaptin appendage domain. *Cell* **97**, 805–815.

Padovani, D., Folly-Klan, M., Labarde, A., Boulakirba, S., Campanacci, V., Franco, M., Zeghouf, M. and Cherfils, J. (2014). EFA6 controls Arf1 and Arf6 activation through a negative feedback loop. *Proc. Natl. Acad. Sci. U. S. A.* 2–7.

Pagano, A., Letourneur, F., Garcia-Estefania, D., Carpentier, J.-L., Orci, L. and Paccaud, J.-P. (1999). Sec24 Proteins and Sorting at the Endoplasmic Reticulum. *J. Biol. Chem.* **274**, 7833–7840.

Page, F. C. (1987). The classification of "naked" amoebae (Phylum Rhizopoda). *Arch. fur Protisenkd.* **133**, 199–217.

Palade, G. (1975). Intracellular aspects of the process of protein synthesis. *Science (80-.).* **189**, 347–358.

Paladino, S., Sarnataro, D., Pillich, R., Tivodar, S., Nitsch, L. and Zurzolo, C. (2004). Protein oligomerization modulates raft partitioning and apical sorting of GPI-anchored proteins. *J. Cell Biol.* **167**, 699–709.

Paleotti, O., Macia, E., Luton, F., Klein, S., Partisani, M., Chardin, P., Kirchhausen, T. and Franco, M. (2005). The Small G-Protein ARF6-GTP Recruits the AP-2 Adaptor Complex to Membranes. *J. Cell Biol.* **280**, 21661–2166.

Parfrey, L. W. and Lahr, D. J. G. (2013). Multicellularity arose several times in the evolution of eukaryotes. *BioEssays* **35**, 339–347.

Parfrey, L. W., Grant, J., Tekle, Y. I., Lasek-Nesselquist, E., Morrison, H. G., Sogin, M. L., Patterson, D. J. and Katz, L. a (2010). Broadly sampled multigene analyses yield a well-resolved eukaryotic tree of life. *Syst. Biol.* **59**, 518–33.

Parsons, J. T., Martin, K. H., Slack, J. K., Taylor, J. M. and Weed, S. a (2000). Focal adhesion kinase: a regulator of focal adhesion dynamics and cell movement. *Oncogene* **19**, 5606–5613.

Payne, W. E., Kaiser, C. a, Bevis, B. J., Soderholm, J., Fu, D., Sears, I. B. and Glick, B. S. (2000). Isolation of Pichia pastoris genes involved in ER-to-Golgi transport. *Yeast* **16**, 979–93.

Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U. S. A.* **85**, 2444–2448.

Peden, A. a., Rudge, R. E., Lui, W. W. Y. and Robinson, M. S. (2002). Assembly and function of AP-3 complexes in cells expressing mutant subunits. *J. Cell Biol.* **156**, 327–336.

Peplowska, K., Markgraf, D. F., Ostrowicz, C. W., Bange, G. and Ungermann, C. (2007). The CORVET Tethering Complex Interacts with the Yeast Rab5 Homolog Vps21 and Is Involved in Endo-Lysosomal Biogenesis. *Dev. Cell* **12**, 739–750.

Pereira-Leal, J. B. and Seabra, M. C. (2001). Evolution of the Rab family of small GTP-binding proteins. *J. Mol. Biol.* **313**, 889–901.

Philippe, H. and Germot, a (2000). Phylogeny of eukaryotes based on ribosomal RNA: long-branch attraction and models of sequence evolution. *Mol. Biol. Evol.* **17**, 830–834.

Plemel, R. L., Lobingier, B. T., Brett, C. L., Angers, C. G., Nickerson, D. P., Paulsel, A., Sprague, D. and Merz, A. J. (2011). Subunit organization and Rab interactions of Vps-C protein complexes that control endolysosomal membrane traffic. *Mol. Biol. Cell* **22**, 1353–1363.

Poole, A. M. and Penny, D. (2007). Evaluating hypotheses for the origin of eukaryotes. *BioEssays* **29**, 74–84.

Premont, R. T., Claing, a, Vitale, N., Freeman, J. L., Pitcher, J. a, Patton, W. a, Moss, J., Vaughan, M. and Lefkowitz, R. J. (1998). beta2-Adrenergic receptor

regulation by GIT1, a G protein-coupled receptor kinase-associated ADP ribosylation factor GTPase-activating protein. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 14082–14087.

Premont, R. T., Perry, S. J., Schmalzigaug, R., Roseman, J. T., Xing, Y. and Claing, A. (2004). The GIT/PIX complex: An oligomeric assembly of GIT family ARF GTPase-activating proteins and PIX family Rac1/Cdc42 guanine nucleotide exchange factors. *Cell. Signal.* **16**, 1001–1011.

Price, D. C., Chan, C. X., Yoon, H. S., Yang, E. C., Qiu, H., Weber, a. P. M., Schwacke, R., Gross, J., Blouin, N. A., Lane, C., et al. (2012). Cyanophora paradoxa Genome Elucidates Origin of Photosynthesis in Algae and Plants. *Science (80-.).* **335**, 843–847.

Pryor, P. R., Jackson, L., Gray, S. R., Edeling, M. a, Thompson, A., Sanderson, C. M., Evans, P. R., Owen, D. J. and Luzio, J. P. (2008). Molecular basis for the sorting of the SNARE VAMP7 into endocytic clathrin-coated vesicles by the ArfGAP Hrb. *Cell* **134**, 817–27.

Puertollano, R. (2005). Interactions of TOM1L1 with the multivesicular body sorting machinery. *J. Biol. Chem.* **280**, 9258–9264.

Pusnik, M., Schmidt, O., Perry, A. J., Oeljeklaus, S., Niemann, M., Warscheid, B., Lithgow, T., Meisinger, C. and Schneider, A. (2011). Mitochondrial preprotein translocase of trypanosomatids has a bacterial origin. *Curr. Biol.* **21**, 1738–1743.

Radakovits, R., Jinkerson, R. E., Fuerstenberg, S. I., Tae, H., Settlage, R. E., Boore, J. L. and Posewitz, M. C. (2012). Draft genome sequence and genetic transformation of the oleaginous alga Nannochloropis gaditana. *Nat. Commun.* **3**, 686.

Radisky, D. C., Snyder, W. B., Emr, S. D. and Kaplan, J. (1997). Characterization of VPS41, a gene required for vacuolar trafficking and high-affinity iron transport in yeast. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 5662–5666.

Ramadas, R. and Thattai, M. (2013). New organelles by gene duplication in a biophysical model of eukaryote endomembrane evolution. *Biophys. J.* **104**, 2553–2563.

Randazzo, P. a., Inoue, H. and Bharti, S. (2007). Arf GAPs as regulators of the actin cytoskeleton. *Biol. Cell* **99**, 583–600.

Rappoport, J. Z. and Simon, S. M. (2009). Endocytic trafficking of activated EGFR is

AP-2 dependent and occurs through preformed clathrin spots. *J. Cell Sci.* **122**, 1301–1305.

Rawet, M., Levi-Tal, S., Szafer-Glusman, E., Parnis, A. and Cassel, D. (2010). ArfGAP1 interacts with coat proteins through tryptophan-based motifs. *Biochem. Biophys. Res. Commun.* **394**, 553–557.

Read, B. A., Kegel, J., Klute, M. J., Kuo, A., Lefebvre, S. C., Maumus, F., Mayer, C., Miller, J., Monier, A., Salamov, A., et al. (2013). Pan genome of the phytoplankton Emiliania underpins its global distribution. *Nature* **499**, 209–13.

Rehling, P., Darsow, T., Katzmann, D. J. and Emr, S. D. (1999). Formation of AP-3 transport intermediates requires Vps41 function. *Nat. Cell Biol.* **1**, 346–353.

Reider, A., Barker, S. L., Mishra, S. K., Im, Y. J., Maldonado-Báez, L., Hurley, J. H., Traub, L. M. and Wendland, B. (2009). Syp1 is a conserved endocytic adaptor that contains domains involved in cargo selection and membrane tubulation. *EMBO J.* **28**, 3103–3116.

Rein, U., Andag, U., Duden, R., Schmitt, H. D. and Spang, A. (2002). ARF-GAP-mediated interaction between the ER-Golgi v-SNAREs and the COPI coat. *J. Cell Biol.* **157**, 395–404.

Remmert, M., Biegert, A., Hauser, A. and Söding, J. (2011). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173–175.

Ren, X., Farías, G. G., Canagarajah, B. J., Bonifacino, J. S. and Hurley, J. H. (2013). Structural basis for recruitment and activation of the AP-1 clathrin adaptor complex by Arf1. *Cell* **152**, 755–767.

Rice, L. M. and Brunger, A. T. (1999). Crystal Structure of the Vesicular Transport Protein Sec17. *Mol. Cell* **4**, 85–95.

Richards, T. a and Cavalier-Smith, T. (2005). Myosin domain evolution and the primary divergence of eukaryotes. *Nature* **436**, 1113–8.

Richardson, B. C., McDonold, C. M. and Fromme, C. J. (2012). The Sec7 Arf-GEF Is Recruited to the trans-Golgi Network by Positive Feedback. *Dev. Cell* **22**, 799–810.

Rizzuto, R., Marchi, S., Bonora, M., Aguiari, P., Bononi, A., De Stefani, D., Giorgi, C., Leo, S., Rimessi, A., Siviero, R., et al. (2009). Ca2+ transfer from the ER to mitochondria: When, how and why. *Biochim. Biophys. Acta - Bioenerg.* **1787**,

1342–1351.

Roberg, K. J., Crotwell, M., Espenshade, P., Gimeno, R. and Kaiser, C. a. (1999). LST1 is a SEC24 homologue used for selective export of the plasma membrane ATPase from the endoplasmic reticulum. *J. Cell Biol.* **145**, 659–672.

Robinson, M. S. and Bonifacino, J. S. (2001). Adaptor-related proteins. *Curr. Opin. Cell Biol.* **13**, 444–453.

Rodríguez-Ezpeleta, N., Brinkmann, H., Burey, S. C., Roure, B., Burger, G., Löffelhardt, W., Bohnert, H. J., Philippe, H. and Lang, B. F. (2005). Monophyly of primary photosynthetic eukaryotes: Green plants, red algae, and glaucophytes. *Curr. Biol.* **15**, 1325–1330.

Rogozin, I. B., Basu, M. K., Csürös, M. and Koonin, E. V (2009). Analysis of rare genomic changes does not support the unikont-bikont phylogeny and suggests cyanobacterial symbiosis as the point of primary radiation of eukaryotes. *Genome Biol. Evol.* **1**, 99–113.

Ronquist, F. and Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574.

Rossi, V., Banfield, D. K., Vacca, M., Dietrich, L. E. P., Ungermann, C., D'Esposito, M., Galli, T. and Filippini, F. (2004). Longins and their longin domains: Regulated SNAREs and multifunctional SNARE regulators. *Trends Biochem. Sci.* **29**, 682–688.

Rutherford, S. and Moore, I. (2002). The Arabidopsis Rab GTPase family: another enigma variation. *Curr. Opin. Plant Biol.* **5**, 518–528.

Sacher, M., Jiang, Y., Barrowman, J., Scarpa, A., Burston, J., Zhang, L., Schieltz, D., Yates, J. R. I. and Abeliovich, H. (1998). TRAPP, a highly conserved novel complex on the cis-Golgi that mediates vesicle docking and fusion. *EMBO J.* **17**, 2494–2503.

Sacher, M., Barrowman, J., Wang, W., Horecka, J., Zhang, Y., Pypaert, M. and Ferro-Novick, S. (2001). TRAPP I implicated in the specificity of tethering in ER-to-Golgi transport. *Mol. Cell* **7**, 433–42.

Saeki, N., Tokuo, H. and Ikebe, M. (2005). BIG1 is a binding partner of myosin IXb and regulates its Rho-GTPase activating protein activity. *J. Biol. Chem.* **280**, 10128–10134.

Saito-Nakano, Y. and Nakano, a (2000). Sed4p functions as a positive regulator of

Sar1p probably through inhibition of the GTPase activation by Sec23p. *Genes Cells* **5**, 1039–48.

Salama, N. R., Yeung, T. and Schekman, R. W. (1993). The Sec 13p complex and reconstitution of vesicle budding from the ER with purified cytosolic proteins. *EMBO J.* **12**, 4073–4082.

Sánchez-Velar, N., Udofia, E. B., Yu, Z. and Zapp, M. L. (2004). hRIP, a cellular cofactor for Rev function, promotes release of HIV RNAs from the perinuclear region. *Genes Dev.* **18**, 23–34.

Sanderfoot, A. (2007). Increases in the number of SNARE genes parallels the rise of multicellularity among the green plants. *Plant Physiol.* **144**, 6–17.

Santarella-Mellwig, R., Franke, J., Jaedicke, A., Gorjanacz, M., Bauer, U., Budd, A., Mattaj, I. W. and Devos, D. P. (2010). The compartmentalized bacteria of the planctomycetes-verrucomicrobia- chlamydiae superphylum have membrane coat-like proteins. *PLoS Biol.* **8**,.

Sapp, J. (2005). The prokaryote-eukaryote dichotomy: meanings and mythology. *Microbiol. Mol. Biol. Rev.* **69**, 292–305.

Sauer, M., Delgadillo, M. O., Zouhar, J., Reynolds, G. D., Pennington, J. G., Jiang, L., Liljegren, S. J., Stierhof, Y.-D., De Jaeger, G., Otegui, M. S., et al. (2013). MTV1 and MTV4 encode plant-specific ENTH and ARF GAP proteins that mediate clathrin-dependent trafficking of vacuolar cargo from the trans-Golgi network. *Plant Cell* **25**, 2217–35.

Scheffzek, K., Ahmadian, M. R. and Wittinghofer, A. (1998). GTPase-activating proteins: Helping hands to complement an active site. *Trends Biochem. Sci.* **23**, 257–262.

Schlacht, A. and Dacks, J. B. (2015). Unexpected ancient paralogues and an evolutionary model for the COPII coat complex. *Genome Biol. Evol.* **7**, 1098–1109.

Schlacht, A., Mowbrey, K., Elias, M., Kahn, R. A. and Dacks, J. B. (2013). Ancient Complexity, Opisthokont Plasticity, and Discovery of the 11th Subfamily of Arf GAP Proteins. *Traffic* **14**, 636–49.

Schlacht, A., Herman, E. K., Klute, M. J., Field, M. C. and Dacks, J. B. (2014). Missing Pieces of an Ancient Puzzle: Evolution of the Eukaryotic Membrane-Trafficking System. In *Cold Spring Harbor perspectives in biology* (ed. Keeling, P. J.) and Koonin, E. V), Cold Spring Harbor Press.

Schledzewski, K., Brinkmann, H. and Mendel, R. R. (1999). Phylogenetic analysis of components of the eukaryotic vesicle transport system reveals a common origin of adaptor protein complexes 1, 2, and 3 and the F subcomplex of the coatomer COPI. *J. Mol. Evol.* **48**, 770–8.

Schneider, N., Schwartz, J. M., Köhler, J., Becker, M., Schwarz, H. and Gerisch, G. (2000). Golvesin-GFP fusions as distinct markers for Golgi and post-Golgi vesicles in Dictyostelium cells. *Biol. Cell* **92**, 495–511.

Sealey-Cardona, M., Schmidt, K., Demmel, L., Hirschmugl, T., Gesell, T., Dong, G. and Warren, G. (2014). Sec16 Determines the Size and Functioning of the Golgi in the Protist Parasite, Trypanosoma brucei. *Traffic* **31**, 1–17.

Seaman, M. N. J. (2012). The retromer complex - endosomal protein recycling and beyond. *J. Cell Sci.* **125**, 4693–702.

Sebé-pedrós, A., Roger, A. J., Lang, F. B., King, N. and Ruiz-trillo, I. (2010). Ancient origin of the integrin-mediated adhesion and signaling machinery. *Proc. Natl. Acad. Sci.*

Sebé-Pedrós, A., Irimia, M., Del Campo, J., Parra-Acero, H., Russ, C., Nusbaum, C., Blencowe, B. J. and Ruiz-Trillo, I. (2013). Regulated aggregative multicellularity in a close unicellular relative of metazoa. *Elife* **2**, e01287.

Serfontein, J., Nisbet, R. E. R., Howe, C. J. and de Vries, P. J. (2010). Evolution of the TSC1/TSC2-TOR signaling pathway. *Sci. Signal.* **3**, ra49.

Shadwick, L. L., Spiegel, F. W., Shadwick, J. D., Brown, M. W. and Silberman, J. D. (2009). Eumycetozoa=Amoebozoa?: SSUrDNA phylo- geny of protosteloid slime molds and its significance for the amoebozoan supergroup. *PLoS One* **4**, e6754.

Shaywitz, D. a., Espenshade, P. J., Gimeno, R. E. and Kaiser, C. a. (1997). COPII Subunit Interactions in the Assembly of the Vesicle Coat. *J. Biol. Chem.* **272**, 25413–25416.

Shiba, Y., Kametaka, S., Waguri, S., Presley, J. F. and Randazzo, P. A. (2013). ArfGAP3 regulates the transport of cation-independent mannose 6-phosphate receptor in the post-golgi compartment. *Curr. Biol.* **23**, 1945–1951.

Shinotsuka, C., Waguri, S., Wakasugi, M., Uchiyama, Y. and Nakayama, K. (2002a). Dominant-negative mutant of BIG2, an ARF-guanine nucleotide exchange factor, specifically affects membrane trafficking from the trans-Golgi network through inhibiting membrane association of AP-1 and GGA coat proteins. *Biochem. Biophys. Res. Commun.* **294**, 254–260.

Shinotsuka, C., Yoshida, Y., Kawamoto, K., Takatsu, H. and Nakayama, K. (2002b). Overexpression of an ADP-ribosylation factor-guanine nucleotide exchange factor, BIG2, uncouples brefeldin A-induced adaptor protein-1 coat dissociation and membrane tubulation. *J. Biol. Chem.* **277**, 9468–9473.

Shirakura, T., Maki, Y., Yoshida, H., Arisue, N., Wada, A., Sánchez, L. B., Nakamura, F., Müller, M. and Hashimoto, T. (2001). Characterization of the ribosomal proteins of the amitochondriate protist, Giardia lamblia. *Mol. Biochem. Parasitol.* **112**, 153–156.

Simmen, T., Höning, S., Icking, A., Tikkanen, R. and Hunziker, W. (2002). AP-4 binds basolateral signals and participates in basolateral sorting in epithelial MDCK cells. *Nat. Cell Biol.* **4**, 154–159.

Simpson, F., Peden, A. A., Christopoulou, L. and Robinson, M. S. (1997). Characterization of the Adaptor-related Protein Complex, AP-3. *J. Cell Biol.* **137**, 835–845.

Simpson, A. G. B., Radek, R., Dacks, J. B. and O'Kelly, C. J. (2002). How Oxymonads Lost Their Groove: An Ultrastructural Comparison of Monocercomonoides and Excavate Taxa. *J. Eukaryot. Microbiol.* **49**, 239–248.

Sjöblom, B., Salmazo, a. and Djinović-Carugo, K. (2008). α-Actinin structure and regulation. *Cell. Mol. Life Sci.* **65**, 2688–2701.

Skruzny, M., Desfosses, A., Prinz, S., Dodonova, S. O., Gieras, A., Uetrecht, C., Jakobi, A. J., Abella, M., Hagen, W. J. H., Schulz, J., et al. (2015). An Organized Co-assembly of Clathrin Adaptors Is Essential for Endocytosis. *Dev. Cell* **33**, 150–162.

Söding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951–960.

Sogin, M. L. (1989). Evolution of Eukaryotic Microorganims and Their Small Subunit Ribosomal RNAs. *Amer. Zool.* **29**, 487–499.

Someya, A., Sata, M., Takeda, K., Pacheco-rodriguez, G., Ferrans, V. J., Moss, J. and Vaughan, M. (2001). ARF-GEP100, a guanine nucleotide-exchange protein for ADP-ribosylation factor 6. *Proc. Natl. Acad. Sci.* **98**, 2413–2418.

Spang, A., Matsuoka, K., Hamamoto, S., Schekman, R. and Orci, L. (1998). Coatomer, Arf1p, and nucleotide are required to bud coat protein complex I-coated vesicles from large synthetic liposomes. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 11199–11204.

Stacey, T. T. I., Nie, Z., Stewart, A., Najdovska, M., Hall, N. E., He, H., Randazzo, P. a and Lock, P. (2004). ARAP3 is transiently tyrosine phosphorylated in cells attaching to fibronectin and inhibits cell spreading in a RhoGAP-dependent manner. *J. Cell Sci.* **117**, 6071–6084.

Stagg, S. M., Gürkan, C., Fowler, D. M., LaPointe, P., Foss, T. R., Potter, C. S., Carragher, B. and Balch, W. E. (2006). Structure of the Sec13/31 COPII coat cage. *Nature* **439**, 234–238.

Stagg, S. M., LaPointe, P., Razvi, A., Gürkan, C., Potter, C. S., Carragher, B. and Balch, W. E. (2008). Structural Basis for Cargo Regulation of COPII Coat Assembly. *Cell* **134**, 474–484.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–90.

Stanier, R. Y. (1970). Some aspects of the biology of cells and their possible evolutionary significance. *Symp. Soc. Gen. Microbiol* **20**, 1–38.

Stanier, R. Y. and van Niel, C. B. (1962). The concept of a bacterium. *Arch. Mikrobiol.* **42**, 17–35.

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

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

Steel, M. A., Lockhart, P. J. and Penny, D. (1993). Confidence in evolutionary trees from biological sequence data. *Nature* **363**, 440–442.

Stenmark, H. (2009). Rab GTPases as coordinators of vesicle traffic. *Nat. Rev. Mol. Cell Biol.* **10**, 513–525.

Stirling, C. J., Rothblatt, J., Hosobuchi, M., Deshaies, R. and Schekman, R. (1992). Protein translocation mutants defective in the insertion of integral membrane proteins into the endoplasmic reticulum. *Mol. Biol. Cell* **3**, 129–142.

Sucgang, R., Kuo, A., Tian, X., Salerno, W., Parikh, A., Feasley, C. L., Dalin, E., Tu, H., Huang, E., Barry, K., et al. (2011). Comparative genomics of the social amoebae Dictyostelium discoideum and Dictyostelium purpureum. *Genome Biol.* **12**, R20.

Sucic, S., El-Kasaby, A., Kudlacek, O., Sarker, S., Sitte, H. H., Marin, P. and Freissmuth, M. (2011). The serotonin transporter is an exclusive client of the coat protein complex II (COPII) component SEC24C. *J. Biol. Chem.* **286**, 16482–

90.

Suga, H., Chen, Z., de Mendoza, A., Sebé-Pedrós, A., Brown, M. W., Kramer, E., Carr, M., Kerner, P., Vervoort, M., Sánchez-Pons, N., et al. (2013). The Capsaspora genome reveals a complex unicellular prehistory of animals. *Nat Commun* **4**,.

Sutton, R. B., Fasshauer, D., Jahn, R. and Brunger, A. T. (1998). Crystal structure of a SNARE complex involved in synaptic ˚ resolution exocytosis at 2 . 4 A. *Nature* **395**, 347–353.

Suzan-Monti, M., La Scola, B., Barrassi, L., Espinosa, L. and Raoult, D. (2007). Ultrastructural characterization of the giant volcano-like virus factory of Acanthamoeba polyphaga Mimivirus. *PLoS One* **2**,.

Tang, B. L., Kausalya, J., Low, D. Y., Lock, M. L. and Hong, W. (1999). A family of mammalian proteins homologous to yeast Sec24p. *Biochem. Biophys. Res. Commun.* **258**, 679–84.

Tatusov, R. L. (1997). A Genomic Perspective on Protein Families. *Science (80-.).* **278**, 631–637.

Teh, O.-K. and Moore, I. (2007). An ARF-GEF acting at the Golgi and in selective endocytosis in polarized plant cells. *Nature* **448**, 493–496.

ter Haar, E., Musacchio, a, Harrison, S. C. and Kirchhausen, T. (1998). Atomic structure of clathrin: a beta propeller terminal domain joins an alpha zigzag linker. *Cell* **95**, 563–573.

TerBush, D. R., Maurice, T., Roth, D. and Novick, P. (1996). The Exocyst is a multiprotein complex required for exocytosis in Saccharomyces cerevisiae. *EMBO J.* **15**, 6483–6494.

Thacker, E., Kearns, B., Chapman, C., Hammond, J., Howell, A. and Theibert, A. (2004). The arf6 GAP centaurin alpha-1 is a neuronal actin-binding protein which also functions via GAP-independent activity to regulate the actin cytoskeleton. *Eur. J. Cell Biol.* **83**, 541–54.

Torruella, G., Derelle, R., Paps, J., Lang, B. F., Roger, A. J., Shalchian-Tabrizi, K. and Ruiz-Trillo, I. (2012). Phylogenetic relationships within the Opisthokonta based on phylogenomic analyses of conserved single-copy protein domains. *Mol. Biol. Evol.* **29**, 531–544.

Traub, L. M., Downs, M. a, Westrich, J. L. and Fremont, D. H. (1999). Crystal

structure of the alpha appendage of AP-2 reveals a recruitment platform for clathrin-coat assembly. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 8907–8912.

Tsuchiya, M., Price, S. R., Tsai, S. C., Moss, J. and Vaughan, M. (1991). Molecular identification of ADP-ribosylation factor mRNAs and their expression in mammalian cells. *J. Biol. Chem.* **266**, 2772–2777.

Turkewitz, A. P. and Bright, L. J. (2011). A Rab-based view of membrane traffic in the ciliate Tetrahymena thermophila. *Small GTPases* **2**, 222–226.

Turner, C. E., Brown, M. C., Perrotta, J. a, Riedy, M. C., Nikolopoulos, S. N., Mcdonald, a R., Bagrodia, S., Thomas, S. and Leventhal, P. S. (1999). Paxillin LD4 Motif Binds PAK and PIX through a Novel 95-kD Ankyrin Repeat, ARF-GAP Protein: A Role in Cytoskeletal Remodeling. *J. Biol. Chem.* **145**, 851–863.

Umasankar, P. K., Ma, L., Thieman, J. R., Jha, A., Doray, B., Watkins, S. C. and Traub, L. M. (2014). A clathrin coat assembly role for the muniscin protein central linker revealed by TALEN-mediated gene editing. *Elife* **3**, 1–33.

Usami, Y., Popov, S. and Gottlinger, H. G. (2014). The Nef-like Effect of Murine Leukemia Virus Glycosylated Gag on HIV-1 Infectivity is Mediated by its Cytoplasmic Domain and Depends on the AP-2 Adaptor Complex. *J Virol* **88**, 3443–3454.

van Dam, T. J. ., Bos, J. L. and Snel, B. (2011). Evolution of the Ras-like small GTPases and their regulators. *Small G* **2**, 4–16.

van Dam, T. J. P., Townsend, M. J., Turk, M., Schlessinger, A., Sali, A., Field, M. C. and Huynen, M. A. (2013). Evolution of modular intraflagellar transport from a coatomer-like progenitor. *PNAS* **110**, 6943–6948.

Van Damme, D., Gadeyne, A., Vanstraelen, M., Inzé, D., Van Montagu, M. C. E., De Jaeger, G., Russinova, E. and Geelen, D. (2011). Adaptin-like protein TPLATE and clathrin recruitment during plant somatic cytokinesis occurs via two distinct pathways. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 615–620.

Vedovato, M., Rossi, V., Dacks, J. B. and Filippini, F. (2009). Comparative analysis of plant genomes allows the definition of the “Phytolongins”: a novel non-SNARE longin domain protein family. *BMC Genomics* **10**, 510.

Veltman, D. M., Auciello, G., Spence, H. J., Machesky, L. M., Rappoport, J. Z. and Insall, R. H. (2011). Functional analysis of Dictyostelium IBARa reveals a conserved role of the I-BAR domain in endocytosis. *Biochem. J.* **436**, 45–52.

Venditti, R., Scanu, T., Santoro, M., Di Tullio, G., Spaar, a., Gaibisso, R., Beznoussenko, G. V., Mironov, a. a., Mironov, a., Zelante, L., et al. (2012). Sedlin Controls the ER Export of Procollagen by Regulating the Sar1 Cycle. *Science (80-.).* **337**, 1668–1672.

Venkateswarlu, K. and Cullen, P. J. (1999). Molecular cloning and functional characterization of a human homologue of centaurin-alpha. *Biochem. Biophys. Res. Commun.* **262**, 237–244.

Venkateswarlu, K., Brandom, K. G. and Lawrence, J. L. (2004). Centaurin-alpha1 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.

Vickerman, K., Darbyshire, J. F. and Ogden, C. G. (1974). Apusomonas proboscidea Alexeieff 1924, an unusual phagotrophic flagellate from soil. *Arch Protistenkd* **116**, 254–269.

Volpicelli-Daley, L. A., Li, Y., Zhang, C.-J. and Kahn, R. A. (2005). Isoform-selective Effects of the Depletion of ADP-Ribosylation Factors 1-5 on Membrane Traffic. *Mol. Biol. Cell* **16**, 4495–4508.

Vossbrinck, C. R. and Woese, C. R. (1986). Eukaryotic ribosomes that lack a 5.8S RNA. *Nature* **320**, 287–288.

Vossbrinck, C. R., Maddox, J. V, Friedman, S., Debrunner-Vossbrinck, B. a and Woese, C. R. (1987). Ribosomal RNA sequence suggests microsporidia are extremely ancient eukaryotes. *Nature* **326**, 411–414.

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

Wang, Y. J., Wang, J., Sun, H. Q., Martinez, M., Sun, Y. X., Macia, E., Kirchhausen, T., Albanesi, J. P., Roth, M. G. and Yin, H. L. (2003). Phosphatidylinositol 4 phosphate regulates targeting of clathrin adaptor AP-1 complexes to the Golgi. *Cell* **114**, 299–310.

Warren, G. and Mellman, I. (2007). Protein Trafficking Between Membranes. In *Cells* (ed. Lewin, B.), Cassimiris, L.), Lingpappa, V. R.), and Plopper, R.), pp. 153–207. Sudbury, USA: Jones and Bartlett Publishers.

Watson, P. J., Frigerio, G., Collins, B. M., Duden, R. and Owen, D. J. (2004). Gamma-COP appendage domain - structure and function. *Traffic* **5**, 79–88.

Weber, T., Zemelman, B. V., McNew, J. a., Westermann, B., Gmachl, M., Parlati, F., Söllner, T. H. and Rothman, J. E. (1998). SNAREpins: Minimal machinery for membrane fusion. *Cell* **92**, 759–772.

Wegener, K. L., Partridge, A. W., Han, J., Pickford, A. R., Liddington, R. C., Ginsberg, M. H. and Campbell, I. D. (2007). Structural Basis of Integrin Activation by Talin. *Cell* **128**, 171–182.

Weimer, C., Beck, R., Eckert, P., Reckmann, I., Moelleken, J., Brügger, B. and Wieland, F. (2008). Differential roles of ArfGAP1, ArfGAP2, and ArfGAP3 in COPI traffi cking. *J. Cell Biol.* **183**, 725–735.

Weissman, J. T., Plutner, H. and Balch, W. E. (2001). The mammalian guanine nucleotide exchange factor mSec12 is essential for activation of the Sar1 GTPase directing endoplasmic reticulum export. *Traffic* **2**, 465–475.

Wendeler, M. W., Paccaud, J.-P. and Hauri, H.-P. (2007). Role of Sec24 isoforms in selective export of membrane proteins from the endoplasmic reticulum. *EMBO Rep.* **8**, 258–64.

West, K. a., Zhang, H., Brown, M. L. C., Nikolopoulos, S. N., Riedy, M. C., Horwitz, A. F. and Turner, C. E. (2001). The LD4 motif of paxillin regulates cell spreading and motility through an interaction with paxillin kinase linker (PKL). *J. Cell Biol.* **154**, 161–176.

Whelan, S. and Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **18**, 691–699.

Whyte, J. R. C. and Munro, S. (2001). The Sec34/35 Golgi Transport Complex Is Related to the Exocyst, Defining a Family of Complexes Involved in Multiple Steps of Membrane Traffic. *Dev. Cell* **1**, 527–537.

Wideman, J. G., Gawryluk, R. M. R., Gray, M. W. and Dacks, J. B. (2013). The ancient and widespread nature of the ER-mitochondria encounter structure. *Mol. Biol. Evol.* **30**, 2044–2049.

Woese, C. R. and Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5088–5090.

Wright, J., Kahn, R. a and Sztul, E. (2014). Regulating the large Sec7 ARF guanine nucleotide exchange factors: the when, where and how of activation. *Cell. Mol. Life Sci.* **71**, 3419–38.

Wuichet, K. and Sogaard-Andersen, L. (2014). Evolution and Diversity of the Ras Superfamily of Small GTPases in Prokaryotes. *Genome Biol. Evol.* **7**, 57–70.

Xia, C., Ma, W., Stafford, L. J., Liu, C., Gong, L., Martin, J. F. and Liu, M. (2003). GGAPs, a new family of bifunctional GTP-binding and GTPase-activating proteins. *Mol. Cell. Biol.* **23**, 2476–2488.

Xiang, Z. (2006). Advances in Homology Protein Structure Modeling. *Curr. Protein Pept. Sci.* **7**, 217–227.

Xie, W. and Sahinidis, N. V. (2006). Residue-rotamer-reduction algorithm for the protein side-chain conformation problem. *Bioinformatics* **22**, 188–194.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39**, 306–314.

Yano, H., Kobayashi, I., Onodera, Y., Luton, F., Franco, M., Mazaki, Y., Hashimoto, S., Iwai, K., Ronai, Z. and Sabe, H. (2008). Fbx8 makes Arf6 refractory to function via ubiquitination. *Mol. Biol. Cell* **19**, 822–832.

Yoon, H.-Y., Miura, K., Cuthbert, E. J., Davis, K. K., Ahvazi, B., Casanova, J. E. and Randazzo, P. a (2006). ARAP2 effects on the actin cytoskeleton are dependent on Arf6-specific GTPase-activating-protein activity and binding to RhoA-GTP. *J. Cell Sci.* **119**, 4650–4666.

Yorimitsu, T., Sato, K. and Takeuchi, M. (2014). Molecular mechanisms of Sar/Arf GTPases in vesicular trafficking in yeast and plants. *Front. Plant Sci.* **5**, 1–12.

Yoshihisa, T., Barlowe, C. and Schekman, R. (1993a). Requirement for a GTPase-activating protein in vesicle budding from the endoplasmic reticulum. *Science* **259**, 1466–1468.

Yoshihisa, T., Barlowe, C. and Schekman, R. (1993b). Requirement for a GTPase-activating protein in vesicle budding from the endoplasmic reticulum. *Science (80-.).* **259**, 1466–8.

Yutin, N., Wolf, M. Y., Wolf, Y. I. and Koonin, E. V (2009). The origins of phagocytosis and eukaryogenesis. *Biol. Direct* **4**, 9.

Zhang, Y., Persson, S., Hirst, J., Robinson, M. S., van Damme, D. and Sánchez-Rodríguez, C. (2015). Change your Tplate, change your fate: plant CME and beyond. *Trends Plant Sci.* **20**, 41–48.

Zhao, X., Claude, A., Chun, J., Shields, D. J., Presley, J. F. and Melançon, P. (2006). GBF1, a cis-Golgi and VTCs-localized ARF-GEF, is implicated in ER-to-Golgi

protein traffic. *J. Cell Sci.* **119**, 3743–3753.

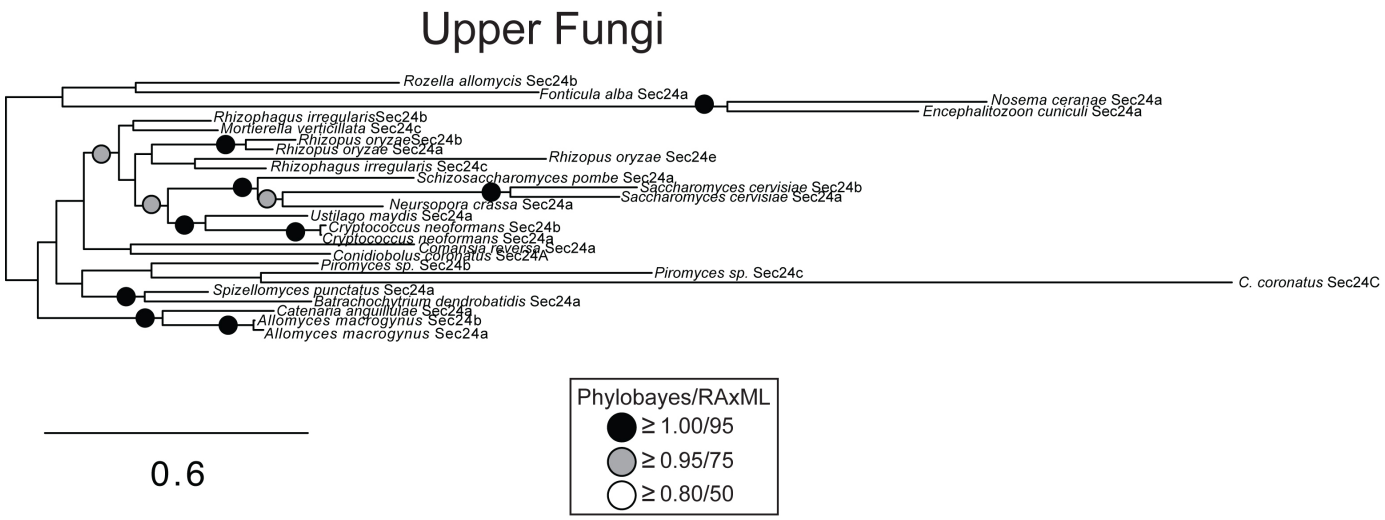
Zhu, Y., Drake, M. T. and Kornfeld, S. (1999). ADP-ribosylation factor 1 dependent clathrin-coat assembly on synthetic liposomes. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 5013–5018.

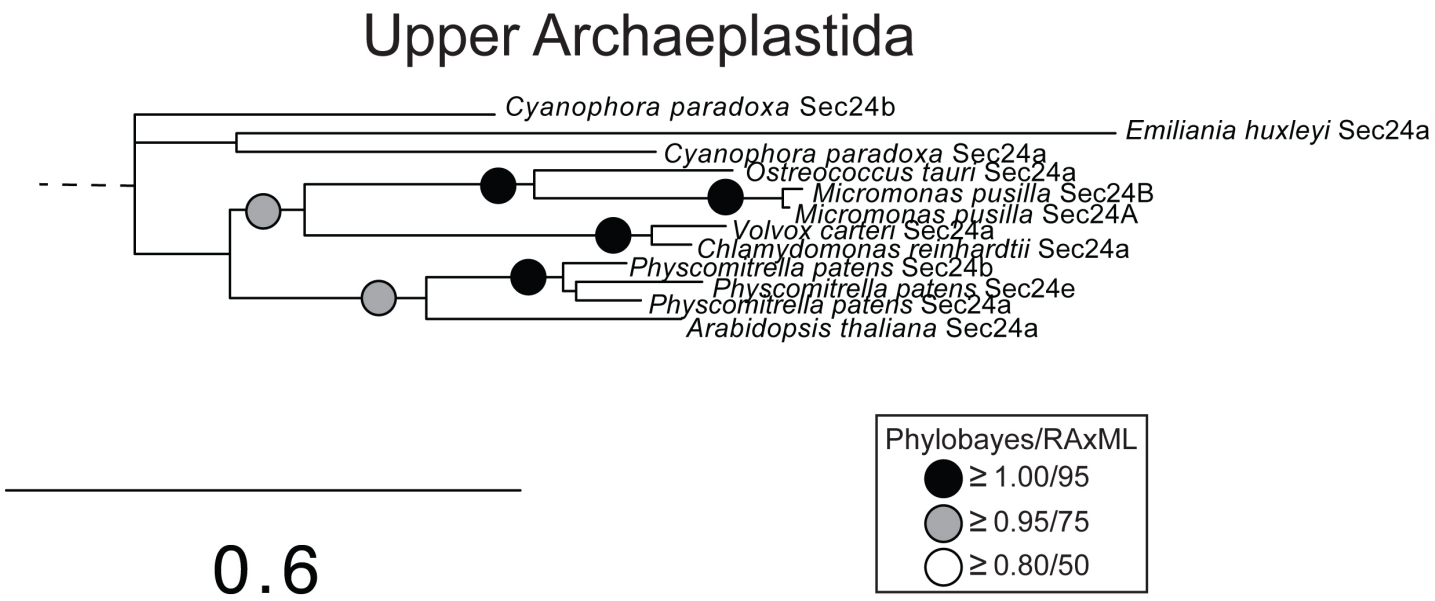
Ziegler, W. H., Liddington, R. C. and Critchley, D. R. (2006). The structure and regulation of vinculin. *Trends Cell Biol.* **16**, 453–460.

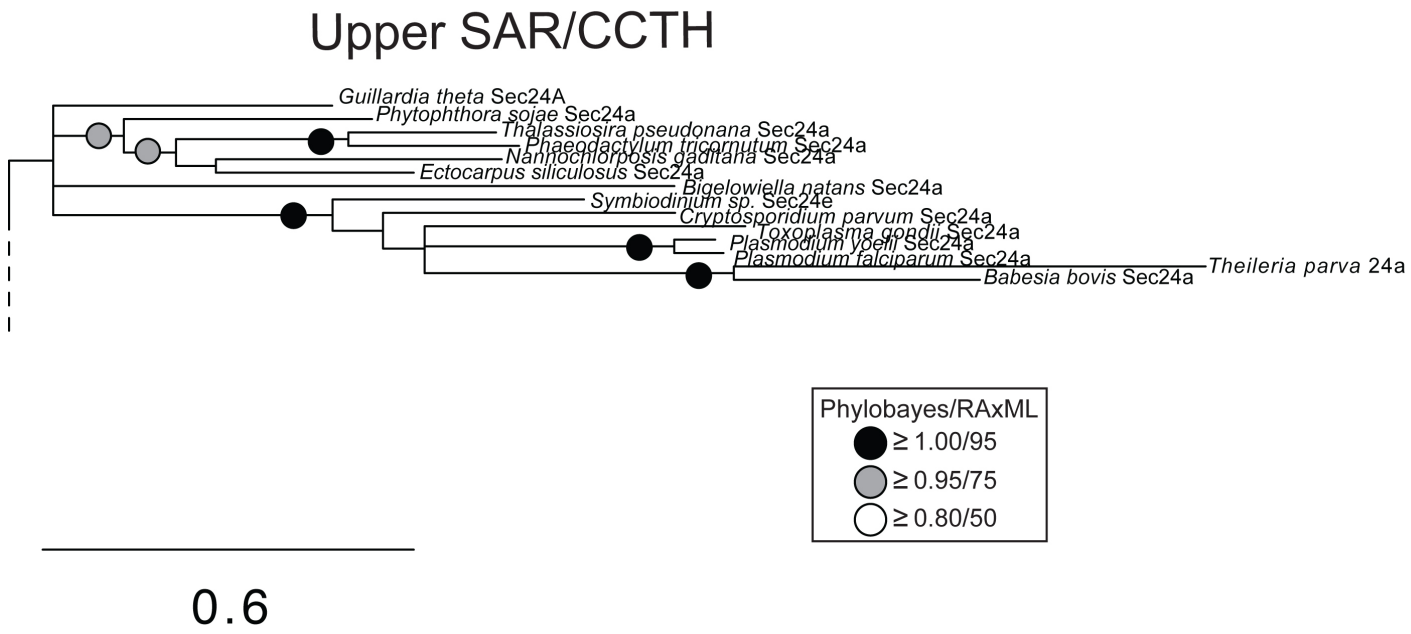
Zizioli, D., Meyer, C., Guhde, G., Saftig, P., von Figura, K. and Schu, P. (1999). Early embryonic death of mice deficient in gamma-adaptin. *J. Biol. Chem.* **274**, 5385–5390.

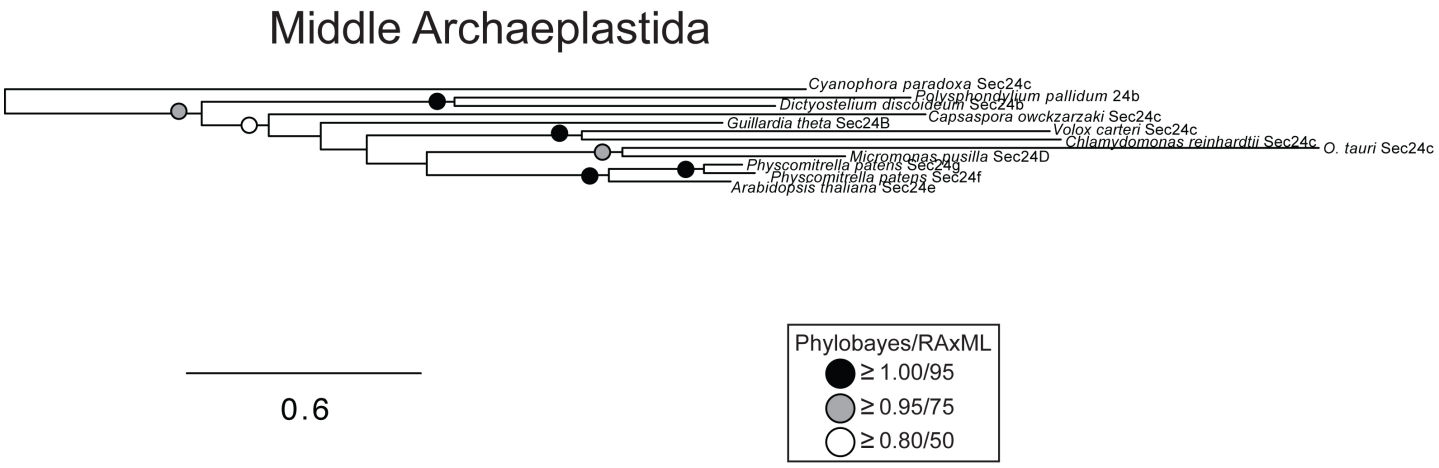
Appendix

The following images are a supplement to Figure 4-9. The large number of sequences included in Figure 4-9 renders it difficult to read the labels attributed to each sequence in the phylogenetic tree. In order to clarify the branching order of sequences in major lineages, each of the major clades were isolated from the image and enlarged here. The order of the clades presented here is from top to bottom of Figure 4-9. Labels are given to clades here (*i.e.,* upper or lower) to denote the relative position of the clade in Figure 4-9.

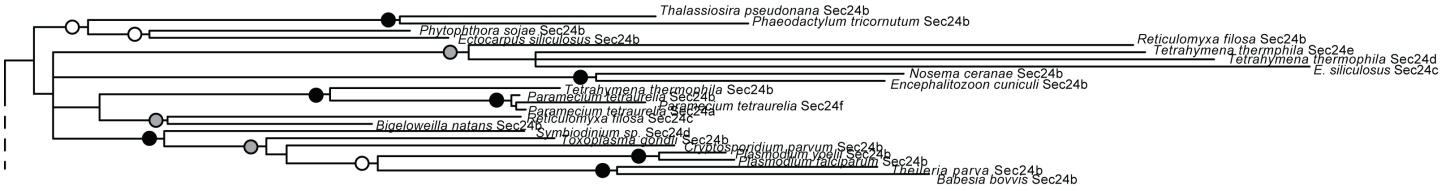

Upper Fungi
Rozella allomycis Sec24b
Fonticula alba Sec24a
Nosema ceranae Sec24a
Encephalitozoon cuniculi Sec24a
Rhizophagus irregularisSec24b
Mortierella verticillata Sec24c
Rhizopus oryzaeSec24b
Rhizopus oryzae Sec24a
Rhizopus oryzae Sec24e
Rhizophagus irregularis Sec24c
Schizosaccharomyces pombe Sec24a
Saccharomyces cervisiae Sec24b
Saccharomyces cervisiae Sec24a
Neursopora crassa Sec24a
Ustilago maydis Sec24a
Cryptococcus neoformans Sec24b
Cryptococcus neoformans Sec24a
Conidiobolus coronatus Sec24A
Piromyces sp. Sec24b
Piromyces sp. Sec24c
C. coronatus Sec24C
Spizellomyces punctatus Sec24a
Batrachochytrium dendrobatidis Sec24a
Catenaria anguillulae Sec24a
Allomyces macrogynus Sec24b
Allomyces macrogynus Sec24a
0.6
Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50


Upper Archaeplastida
Cyanophora paradoxa Sec24b
Emiliania huxleyi Sec24a
Cyanophora paradoxa Sec24a
Ostreococcus tauri Sec24a
Micromonas pusilla Sec24B
Micromonas pusilla Sec24A
Volvox carteri Sec24a
Chlamydomonas reinhardtii Sec24a
Physcomitrella patens Sec24b
Physcomitrella patens Sec24e
Physcomitrella patens Sec24a
Arabidopsis thaliana Sec24a
Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50
0.6


Upper SAR/CCTH
Guillardia theta Sec24A
Phytophthora sojae Sec24a
Thalassiosira pseudonana Sec24a
Phaeodactylum tricornutum Sec24a
Nannochlorpsis gaditana Sec24a
Ectocarpus siliculosus Sec24a
Bigelowiella natans Sec24a
Symbiodinium sp. Sec24e
Cryptosporidium parvum Sec24a
Toxoplasma gondii Sec24a
Plasmodium yoelii Sec24a
Plasmodium falciparum Sec24a
Theileria parva 24a
Babesia bovis Sec24a
Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50
0.6

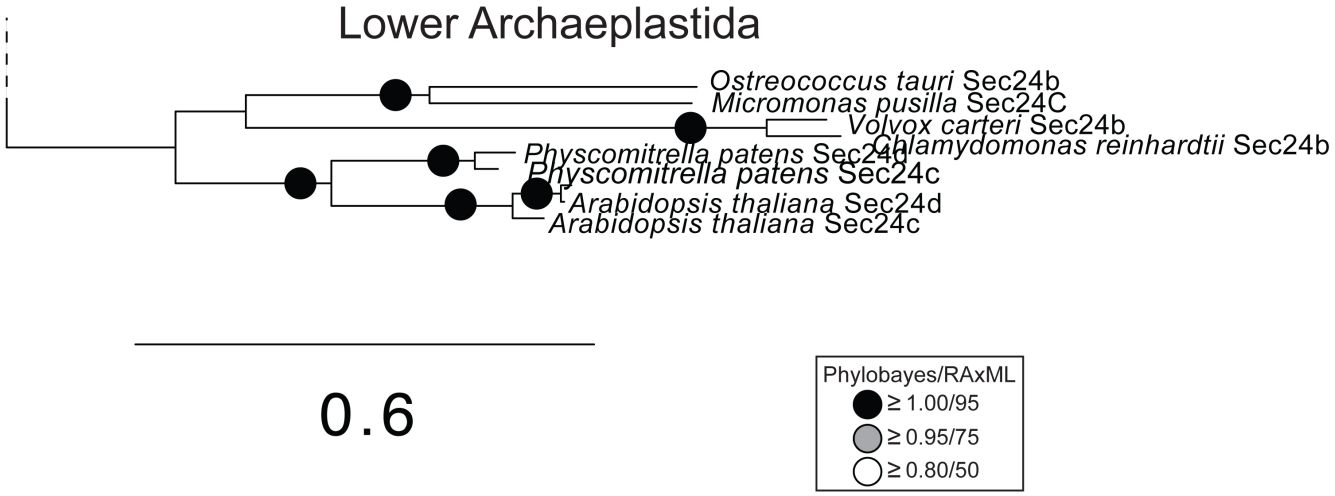

Middle Archaeplastida
Cyanophora paradoxa Sec24c
Polysphondylium pallidum 24b
Dictyostelium discoideum Sec24b
Capsaspora owczarzaki Sec24c
Guillardia theta Sec24B
Volox carteri Sec24c
Chlamydomonas reinhardtii Sec24c
O. tauri Sec24c
Micromonas pusilla Sec24D
Physcomitrella patens Sec24g
Physcomitrella patens Sec24f
Arabidopsis thaliana Sec24e
0.6
Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50

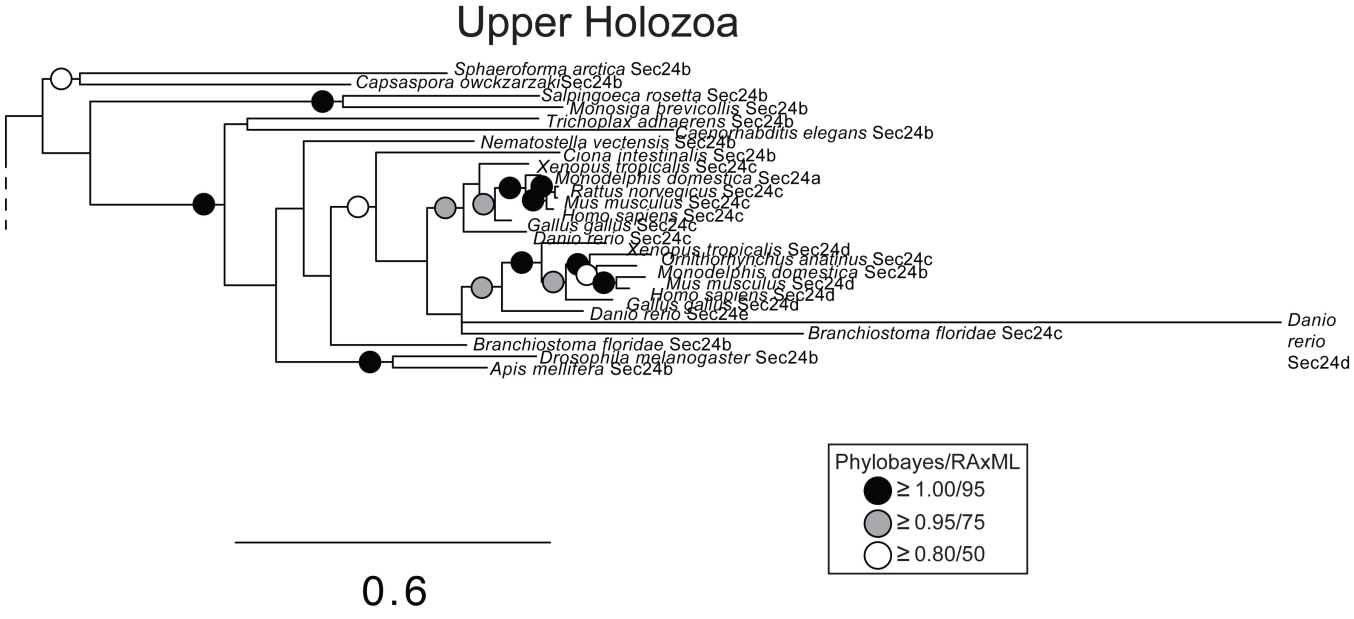
Lower SAR/CCTH

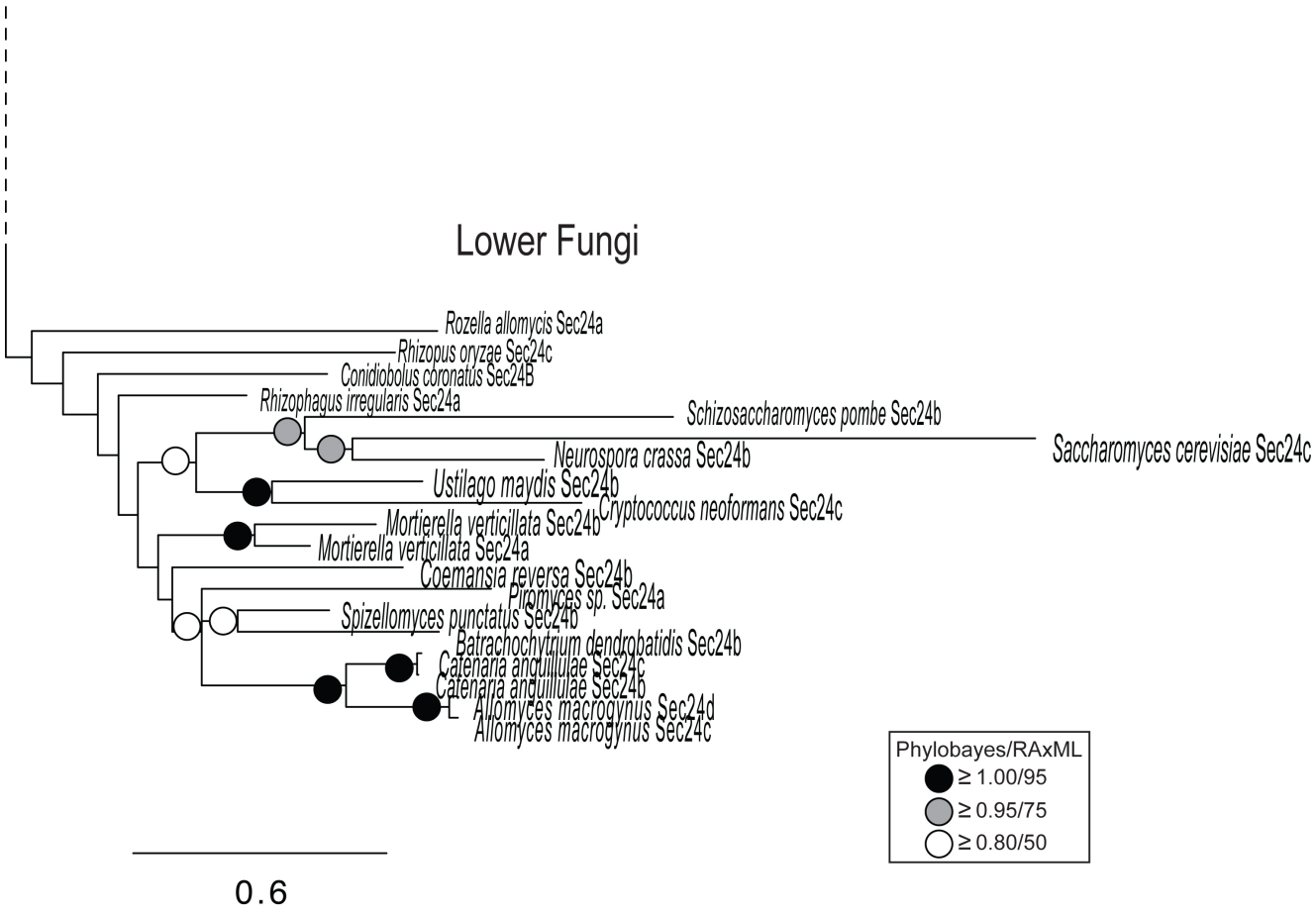


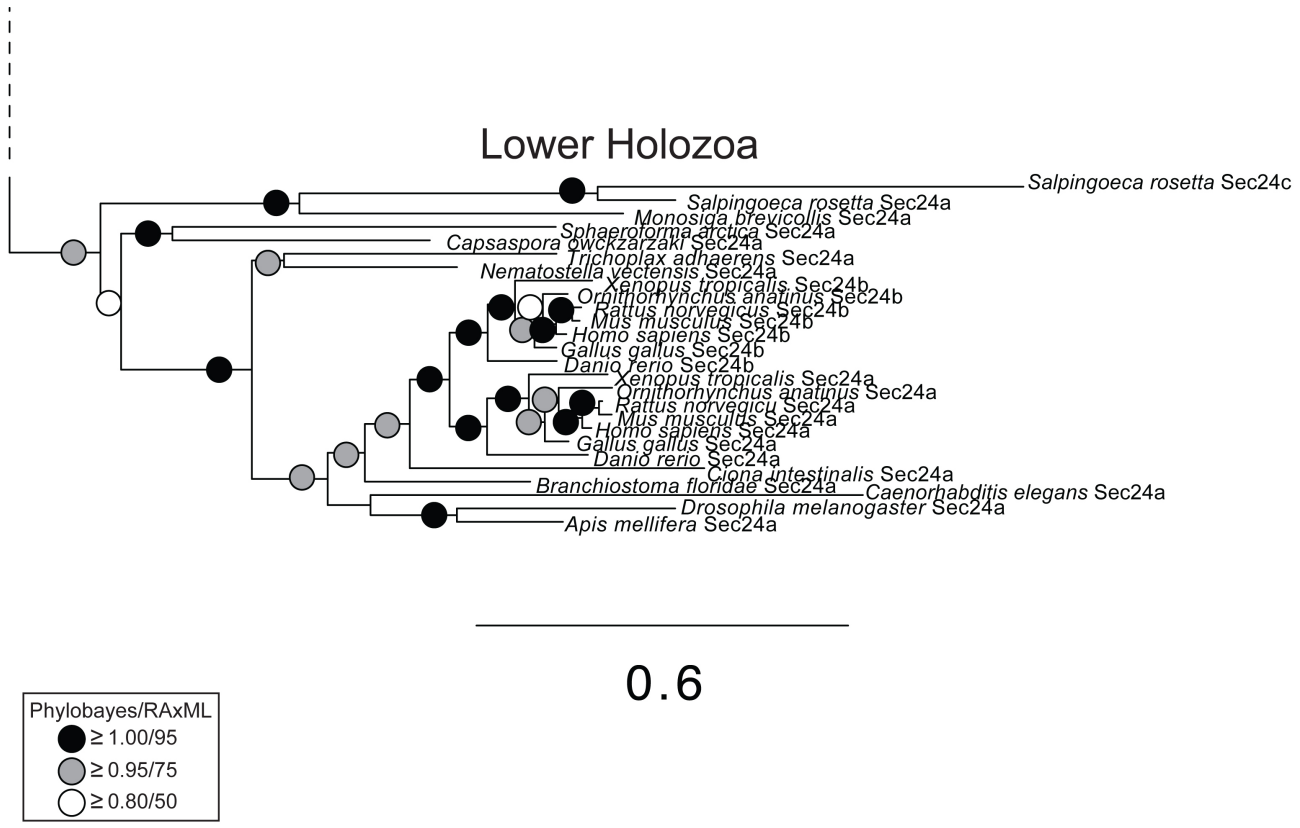
0.6

Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50


Lower Archaeplastida
Ostreococcus tauri Sec24b
Micromonas pusilla Sec24C
Volvox carteri Sec24b
Chlamydomonas reinhardtii Sec24b
Physcomitrella patens Sec24d
Physcomitrella patens Sec24c
Arabidopsis thaliana Sec24d
Arabidopsis thaliana Sec24c
0.6
Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50


Upper Holozoa
Sphaeroforma arctica Sec24b
Capsaspora owckzarzaki Sec24b
Salpingoeca rosetta Sec24b
Monosiga brevicollis Sec24b
Trichoplax adhaerens Sec24b
Caenorhabditis elegans Sec24b
Nematostella vectensis Sec24b
Ciona intestinalis Sec24b
Xenopus tropicalis Sec24c
Monodelphis domestica Sec24a
Rattus norvegicus Sec24c
Mus musculus Sec24c
Homo sapiens Sec24c
Gallus gallus Sec24c
Danio rerio Sec24c
Xenopus tropicalis Sec24d
Ornithorhynchus anatinus Sec24c
Monodelphis domestica Sec24b
Mus musculus Sec24d
Homo sapiens Sec24d
Gallus gallus Sec24d
Danio rerio Sec24e
Branchiostoma floridae Sec24c
Branchiostoma floridae Sec24b
Drosophila melanogaster Sec24b
Apis mellifera Sec24b
Danio rerio Sec24d
Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50
0.6


Lower Fungi
Rozella allomycis Sec24a
Rhizopus oryzae Sec24c
Conidiobolus coronatus Sec24B
Rhizophagus irregularis Sec24a
Schizosaccharomyces pombe Sec24b
Saccharomyces cerevisiae Sec24c
Neurospora crassa Sec24b
Ustilago maydis Sec24b
Cryptococcus neoformans Sec24c
Mortierella verticillata Sec24b
Mortierella verticillata Sec24a
Coemansia reversa Sec24b
Piromyces sp. Sec24a
Spizellomyces punctatus Sec24b
Batrachochytrium dendrobatidis Sec24b
Catenaria anguillulae Sec24c
Catenaria anguillulae Sec24b
Allomyces macrogynus Sec24d
Allomyces macrogynus Sec24c
Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50
0.6


Lower Holozoa
Salpingoeca rosetta Sec24c
Salpingoeca rosetta Sec24a
Monosiga brevicollis Sec24a
Sphaeroforma arctica Sec24a
Capsaspora owczarzaki Sec24a
Trichoplax adhaerens Sec24a
Nematostella vectensis Sec24a
Xenopus tropicalis Sec24b
Ornithorhynchus anatinus Sec24b
Rattus norvegicus Sec24b
Mus musculus Sec24b
Homo sapiens Sec24b
Gallus gallus Sec24b
Danio rerio Sec24b
Xenopus tropicalis Sec24a
Ornithorhynchus anatinus Sec24a
Rattus norvegicu Sec24a
Mus musculus Sec24a
Homo sapiens Sec24a
Gallus gallus Sec24a
Danio rerio Sec24a
Ciona intestinalis Sec24a
Branchiostoma floridae Sec24a
Caenorhabditis elegans Sec24a
Drosophila melanogaster Sec24a
Apis mellifera Sec24a
0.6
Phylobayes/RAxML
≥ 1.00/95
≥ 0.95/75
≥ 0.80/50