Molecular evolution of the vesicular transport machinery and the Golgi apparatus

by

Joel B. Dacks

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
April, 2003

National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisitions et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

0-612-79416-4

Canada

DALHOUSIE UNIVERSITY

FACULTY OF GRADUATE STUDIES

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled "Molecular evolution of the vesicular transport machinery and the Golgi apparatus" by Joel B. Dacks in partial fulfillment for the degree of Doctor of Philosophy.

Dated: April 16, 2003

External Examiner: [redacted]

Research Supervisor: [redacted]

Examining Committee: [redacted]

Departmental Representative: [redacted]

DALHOUSIE UNIVERSITY

DATE: April 29, 2003

AUTHOR: Joel B. Dacks

TITLE: Molecular evolution of the vesicular transport machinery and the Golgi apparatus

DEPARTMENT OR SCHOOL: Biochemistry and Molecular Biology

DEGREE: PhD CONVOCATION: October YEAR: 2003

Permission is herewith granted to Dalhousie University to circulate and to have copied for non-commercial purposes, at its discretion, the above title upon the request of individuals or institutions.

[redacted]
Signature of Author

From New Year's Eve to Montreal in August, Highway 2 at dusk to my defending day…thank you.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

ABSTRACT

The system of internal membrane-bound compartments involved in protein transport and degradation plays a crucial role in eukaryotic cell biology and yet there has been relatively little investigation of this system's evolution.

Transport between these compartments is accomplished through a complicated set of machinery for crafting, delivering and fusing vesicles. Comparative genomic methods were used to examine the origin and early evolution of this vesicular-transport machinery. Molecular biological and phylogenetic methods were also used to investigate more detailed evolutionary questions of the syntaxins, a component of this machinery. Together, these studies uncovered aspects of the endomembrane system's prokaryotic origins, the early crystalization of the core machinery and the extensive diversification of the syntaxin family though out the course of eukaryotic evolution.

Involved in both secretion and endocytosis, the Golgi apparatus plays a deeply entrenched role in the life of most eukaryotic cells. There are, however, a few eukaryotes that are thought to lack this organelle. The question of primary versus secondary absence of the Golgi apparatus in these lineages impacts our understanding of its evolution in eukaryotes as a whole. Molecular biological and phylogenetic methods were used to establish the sisterhood of one 'Golgi-lacking' lineage (the oxymonads) with Golgi-possessing taxa. Golgi-specific components of the vesicular-transport machinery were also obtained from a variety of 'Golgi-lacking' eukaryotes. In sum, these data suggest that there are no extant eukaryotes that primitively lack the organelle, and that the Golgi apparatus was present in the Last Common Eukaryotic Ancestor.

It appears that the basic machinery for intracellular trafficking, as well as the complete organellar complement of the endomembrane system, was already established before the diversification of the known eukaryotic lineages. This finding underscores the importance of the endomembrane system's place in our cellular makeup and its possible role in eukaryogenesis.

LIST OF ABBREVIATIONS

Adaptin=AP

Approximately Unbiased=AU

Degrees of freedom=df

Elongation Factor=EF

Endoplasmic Reticulum=ER

Expressed Sequence Tag=EST

General Time Reversible=GTR

Genome Sequence Survey=GSS

GTPase Activating Protein=GAP

Guanine Exchange Factor=GEF

***In situ* Hybridization=ISH**

Invariable=I

Kishino-Hasegawa=KH

Last Common Eukaryotic Ancestor=LCEA

Lateral Gene Transfer=LGT

Long Branch Attraction=LBA

Maximum Likelihood=ML

Nucleotides=nts

Open Reading Frame=ORF

Plasma Membrane=PM

RNA polymerase II, largest subunit=RPB1

Seconds=s

Small subunit ribosomal RNA =ssu rDNA

***trans*-Golgi Network=TGN**

Vesicular-tubular compartment=VTC

ACKNOWLEDGEMENTS

The following genomics initiatives contributed clones or publicly released data which was used in this thesis: the *Dictyostelium* cDNA project, the *Giardia* genome project, the *Phytophthora* Genome Consortium, the *Chlamydomonas*, and *Porphyra* genome projects, the TIGR *Entamoeba histolytica* Genome Project and especially the Protist consortium (Andrew Roger, Martin Embley, Mark Ragan, John Logsdon and Robert Hirt). The following people provided me with purified DNA: Stephen Hadjuk (*Trypanosoma*); Gary Sisson and Paul Hoffman (*Entamoeba/Giardia*); Lesley Davis and Andrew Roger (*Mastigamoeba/Giardia*); Amanda Lohan and Mike Gray (*Hartmanella*); Alastair Simpson and John Archibald (*Reclinomonas*); and Rosie Redfield (*Naegleria/Reticulitermes hesperus* whole gut fauna). The sequencing in this thesis was performed at the NRC, or in-house by Lesley Davis. My thanks go out to everyone who provided me with materials or technical help.

The core of graduate students in this department have provided a tight and supportive environment in which to work and play. Mike Charette, especially, has been there for me. I have lived with him for longer than anyone to whom I wasn't related.

I have had the pleasure of working with a number of excellent collaborators, some briefly (Renate Radek, Mark Field, Alexandra Marinets), and some as a part of long time associations (Tom Cavalier-Smith, Rosie Redfield, Shigeharu Moriya). Whether our collective product has been included as part of this thesis or not, I have valued their input and interaction. Pak Poon has been a terrific source of advice over the course of my thesis. I have come to lean heavily on his insight on my forays into evolutionary cell biology.

The IG9, especially Niko Yiannakoulias, Sasha Viminitz and Amin Bardestani, have been the sounding board and the compass that have kept me for over 15 years. I take comfort in our collective existence.

The Doolittle lab has been an unbelievable place in which to work. When combined with the Roger lab, well, the term "Science Disneyland" has been known to pop up in my efforts to describe this workplace. A lot of people have come and gone over the last four and a bit years and all of them have had a positive influence on my experiments and my experience here. Sandy Baldauf, John Archibald, Kamran Shalchian-Tabrizi, Maureen O'Malley, Wanda Danilchuk, Camilla Nesbo, Ellen Boudreau, Christophe Douady, Christian Blouin, Thane Papke and David Walsh are only some of the people who deserve mention.

A few key standouts-

Jeffrey Silberman shared my enthusiasm for the oxymonad question early on and essentially taught me how to do nucleotide phylogeny.

That this is the second thesis in which I have acknowledged Andrew Roger's help speaks volumes about his influence on my career. Thanks, Andrew.

Alastair Simpson initiated me seriously into the world (society) of protozoologists. His efforts in our extensive collaborations showed collegial respect; his voluntarily staying behind when I was detained at the Czech border showed friendship.

John Logsdon took me under his wing with overwhelming generosity when I first arrived in the Doolittle lab. His guidance in my initial time here very much set the stage for how I continue to think about eukaryotic evolution and how I try to treat new arrivals.

Yuji Inagaki is consistently the person whom I trust and go to for technical help. My collaborations with him have been some of the smoothest and most enjoyable of any that I have had. His willingness to critically read this thesis, all 200 plus pages of it, says a lot about what kind of labmate and friend he is.

Yan Boucher a été un collègue, un co-loc et un trés bon ami dès son arrivé à Halifax. Je me considère chanceux d' avoir partager un lab et un apartement avec lui pour les dernier 4 années.

Ford Doolittle gave me the freedom to study the questions that I thought were worthwhile, in an environment where those questions could be asked. Most importantly he shared his valuable insight into those questions. Thank you for all of your advice, support and supervision.

I want to thank my bubby Toby, my brother Drew and my parents Barbara and Gurston. Whether for the weekly updates, or the all too infrequent visits, your involvement in my life has not diminished because of our separation. I miss Ethel very much.

Lesley Davis. Without your technical skills, much of the data in this thesis would not have been collected. Without your love and support, none of this work means anything.

Chapter 1: Introduction

"...this basic divergence in cellular structure, which separates the bacteria and blue-green algae from all other cellular organisms, probably represents the greatest single evolutionary discontinuity to be found in the present day world." Stanier, Douderoff and Adelberg, 1963 (Stanier, Douderoff et al. 1963)

Although the archaea were unknown at the time, Stanier's quote still rings true 40 years later. One of the most profound divisions in the biological world is between prokaryotic and eukaryotic cells. While organisms in the prokaryotic grade are biochemically and metabolically diverse, eukaryotes have instead expanded their structural diversity and evolved complex cell biological systems. Features such as a membrane-bound nucleus, cytoskeleton, mitochondria, plastids, and a system of functionally connected membrane-bound compartments (collectively referred to as the endomembrane system), are only some of the things that set eukaryotes apart from prokaryotes. Introductory biology textbooks sport large tables detailing the differences in organization between the two cell types (Alberts 2002).

At the same time, the gulf may not be as wide as originally thought. On the one side, sophisticated homology searching programs and structural examinations have identified prokaryotic homologues for proteins once thought to be exclusively eukaryotic (Kasinsky, Lewis et al. 2001; van den Ent, Amos et al. 2001). On the other, eukaryotes are themselves not as uniform in their cellular organization as previously imagined. Organelles such as mitochondria and peroxisomes have been lost or transformed many times in the course of eukaryotic history (Roger 1999). Plastid evolution is an even more sordid tale of

theft, kidnapping and metamorphosis (Delwiche 1999). To make salient generalizations about eukaryotic evolution, a broad comparative approach will be key.

One of the features that most distinctively separates eukaryotes from prokaryotes is the system of internal membrane-bound compartments involved in protein trafficking. This organellar network sorts, modifies, transports and even captures material: it is one of the defining features of eukaryotes. Evolving this endomembrane system would have been a crucial step in the transition from prokaryote to eukaryote. In addition to the mere selective advantage due to efficiency over simple diffusion, effective directed transport allows for cellular size increase, opening up novel ecological niches. Phagocytosis, for instance, allows efficient heterotrophy and its origin may have been a prerequisite for acquisition of mitochondria and plastids. The evolution of the endomembrane system has thus been proposed by some to be the key step in the evolution of eukaryotes (Stanier 1970).

In this introductory chapter, I outline some of the background useful for investigating the evolution of the eukaryotic endomembrane system. The phylogeny of eukaryotes plays heavily in this story and so an overview of historical and current views of eukaryotic relationships, as well as proposed rooting hypotheses of the eukaryotic tree, will be provided. A reasonably detailed overview of the endomembrane system and, in particular, the vesicular-transport machinery is given to familiarize readers with some of the key processes, transport steps and protein components that are referred to extensively over the course of this thesis. I also critique selected papers that have previously addressed the question of endomembrane system origin and

evolution. Finally, I lay out the specific questions formulated and addressed in this thesis.

Historical overview of eukaryotic phylogeny

From very early on it has been assumed that the more complex eukaryotic cells evolved from cytologically more simple prokaryotic ones (Haeckel 1866). This assumption made possible an idea, implicit since the early 1970s (Margulis 1970), and made explicit in 1983 by Tom Cavalier-Smith under the title of the "Archezoa Hypothesis" (Cavalier-Smith 1983). He stated that, if the prokaryote to eukaryote transition was a gradual one, then some early evolving-eukaryotic lineages may have left descendants which today would lack certain key eukaryotic features such as mitochondria, Golgi or introns because they diverged from the main eukaryotic lineage prior to these inventions. There are, in fact, a number of single-celled eukaryotes that appear to lack some or all of these features, which Cavalier-Smith considered descendants of early-evolving eukaryotes. These included the metamonads (diplomonads, retortamonads and oxymonads), the parabasalids, the archamoebae (pelobionts and entamoebids), and microsporidia (Cavalier-Smith 1987).

In addition to being logically cohesive, the Archezoa Hypothesis was bolstered by the initial molecular phylogenies of eukaryotic relationships based on small-subunit ribosomal RNA (ssu rDNA) (Figure 1.1). These showed at least three of the Archezoan lineages emerging at the base of eukaryotes, adjacent to the prokaryotic outgroups (Sogin 1991). This finding implied that they had emerged prior to the invention, at least, of mitochondria and possibly of other

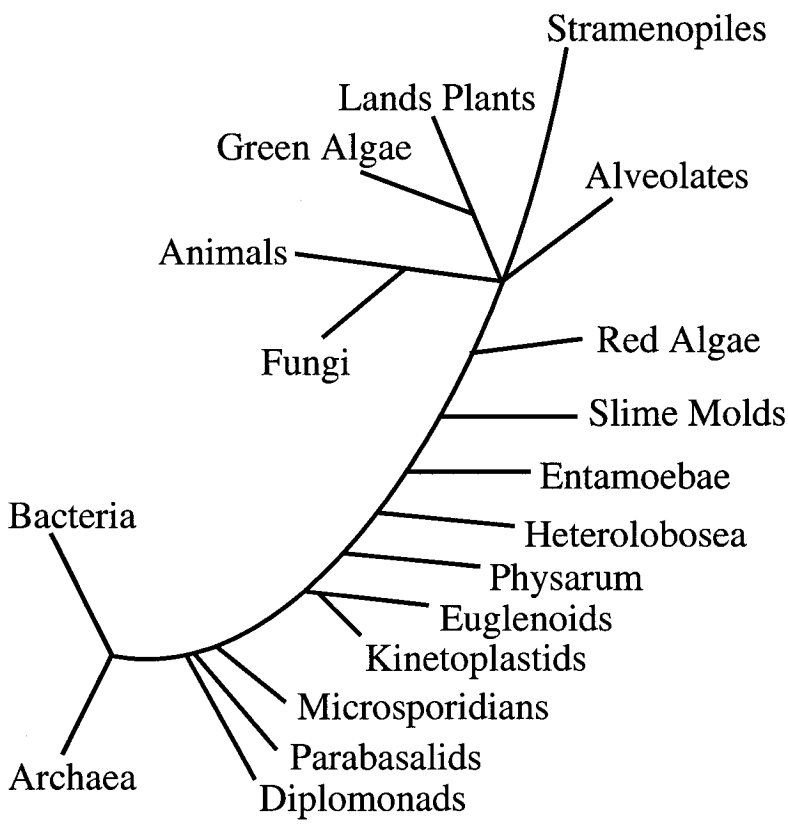


Figure 1.1: **Ssu rDNA model of eukaryotic phylogeny.** Redrawn from Sogin 1991.

eukaryotic novelties. The Archezoa Hypothesis was established as the guiding principle of eukaryotic molecular evolution and reigned virtually unchallenged well into the mid-1990s. The set of relationships based on ssu rDNA, which set forward a robust backbone ladder of sequentially diverging protozoan lineages culminating in an unresolved cluster of plants, animals, fungi and selected protists, similarly held sway as the undisputed arrangement of eukaryotic groups.

Our understanding of eukaryotic relationships is now more textured. Increased taxon sampling has enriched phylogenetic trees that encompass eukaryotic diversity, providing novel fodder for evolutionary scenarios. On the other hand, few people still believe that we have a fully resolved set of eukaryotic relationships.

Phylogenetic analysis of protein sequences developed to be reliable markers of eukaryotic organismal evolution failed to reproduce the small subunit ssu rDNA tree (Embley and Hirt 1998). Although similar disharmony among prokaryotic phylogenies based on different genes is frequently due to lateral gene transfer (LGT) (Doolittle 1999), and eukaryotes are surely not immune to transfer (Andersson, Sjögren et al. 2003), the incongruence in eukaryotes has been largely attributed to failures of phylogenetic reconstruction methods. Early models of phylogenetic analysis made some computationally necessary assumptions that ignored a few important variables. These included the assumption that all aligned positions in a sequence are able to change or that all changes between nucleotides or amino acids occur with the same frequency (Swofford, Olsen et al. 1996), when clearly these are not the case. Most importantly, however, were the faulty assumptions that all sites in a sequence

evolved at the same rate, and that the same gene in different organisms evolved at a constant rate. These assumptions contributed to an artifact in phylogenetic reconstruction called long-branch attraction (LBA) (Felsenstein 1978). LBA causes sequences that evolve at higher rates to be artificially attracted to each other and (for the same reasons) to sequences that are very different because they diverged very long ago. For eukaryotic trees rooted with bacterial or archaeal outgroup sequences (which present long branches because they are indeed anciently diverged) the result will be the artifactual placement of rapidly evolving sequences at the root of the tree.

The recognition of this artifact, and attempts to compensate for it with more sophisticated computer algorithms and biologically accurate models of sequence evolution, yielded some congruence between the ssu rDNA and protein trees but also forced a reassessment of some dearly held "facts" regarding eukaryotic relationships. Principally this has thrown into doubt the ancient nature of many of the organisms held as deeply diverging. Microsporidia were the hardest hit. Phylogenies of tubulins and RNA polymerase (RPB1), as well as re-analyses of other markers, have shown that the microsporidia are not ancient eukaryotes but are either degenerate fungi (Keeling and Doolittle 1996; Hirt, Logsdon et al. 1999; Van de Peer, Ben Ali et al. 2000) or a sister to that clade (Keeling, Luker et al. 2000). The ancient nature of the parabasalids and diplomonads has also come into question: these organisms clearly present long branches in many phylogenetic reconstructions (Hirt, Logsdon et al. 1999; Stiller and Hall 1999). However, no alternative placement has been suggested for these taxa and so their status as deeply diverging lineages is merely in doubt, not disproven.

The robustly resolved backbone of the eukaryotic tree has also been called into question. Philippe et al demonstrated that the long-branch attraction artifact can provide false support for the ladder-like structure of sequentially emerging taxa seen in the eukaryotic phylogenetic tree in Fig. 1.1. Correction for this artifact produces a tree whose deep branching order is unresolved (Philippe, Lopez et al. 2000). They propose this unresolved backbone structure as the result of a real (biological) phenomenon, rather than methodological failure. Their "Big Bang" hypothesis suggests that many of the extant eukaryotic lineages evolved very rapidly from one another: the branching order is unresolved because, for most markers, few mutations occurred between branchings (Philippe, Germot et al. 2000). A rapid radiation does not mean that a phylogeny of eukaryotes is an unattainable goal. It does, however, require that inferences of deep eukaryotic relationships be based on multiple and varied lines of evidence.

Current views of Eukaryotic Phylogeny

Figure 1.2 schematically depicts a synthesis of the latest proposed relationships among major eukaryotic taxa leaving out, for the time being, relationships determined exclusively in this thesis. Many of the higher taxonomic groups, or supergroups, that will be discussed are proposed based on both ssu rDNA and protein gene phylogenies and are also congruent with morphological evidence.

Starting clockwise from the top of Figure 1.2, the Cercozoa unites cercomonads with euglyphid testate amoebae, plasmodiophorids and chlorarachnean algae based on ssu rDNA evidence (Cavalier-Smith 2000). Actin phylogenies (Keeling 2001) confirm this relationship and, in turn, place the

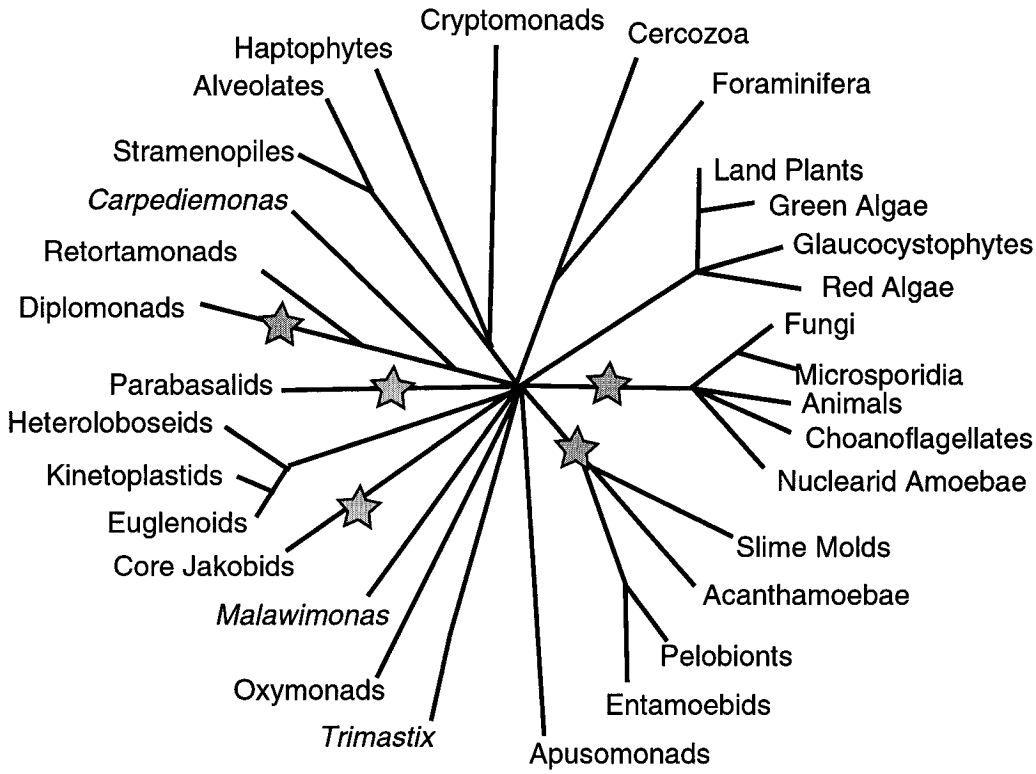


Figure 1.2: **Star phylogeny of current proposed eukaryotic relationships.** This unrooted star phylogeny incorporates morphological, ssu rDNA and protein data (Baldauf and Palmer 1993; Cavalier-Smith and Chao 1996; Baldauf, Roger et al. 2000; Moreira, Le Guyader et al. 2000; Fast, Kissinger et al. 2001; Simpson and Patterson 2001; Arisue, Hashimoto et al. 2002; Bapteste, Brinkmann et al. 2002; Silberman, Simpson et al. 2002; Simpson, Roger et al. 2002). The yellow stars illustrate proposed placements of the eukaryotic root.

Cercozoa as sisters to the Foraminifera, which represent one of the most extensively fossilized eukaryotic groups (Archibald, Longet et al. 2003). The larger 'Cercozoa plus Foraminifera' clade is also reinforced by a unique insertion in the tandemly repeated ubiquitin gene (Archibald, Longet et al. 2003).

The debate as to the number of endosymbiotic events giving rise to primary plastids has generally hinged on whether or not Glaucocystophytes, red and green algae form a monophyletic group (Delwiche 1999). While plastid gene phylogenies seem to support such a clade, RPB1 phylogenies appeared to refute it (Stiller and Hall 1997). Very strong phylogenetic evidence from EF2 genes and combined nuclear gene sequence (Moreira, Le Guyader et al. 2000), as well as a reanalysis of the RPB1 dataset (Moreira, Le Guyader et al. 2000; Dacks, Marinets et al. 2002), however, support the clade of primary plastid-containing organisms.

Some of the most surprising and humbling relationships in the eukaryotic tree involve a rather familiar eukaryotic lineage, ourselves. In 1993, it was reported that animals and fungi share an insertion in their EF1alpha gene, uniting them in a group later dubbed the Opisthokonts (Baldauf and Palmer 1993). This grouping of animals and fungi has been supported by many subsequent protein phylogenies and concatenated gene analyses (Baldauf, Roger et al. 2000), as well as being congruent with morphological features including flagellar arrangement (Cavalier-Smith 1987). Other protozoan lineages have also been attached to this assemblage, or located within it, including: the Choanoflagellata; Ichthyosporea; and the nucleariid amoebae (Zettler, Nerad et al. 2001; Lang, O'Kelly et al. 2002).

The pelobionts and entamoebid are amitochondriate, 'Golgi-lacking', amoebae which were suggested to be primitive based on their cytological

simplicity (Cavalier-Smith 1987). The first ssu rDNA analyses of these organisms seemed to dispute that assessment (Hinkle, Leipe et al. 1994) but, with the current uncertainty of the eukaryotic root and the acknowledgement of artifact in the ssu rDNA tree, their status has come back into question (Stiller and Hall 1999). Another group of amoebae with claims to the primitive label are the slime molds. These organisms were thought to be deep based on ssu rDNA (Hinkle, Leipe et al. 1994) and their very monophyly was also brought into question. However, they have also been suggested as the sister group to the opisthokonts on several occasions (Baldauf and Doolittle 1997; Baldauf, Roger et al. 2000). Several recent analyses of protein genes have robustly placed the pelobionts, entamoebids and slime molds together as a robust monophyletic clade called "Conosa" (Arisue, Hashimoto et al. 2002; Bapteste, Brinkmann et al. 2002).

Many of the organisms from the former archezoa fall into a loose assemblage of primarily flagellated protists termed "the excavates" (Simpson and Patterson 1999; Simpson and Patterson 2001). This group is characterized by the possession of a complex cytoskeletal apparatus underlying a ventral feeding groove. The presence and specific arrangement of the microtubular roots, and several non-microtubular fibers (H, I and B fibers), are thought to be so complex a set of morphological traits that they are highly unlikely to have evolved multiple times. Excluding work done in this thesis, lineages thought to be in the excavate taxa include the Heterolobosea, diplomonads, retortamonads, and some lesser known flagellates including *Trimastix*, *Carpediemonas*, *Malawimonas*, and the "core jakobids", *Reclinomonas* and *Jakoba* (Simpson and Patterson 1999; Simpson and Patterson 2001). Molecular analyses have united several subsets of these taxa or established their relationship to other eukaryotic groups. The

combined data approach used by Baldauf et al. demonstrated the unification of the Heterolobosea and the Euglenozoa into the larger assemblage, the "discicristata"(Baldauf, Roger et al. 2000). Molecular data from ssu rDNA bring together retortamonads, diplomonads and *Carpediemonas* (Silberman, Simpson et al. 2002; Simpson, Roger et al. 2002), with tubulin data also uniting *Carpediemonas* and diplomonads(Simpson, Roger et al. 2002). This clade is also proposed to be related to the parabasalids, from the shared presence of a costal fiber (Simpson and Patterson 2001), and an affiliation of diplomonads with parabasalids seen on many molecular trees (Dacks and Roger 1999). However, all molecular data thus far has failed to unite the excavate taxa into a single monophyletic clade. Consequently, while the subsets of these taxa described above are depicted in Figure 1.2, the excavate taxa are shown as adjacent in the diagram rather than as a clade.

Each a major eukaryotic group in their own right, the ciliates, apicomplexa and dinoflagellates have been grouped into an "alveolate" clade based on morphological and molecular evidence (Gajadhar, Marquardt et al. 1991; Patterson 1999). In 1999, Tom Cavalier-Smith predicted that these groups were related to a protozoan clade called the heterokonts, or stramenopiles in a higher taxon that he called the "chromalveolates" (Cavalier-Smith 1999). This prediction was based on ultrastructural and protein targeting evidence and is strongly supported by the endosymbiotic gene replacement of the nuclear glyceraldehyde-3-phosphate dehydrogenase gene by a plastid version in representatives of these lineages (Fast, Kissinger et al. 2001). This feature not only demonstrates the larger taxonomic affiliation of these groups but also their photosynthetic origins, a revolutionary and controversial idea, particularly for

the ciliates. The cryptophytes and haptophytes are also proposed to belong to this photosynthetic line (Cavalier-Smith 1999; Yoon, Hackett et al. 2002).

Rooting the Tree of Eukaryotes

A rooted phylogeny makes inferences about ancestral character states easier to come by. For features that involve all eukaryotes, such as the endomembrane system, this means finding the root of the eukaryotic tree. Under the ssu rDNA model of eukaryotic evolution it was thought this was a relatively resolved question. However, the acknowledgement that LBA-related artifact could be responsible for this result has opened the door to alternative placements of the root. Figure 1.2 is shown as unresolved and unrooted to convey this uncertainty, but stars have been placed to illustrate some of the more prominent theories about where that root might lie.

Strong arguments have been made for placing the eukaryotic root near the diplomonads, the parabasalids, or both. This has been based on their apparent lack of key eukaryotic features, as well as their consistent placement near the base of outgroup rooted phylogenies (Sogin 1991; Sogin 1997). While both groups surely represent long branches in these phylogenies (Embley and Hirt 1998), the lack of an alternative placement for them among a clearly derived lineage (as exists for microsporidia) allows both the diplomonads and parabasalids to remain as viable candidates for deep-branching eukaryotes.

An additional piece of evidence supports the deep branching status of parabasalids. A single amino acid indel in the enolase gene is shared between parabasalids and prokaryotes, to the exclusion of all other eukaryotic lineages examined (Keeling and Palmer 2000). Partial gene conversion combined with

LGT, as well as simple mis-alignment, have been proposed to explain this character (Bapteste, Brinkmann et al. 2002). However, it remains possible that the indel is a shared derived state for eukaryotes, with parabasalids diverging prior to this evolutionary event.

The largest sets of genes contained in mitochondrial genomes are found in *Reclinomonas americana, Malawimonas jakobiformis,* and *Jakoba libera,* and these most closely resemble the hypothesized genome of a bacterial mitochondrial ancestor (Lang, Burger et al. 1997). With the demonstration that most eukaryotic lineages did, at one point, possess mitochondria (Roger 1999, Silberman and Roger 2002), the presence of pleisiomorphic mitochondrial genomes in the jakobids and *Malawimonas* suggests that these organisms may be near the root of the eukaryotic tree. They would have then diverged away from the rest of eukaryotes prior to several endosymbiotic gene transfers and the replacement of RNA polymerase genes in the mitochondrial genomes of most eukaryotes.

The phylogenetic distribution of a derived gene fusion between dihydrofolate reductase and thymidylate synthase (Stechmann and Cavalier-Smith 2002) forms the rationale for the most recent suggested placement of the eukaryotic root. This proposed rooting is between opisthokonts, in which the genes are separate as in prokaryotes, and most or all other well-characterized lineages, many of which were shown to have the genes fused. The taxon sampling of this study is limited, with few Conosa and especially few excavate taxa, where other rootings have been proposed. It is also vulnerable to possibility of convergent occurrences of the gene fusion or to inter-eukaryote lateral gene transfer events.

The rooting of the eukaryotic tree remains an unanswered question. As much as possible in this thesis, deductions are made that are independent of the exact placement of the eukaryotic root. In cases where this is not possible, I take into account all proposed placements and address the various possibilities.

Endomembrane system review

Before approaching the question of its origin and evolution, a cell-biological tour of the endomembrane system is warranted. Unlike the nucleus or mitochondria, this system is spread across multiple organelles that span the cell. It also involves a large number of molecular components, each specialized and often cryptically named. A better understanding of the organelles and protein components of this system should help to navigate the proceeding questions of its origins, evolution and complexification.

Organelles in the endomembrane system

The endomembrane system is a series of membrane-bound compartments, connected by vesicular transport, which functions as an assembly line for both protein transport and ingestion of extracellular material. Although traffic between organelles is often bi-directional, the description given here follows the path taken by a protein, first being synthesized, then modified and targeted for secretion. The organelles and progression of material in the digestive or endocytic pathway are then described. Movement forward along these pathways is termed anterograde transport, while movement to the previous organelle in the process is called retrograde.

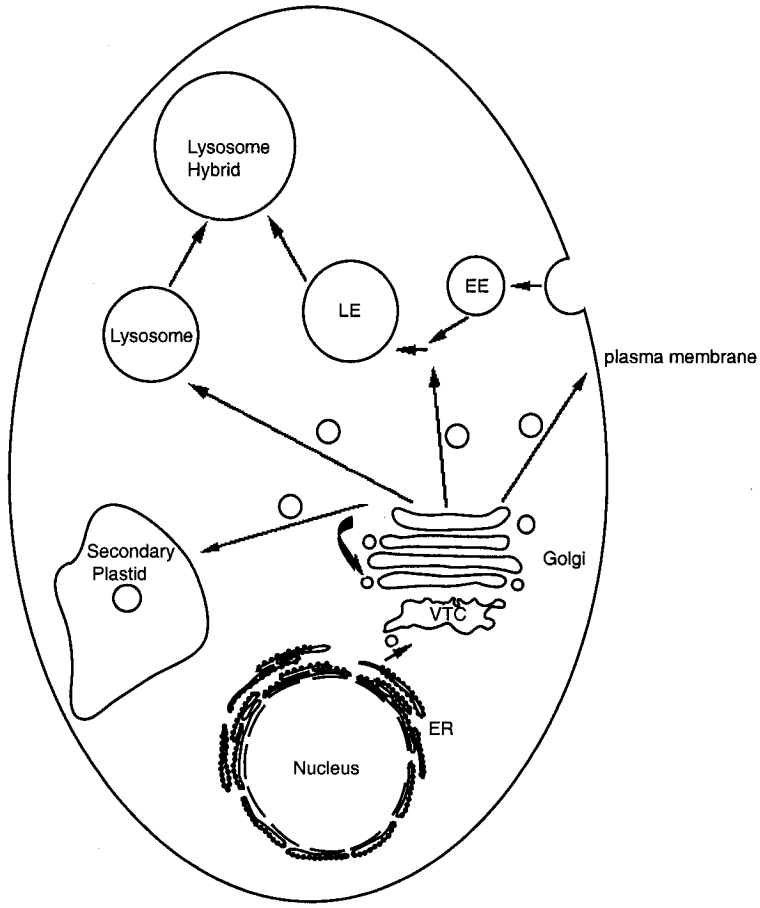


Figure 1.3: **Organelles and direction of vesicular transport in a hypothetical eukaryotic cell.** Straight arrows in this cartoon show the anterograde movement of vesicles between membrane compartments. The curved arrow illustrates retrograde transport in this case alone, as retrograde transport is the only non-controversial function of COPI vesicles. Small circles represent transport vesicles, while larger ones represent digestive organelles. EE and LE denote the early and late endosomes, respectively. ER = endoplasmic reticulum, VTC = vesicular-tubular compartment. Much of the information shown here was derived from studies of mammalian and yeast cells, but the movement of Golgi-derived vesicles to a secondary plastid (as in *Euglena*) is also depicted.

The endoplasmic reticulum (ER) is contiguous with the nuclear envelope (Fig 1.3). Rough ER has a studded appearance due to bound ribosomes, and is the site of synthesis for proteins destined to travel *via* vesicular transport. Transport vesicles bud from ribosome-free regions of the rough ER called transitional elements (Klumperman 2000) and quickly fuse, either with other vesicles derived from the same source, or with a network of tubules termed the vesicular-tubular compartment (VTC). The VTC moves material along to the *cis*-Golgi and is sometimes contiguous with it (being referred to as the *cis*-Golgi network).

The Golgi apparatus (also called Golgi body, Golgi complex or dictyosome) is the next distinct organelle in the endomembrane system. Most familiar as parallel stacks of flattened membrane-bound compartments, Golgi body morphology is actually quite varied among eukaryotes-with flattened stacks in animals, plants and many protozoa, punctate vesicles in most fungi (but not chytrids) and smaller, but numerous, stacks in some alveolates (Becker and Melkonian 1996). Given this diversity, a definition that relies upon classical stacked morphology is not appropriate. Rather the definition should be functional and encompassing all morphologies in eukaryotes. The Golgi apparatus may be described as a series of membrane compartments which receives material from the ER and in which proteins are modified and sorted for later transport to various organelles. Compartments that receive material from the ER are called *cis*-Golgi. Subsequent compartments to the *cis*-Golgi are called medial Golgi. Finally, Golgi compartments which produce vesicles bound for further transport are called the *trans*-Golgi network (TGN).

Progression of material becomes less linear upon exiting the Golgi apparatus (Fig 1.3). In mammals and yeast, vesicles emerge from the TGN and may travel in four possible directions (Bryant and Stevens 1998). They can follow a retrograde path backward to previous compartments, recycling material back to earlier stages in the Golgi apparatus or to the ER. They may journey in an anterograde direction either to the plasma membrane (PM) or to intersect with the endocytic pathway.

The plasma membrane represents the end point of secretion and the beginning of the endocytic pathway. Vesicles leave the TGN and travel to the PM where they fuse, either releasing their soluble contents, or presenting their membrane bound cargo at the surface. At the plasma membrane, endocytic vesicles are created to entrap food or internalize ligand-bound cell surface receptors (Fig 1.3).

Vesicles derived from the TGN fuse, either with endocytic vesicles derived from the plasma membrane called early endosomes, or with a pre-existing late endosome. The late endosome then fuses with lysosomes to create a larger hydrolytic organelle involved in protein degradation. This late endosome+lysosome hybrid has been considered by some to be a separate organelle. The lysosome then is defined as the organelle containing concentrated hydrolytic enzymes and is reconstituted after protein degradation occurs in the hybrid (Fig 1.3).

While this description holds for animals and fungi, additional complexities certainly exist in the organization of the endomembrane system in other eukaryotes. The secondary plastid of euglena, for example receives proteins through vesicular transport (Sulli and Schwartzbach 1995; Sulli, Fang et

al. 1999). The organization of the endomembrane system in *Giardia intestinalis* is a matter of significant controversy. It was been proposed that protein sorting occurs at the ER, and that some proteins may be transported directly to the plasma membrane, while others are processed through transient Golgi-like structures (Encystation Specific Vesicles) (Marti, Li et al. 2003). Other workers have reported stack-like structures in *Giardia*, but only at encystations and excystation (Lujan, Marotta et al. 1995).

Vesicular transport and the machinery that runs it

Cell-biological studies in model systems have shown that, regardless of the donating and receiving organelles, the mechanistic process of vesicular transport has some shared features (Springer, Spang et al. 1999). These, therefore, can be described in a generalized model with three basic steps: vesicle formation/budding from the donor organelle, vesicle movement and fusion of the vesicle with the target organelle. The machinery used for vesicular transport between the different organelles is a mixture of components common to a process (regardless of location), members of protein families with paralogues specific for transport between two given organelles and uniquely organelle-specific complexes.

Vesicle budding

The process begins by recruitment of a small GTPase to the cytosolic side of the membrane at the site of vesicle formation. This GTPase is initially GDP-bound, but Guanine-Exchange Factor proteins (GEFs) replace GDP with GTP. The GTPase aids in regulating vesicle formation, and assembling the cytosolic coat proteins required for vesicle budding. Cargo proteins destined for

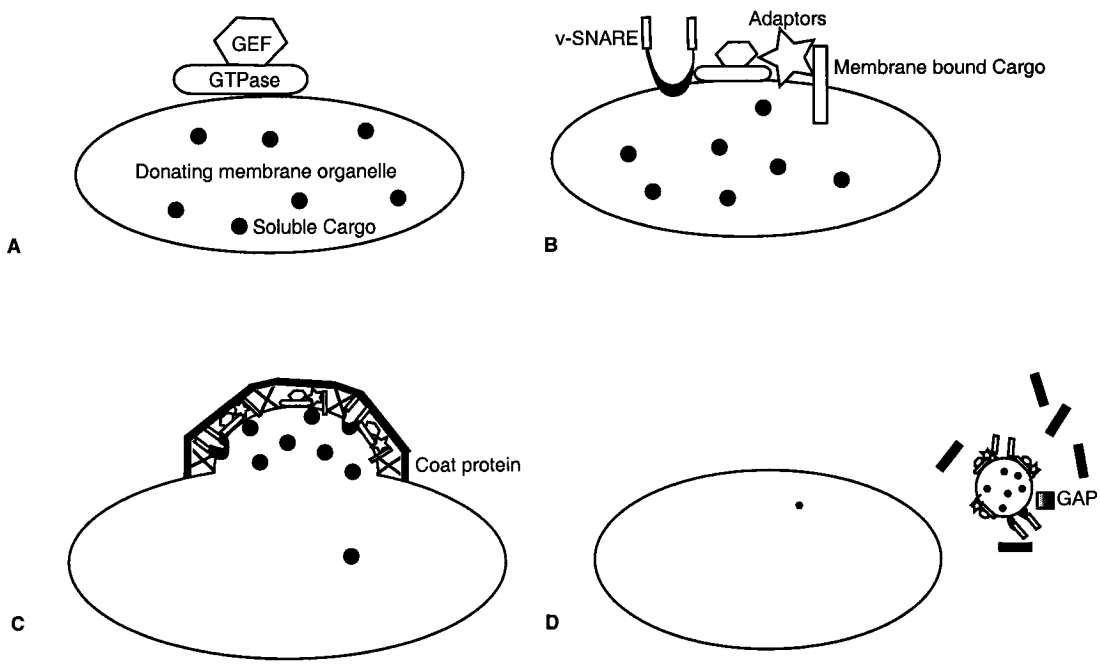


Figure 1.4: **Generalized cartoon of vesicle formation, budding and movement.** (A) A GTPase attaches to the membrane and a GEF swaps GDP for GTP. (B) Adaptor proteins and cargo attach to the nucleating site of vesicle formation. (C) Coat proteins arrive and form a scaffolding complex for vesicle formation. Soluble cargo may be incorporated into the vesicle *via* adaptors or by bulk flow. (D) The vesicle has budded away from the donating membrane, a GTPase-activating protein (GAP) hydrolyzed GTP and the vesicle uncoats. All shapes, once named in a panel, retain their assignment in subsequent panels.

transport by the vesicle may be packaged as a result of bulk flow, direct interaction with coat proteins *via* amino acid motifs in the cargo, or *via* adaptor proteins. After cargo selection, the protein coat polymerizes and the ensuing vesicle buds, beginning the journey to its target destination. This generalized model is illustrated in Figure 1.4 and was strongly influenced by that laid out in Springer et al. 1999 (Springer, Spang et al. 1999). Although the three well-characterized types of vesicles built within the cell conform to the generalized model of vesicle formation, their protein components differ significantly.

Vesicles involved in anterograde ER-to-Golgi body transport are coated with a complex termed COPII (Kaiser and Ferro-Novick 1998; Springer, Spang et al. 1999). In the creation of COPII vesicles, the GTPase Sar1 binds to the cytosolic face of the ER with Sec12 acting as its GEF. The Sec23/24 complex interacts with membrane-bound cargo proteins, either directly (usually by KKXX motifs in the cytosolic tails of the cargo), or indirectly *via* protein cargo adaptors such as p24 or ERGIC-55. Cargo is concentrated into these exit regions and incorporated into transport vesicles, presumably *via* retention though proteins of the Emp24 family (Muniz, Nuoffer et al. 2000) and also through bulk flow (Klumperman 2000). The Sec23/24 complex, along with any attached cargo, will bind to the vesicle formation site followed by the Sec13/31 complex.
Several complexes, of all of the above proteins, polymerize to cause vesicle budding.

COPI vesicles recycle material from the Golgi apparatus back to the ER. In formation of the COPI complex, a member of the Arf protein family binds to the cytosolic portion of the membrane. The Arf attaches to the membrane in GDP-bound form, which is then exchanged for a GTP moiety *via* the action of an

ArfGEF. The vesicle coat itself is composed of a heteroheptameric complex entitled Coatomer. The seven genes encoding these products are dubbed COP genes and listed alpha, beta, beta-prime, gamma, delta, epsilon, and zeta (Waters, Serafini et al. 1991; Chow, Sakharkar et al. 2001). Some membrane-bound cargo may attach, *via* a dihydrophobic amino acid motif, to a vesicle coat forming coatomer complex. Soluble ER-resident proteins displaced to the Golgi body by anterograde transport recycle by attaching to KDEL-receptor proteins. These, in turn, bind a GTPase activating protein (GAP) called ArfGAP. ArfGAP then binds Arf, to get into the complex. Coatomer, Arf, and ArfGAP complex to form the polymeric coat and vesicle budding occurs (Springer, Spang et al. 1999).

While COPI vesicles are clearly involved in retrograde transport from the Golgi apparatus back to the ER, they may also be involved in anterograde transport forward to later Golgi compartments (Orci, Stamnes et al. 1997; Schekman and Mellman 1997). This possibility gets to the heart of whether transport in the Golgi apparatus occurs by vesicular transport between stable organellar structures or by an assembly-line type movement of the structures progressively maturing into new organellar identities. In either model, however, COPI vesicles are important.

Many of the remaining vesicles formed in the cell, including those for transport from TGN to both the endosome and the plasma membrane, are coated with clathrin or clathrin-related proteins. In the formation of clathrin-coated vesicles, an Arf paralogue again acts as the GTPase with an ArfGEF again providing the GTP exchange (Springer, Spang et al. 1999; Kirchhausen 2000). Proteins called adaptins (AP) bind cargo and provide specificity for particular organellar destinations. AP1 and AP3 are involved in transport steps derived

from the TGN, targeting material to the late endosome and lysosome, respectively. AP2 is involved in cargo selection for plasma membrane-derived vesicles entering the endocytic pathway (Kirchhausen 2000; Robinson and Bonifacino 2001). Adaptors bind *via* cis-acting amino acid motifs in the cargo or *via* additional adaptor proteins, such as the mannose-6-phosphate receptors in mammals for the transport of material to the late endosome (Kirchhausen 2000). Clathrin itself acts as the protein coat, polymerizing and forming the vesicle. In the case of AP3, clathrin is not involved but Vps41, which appears to have a homologous clathrin domain, acts as the protein coat (Kirchhausen 2000; Robinson and Bonifacino 2001).

Vesicle movement

After vesicle formation and budding, the vesicle is transported to its eventual target. As with intracellular transport of most cellular material, the cytoskeleton is involved. Both actin- and microtubule-based networks, as well as isoforms of kinesin, dynein and myosin, have been implicated in endocytosis and intracellular trafficking (Ma, Fey et al. 2001). This has been demonstrated by genetic studies in *Dictyostelium discoideum* (Ma, Fey et al. 2001), as well as by the identification of various proteins associated with cytoskeletal function in genetic screens for secretion and vacuolar protein sorting yeast mutants (Bryant and Stevens 1998).

At some point after vesicle formation, the GTP on the GTPase is hydrolyzed back to GDP *via* the action of an ArfGAP homologue. The role of this hydrolysis is unclear, although ArfGAPs have been suggested to interact with the cytoskeleton and signaling proteins (Donaldson and Lippincott-Schwartz 2000).

Figure 1.5: **Generalized cartoon of vesicle fusion.** (A) An incoming vesicle containing cargo and v-SNARE homologue approaches receiving organelle possessing a Snap-25 and a syntaxin homologue complexed with a Sec1 homologue. (B) Sec1 releases syntaxin, which forms a coiled-coil structure with the v-SNARE and Snap25 homologues, prompting vesicle docking and tethering. (C) Vesicle fusion begins with the SNARE complex and other proteins (V-ATPase subunits) being implicated in creating a fusion pore. (D) NSF hydrolyzes ATP to dissociate the SNARE complex and recycle components for future rounds of vesicle fusion. Rab is implicated at various steps in the process. All shapes, once named in a panel, retain their assignment in subsequent panels.

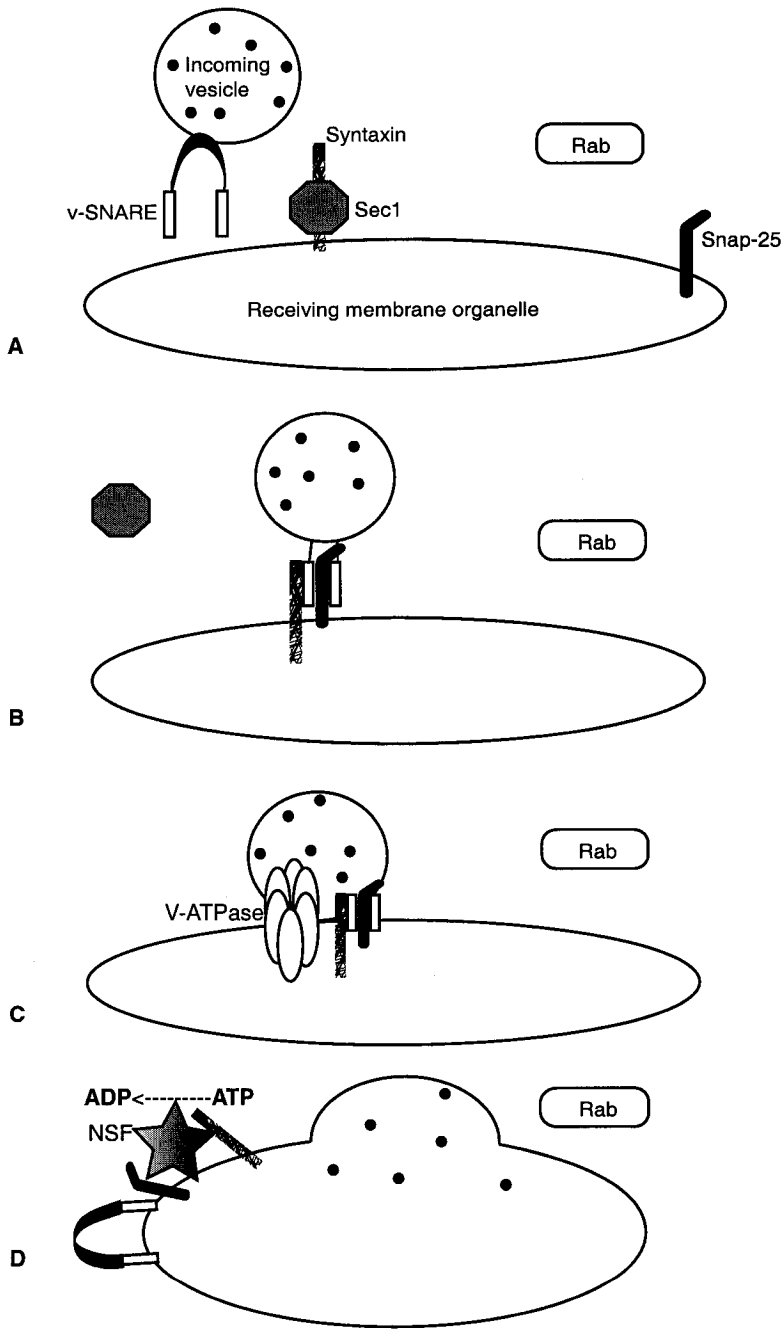


Figure 1.5: **Generalized cartoon of vesicle fusion**

It has also been demonstrated that, when GTP hydrolysis is blocked, intracellular transport vesicles appear unable to uncoat (Tanigawa, Orci et al. 1993). Whether this is a causal relationship or not, uncoating of the transport vesicles does occur after leaving the donor membrane and before vesicle fusion.

Vesicle fusion

The final stage of vesicular transport is the fusion of the cargo-laden vesicle with its intended target, as illustrated in Figure 1.5. A v-SNARE homologue was incorporated as membrane cargo during vesicle formation. On the receiving side, each target membrane possesses at least one member of the t-SNARE protein family, syntaxin (Edwardson 1998). The cytosolic portion of this protein, usually the very N-terminus only (Dulubova, Yamaguchi et al. 2003), becomes complexed by a member of the Sec1 family (other names for members of this family include Munc-18, Vps45, Vps33, syntaxin binding protein, and Sly1). These proteins appear to regulate interaction of the syntaxin with other pieces of the fusion machinery (Dulubova, Yamaguchi et al. 2003). As the incoming vesicle reaches the target membrane (Fig 1.5B), the syntaxin, v-SNARE and a homologue of SNAP-25 form a four-helix coiled-coil bundle along their (approximately 57 residue) SNARE motifs. The SNAP-25 and syntaxin homologues each contribute one copy of the motif and the v-SNARE contributes two (Hay 2001). The exact contribution of the SNARE complex is highly controversial. It has been implicated in a variety of vesicle-fusion stages (Ungermann, Sato et al. 1998; Nickel, Weber et al. 1999; Ungermann, Price et al. 2000). The SNARE hypothesis (Sollner, Whiteheart et al. 1993) proposes that the

interactions between the different SNAREs encode the specificity for vesicular transport (McNew, Parlati et al. 2000). Regardless of its exact role, the SNARE complex is essential for vesicular fusion, which can only begin after the complex is formed. SNAREs have additionally been implicated in the production of the physical fusion pore (Fig 1.5C). The V_0-subunit of the vacuolar H^+-ATPase is also involved in pore creation, at least during vacuolar fusion (Peters, Bayer et al. 2001). The universality of this feature is unclear. After fusion (Fig 1.5D), the four-helix SNARE bundle is disassembled and recycled *via* the action of an ATPase, NSF (Edwardson 1998) which seems to be involved in all vesicular transport fusion events. An NSF paralogue, named p97, acts in cases of post-mitotic reassembly of organelles, also a vesicular fusion event (Rabouille, Kondo et al. 1998).

GTPases of the Rab protein family are involved in the vesicular transport process in a variety of steps and in a variety of ways. They have been implicated in vesicle formation, movement and fusion. They are clearly essential and have a large number of proteins with which they interact, both physically and genetically, including SNAREs. Rabs, it seems, regulate the various steps in vesicular transport, possibly ordering the events in the process (Zerial and McBride 2001).

The above descriptions have been primarily based on the extensive studies performed in a very few model systems, primarily *Saccharomyces cereviseae* and mammalian cells. Whenever possible, comparative data with other systems have been incorporated to obtain a wider and more general picture. Indeed one of the benefits of the work in this thesis is to provide, at least on a gene sequence level, additional information that can be used in exactly such a

way. Although the descriptions above have, by necessity of space and clarity, been simplified they should provide a reference for the complex set of proteins and processes that are dealt with throughout this thesis.

Origin of the endomembrane system

Although the endomembrane system has not received as much attention as mitochondria or plastids, a number of papers have attempted to explain its origin and evolution. Both endosymbiotic and autogenous origins have been proposed.

The endosymbiotic theories are primarily aimed at explaining the origin of the nucleus, with the creation of an endomembrane system occurring as a by-product. The "chimeric fusion" hypothesis (Gupta and Golding 1996) of the eukaryotic nucleus postulates ER as remnants from a mottled cell membrane that fused to create vesicles. The "syntrophy hypothesis" (Moreira and Lopez-Garcia 1998) proposes the origin of eukaryotic cells as a fusion of a methanogenic archaeon (giving rise to the nuclear genetic material) and delta-proteobacteria (giving rise to the cytoplasm and plasma membrane, ER and nuclear membrane).

Arguments against the endosymbiotic origins of the nucleus have been eloquently set out elsewhere (Martin 1999) and will not be dealt with in detail here. Even if the endosymbiotic theories were true with respect to the origin of the nucleus, their explanations of the origin of the endomembrane system still fall short. Both of the above theories treat the endomembrane system as if it were simply a series of static vesicles in the cell, and ignore mechanistic problems such as maintenance of the system, evolution of its components or even maintaining the shape of the vesicular apparatus that they postulate.

The autogenous theories are far more explicit and have more explanatory power. In his 1999 paper, William Martin explains the evolution of the endomembrane system due to a lateral gene transfer of a lipid-synthesizing enzyme to the protoeukaryote (Martin 1999). Out of its cellular context, this enzyme would begin to produce lipids in the cytoplasm, causing an increase in local concentration and spontaneous vesicle formation. These internal vesicles would then provide the seed for an eventual endomembrane system.

This proposal has several potential flaws. The first is that lipid biosynthesis is a complicated process. A single lateral gene transfer is unlikely to cause the synthesis of lipids that would be unable to integrate themselves into the existing membranes, as is suggested. The transfer of the entire pathway would be necessary, rendering the event less likely. Secondly, the spontaneous creation of vesicles is a matter of concentration. Since the cell would not be compartmentalized at this point the lipids would freely diffuse and thus the vesicle, when it gets created, will be at best randomly formed in the cytoplasm. There is little reason to presume that these lipids would fuse into a membrane coat that surrounds the genetic material (i.e. a nuclear envelope and ER). Finally, even should this proposal be true, it merely explains the creation of vesicles in the cytoplasm, and possibly a nucleus. How these vesicles become a full-fledged endomembrane system is not examined.

In general the published theories explaining the evolution of the endomembrane system have been highly speculative and vague. The topic is included as an afterthought in stories of nuclear origins and, except for speculating that the ER was the first compartment to evolve, they fail to explain details of how a system might be evolved or be maintained. One exception is the

treatment by Tom Cavalier-Smith. The most recent version of his hypothesis (Cavalier-Smith 2002) begins with the logical step of invaginations from the plasma membrane. From there, there are several critical points required for the evolution of an endomembrane system: the cell wall must be lost from the archeon-derived proto-eukaryote; the ability to exocytose and phagocytose must evolve concurrently; the ER must differentiate itself as an organelle separate from the plasma membrane; and finally the prokaryotic SecA pathway of membrane protein insertion must be converted to the more eukaryotic SRP pathway in the proto-eukaryote. This switch to SRP insertion, along with the ability to phagocytose and the ability to bud vesicles off the ER, is sufficient in this model to define an endomembrane system as 'present'. Once the endomembrane system was 'present' then the various organelles could differentiate and establish themselves, based on the evolution of the three types of coated vesicles. COPII vesicles are the oldest in his model because they are exclusively ER localized. Clathrin-coated vesicles and finally COPI vesicles would then have evolved (Cavalier-Smith 2002). The origin of a few protein components of the vesicular-transport system is proposed. For example, Rabs are proposed to have evolved from myxobacterial small GTPases. Although this schema for the evolution of the system is highly speculative, it is far and away the most detailed and complete explanation available.

Questions addressed in this thesis

In the broadest sense, this thesis examines the general evolution of the eukaryotic endomembrane system. This question, however, is far too expansive to be fully addressed here and will likely take many years work by many

different investigators. I have chosen to focus on two aspects, each of which encompasses a reasonably broad sub-question in its own right.

The first section of the thesis addresses the origin, evolution and complexification of the machinery involved in vesicular transport. Chapter 2 approaches this question from a bio-informatic perspective, using publicly available genome data to find prokaryotic homologues of vesicular-transport components and to reconstruct the minimal ancestral protein complement of the vesicular-transport machinery likely present in an early common eukaryotic ancestor. Chapter 3 addresses a more detailed question involving the complexification of one of the gene families important in vesicular fusion, the syntaxins.

The second section of the thesis addresses the evolution of the Golgi apparatus in light of the Archezoa Hypothesis. Several putatively ancient eukaryotic lineages lack visible Golgi organelles and so I address whether they primarily lack the organelle or whether the absence of the visible organelle is secondary. Chapter 4 addresses this question indirectly, through the organismal phylogeny of a putatively 'Golgi-lacking' lineage, the oxymonads. Chapter 5 describes genes encoding putatively Golgi-associated proteins from a diverse array of non-model eukaryotes, particularly from lineages that have been proposed to primitively lack Golgi bodies. These genes are proffered as a more direct form of evidence for the presence of a Golgi apparatus, even in the absence of a visible stacked organelle.

In the Conclusions (Chapter 6), I summarize the data obtained in my projects and address its implications, for the two major questions in my thesis. I

also point out future directions that might be fruitful in further investigating the overall evolution of the eukaryotic endomembrane system.

Section 1-Molecular evolution of the vesicular-transport machinery

Although the organelles of the endomembrane system are its most familiar aspect, the protein machinery involved in moving material between the various organelles is just as important a facet. Evolving this vesicular-transport machinery would have represented an important step in the establishment of the system and, as such, provides an interesting subject for rigorous evolutionary examination.

One can envision three stages in the evolution of a complex eukaryotic system from prokaryotic antecedents. Approaching temporally from prokaryotic origins, one reaches the proto-eukaryote, the lineage that split from the rest of prokaryotes to eventually give rise to the eukaryotic clade. The origin of the system then would be the forging, from prokaryotic precursors in the proto-eukaryote, of some unique network of structures or functions. This novelty would not have been present in prokaryotes and would constitute a minimal version of the endomembrane system. While examinations into the origins of a system may be highly speculative (see the section in the introduction on theories of endomembrane evolution), there are more tangible ways to address the problem. Finding prokaryotic homologues addresses the system before its coalescence.

The Last Common Eukaryotic Ancestor (LCEA) is a cell that would have been clearly eukaryotic and that must possess, at a minimum, the characteristics shared by all extant eukaryotic cells. The form of the endomembrane system, as it was present in the LCEA, represents a definite and tractable stage to be examined. Reconstructing the minimal complexity that might have been present

at this node allows us to determine how far along its evolutionary trajectory the system had progressed from its embryonic prokaryotic state to its current form.

The LCEA, however, was nowhere near the final word in vesicular-transport evolution. General expansion of protein families, establishment of paralogues that act at specific locations or transport steps, and lineage-specific duplications have all occurred (Schledzewski, Brinkmann et al. 1999). The vesicular-transport machinery is highly complex, and the evolutionary events involved in its complexification will be some of the most interesting and tractable of all.

Advances in two seemingly disparate fields have greatly improved the opportunities for the evolutionary study of a cell-biological system such as the vesicular-transport machinery. Without the careful identification of the protein components that compose the vesicular-transport machinery, we would have no idea which genes to examine to reconstruct any of the events at the aforementioned time points. As detailed in the Introduction, there is now a great deal known about vesicular transport and the proteins involved. This information has come about due to rapid breakthroughs in the field of cell biology, particular the study of yeast secretion mutants and mammalian neuronal cells (Jahn and Sudhof 1999). The second major advance allowing the study of the evolution of the vesicular-transport machinery has been the recent availability of prokaryotic and eukaryotic genomes. The scale of questions such as the origin of the endomembrane system is so large that, up until recently, they have necessarily been dealt with either superficially (Gupta and Golding 1996; Lopez-Garcia and Moreira 1999; Martin 1999) or in a highly speculative manner (Cavalier-Smith 2002). As explained in more detail in the introduction to Chapter

2, genomics allows these questions to be posed in much more detail, while still maintaining their scope (Schledzewski, Brinkmann et al. 1999).

Reconstructing aspects of the origin, evolution and complexification of the vesicular-transport machinery is done in two ways in this thesis.
Chapter 2 is strictly bioinformatic. Chapter 3 takes a more traditional approach consisting of molecular biological and phylogenetic analyses aided by genomic data.

Chapter 2: A Bioinformatic examination of the origin and evolution of the vesicular-transport machinery.

There has been a great deal of public attention paid to genomics (Enserink 2002; Pennisi 2002). Whether or not the field will eventually live up to its publicity, it has provided researchers with a tremendous amount of raw data with which to work. This vast amount of DNA sequence can be used to answer questions of a scope that was previously intractable (Wolfe and Shields 1997; Wolf, Kondrashov et al. 2001; Anantharaman, Koonin et al. 2002; Bapteste, Brinkmann et al. 2002). Reconstructing a given cell-biological system present in the Last Common Eukaryotic Ancestor (LCEA), for example, can be accomplished by posing the following query: " Are representatives of the major protein families involved in the function of that system present in a well-sampled diversity of eukaryotic taxa?" Such a study might involve hundreds of genes and would be impractical for a single researcher, if each gene had to be experimentally obtained and characterized by standard molecular biological methods. The project would have to be limited, either in the diversity of organisms sampled, or in the number of components studied. More likely a question of that scope would simply not have been undertaken. Having the genome sequence of a given organism already available drastically reduces the amount of time that it takes to identify the homologue of interest, allowing large-scale questions to be addressed concretely and with a data-driven approach. This is particularly true when looking across evolutionarily diverse taxa or for poorly conserved components. Standard experimental strategies, if successful at all, might take months to find and sequence such genes. The sensitivity of homology

searching programs, however, can overcome this sequence divergence and identify the homologues, providing that the sequence data are already available (van den Ent, Amos et al. 2001).

Many broad-scale questions could be asked about the evolution of the vesicular-transport system. I have asked two.

1) Prokaryotic homologues of vesicular-transport machinery components and origin of the system.

Since prokaryotes lack an endomembrane system, and most eukaryotes have one fully formed, the origin of that system must have occurred after appearance of the (prokaryote-like) proto-eukaryote and before the LCEA. While an autogenous origin of the system certainly seems likely, the question still remains, from what proteins did the components of the vesicular-transport system evolve? Some components may have been created well after the split from the prokaryotic lines, from functionally unrelated eukaryotic machinery. However, other components may have been derived directly from proteins present in the prokaryotic ancestor. These precursor proteins might then be identifiable as homologues of vesicular-transport components by using "eukaryote-specific" queries in searching prokaryotic genomes.

2) Reconstructing the vesicular-transport machinery present in the Last Common Eukaryotic Ancestor.

After the split of the proto-eukaryote from our prokaryotic ancestors, but before the establishment of the currently existing eukaryotic lineages lies the last common ancestor of extant eukaryotes. This is an important evolutionary point

after the origin of the endomembrane system, and so knowing how well-developed the system was, i.e. the complexity of the vesicular-transport machinery, at this point would help to develop theories on how the system as a whole evolved. It also identifies a minimal set of vesicular-transport components that underlies the machinery in general.

Each question is addressed in a comparative genomics survey using queries and methods whose rationale are described below.

Protein components of the vesicular-transport machinery to be used as search queries

There are key proteins or protein families involved in vesicular transport. The machinery involved in vesicle formation has as its common components, Arf/Sar1 GTPases, GAPs and GEFs (Springer, Spang et al. 1999). As there are, at least, three major types of vesicles that share some of these common components (Rothman and Wieland 1996), it is possible to search for these types of vesicles by looking for a representative component of their respective coat polymers. Clathrin (the heavy chain) is the obvious representative for clathrin-coated vesicles, while α-COP and Sec31 will be used as representatives of COP I and II vesicles, respectively. The fusion machinery also provides several attractive search query candidates. The Sec1 protein family contains multiple paralogues involved in the same role at various intracellular locations (Jahn and Sudhof 1999). The same can be said for syntaxins and v-SNAREs (Edwardson 1998; Hay 2001). NSF and p97 play key roles in membrane fusion events and as such are

good candidates (Jahn and Sudhof 1999). As general regulators of vesicular transport, Rabs are also important components to examine (Armstrong 2000).

Bio-informatic search methods

The BLAST search algorithm (Altschul, Madden et al. 1997) can be used to find homologues of DNA or protein sequences (queries) by searching genomic databases containing either sequence type. This algorithm aligns the query sequence with others in the database and assigns the alignment a score based on how similar the two sequences are. The reliability of a match by BLAST is measured in expectation (E) values and is usually expressed as a negative exponent. This corresponds to the expected number of alignments that score the same (or better than) as the alignment between the query and a retrieved database entry based on chance alone. This value is corrected for the size of the database and a scoring matrix (Altschul, Madden et al. 1997). The lower the E value; the more significant the match. At some point, the E value drops so low that the server merely states the value as 0. The sequence retrieved from the database as having the most significant E value is commonly referred to as the "top BLAST hit" and this term may be used interchangeably in this thesis.

PSI-BLAST is an iterative BLAST program that uses a scoring matrix based on a consensus of retrieved homologues to increase the sensitivity of the subsequent search. This method can also counteract lineage-specific peculiarities for a given search query such as amino acid composition, rapid evolutionary rate or divergence of a key motif (Altschul, Madden et al. 1997).

There are a variety of reasons why a particular protein may not appear in a genome database, other than its true absence from a genome. Expressed-

sequence-tag (EST) projects provide a snapshot of genes expressed at a given time and under given conditions. If a gene is not expressed at that life-cycle stage, or is expressed in low abundance, then it may not be represented in the database. Genome-sequence survey (GSS) projects are random samplings of a genome, and so by chance a gene of interest simply may not have been encountered at the time that search is performed. Finally, when looking amongst diverse eukaryotes, the gene of interest may have diverged so much in that taxon as to be unrecognizable by a BLAST search. If no homologue can be identified in response to a particular query, then stating simply that a homologue was "Not Identified" is the most prudent response in the majority of cases.

The conservative nature of the "Not Identified" label released me from having to use a method that rigorously excludes claims about the lack of a homologue in a genome. Instead, I was able to use a search strategy that was biased against the other major pitfall, false positive identification of homologues. For each protein component, the relevant query sequence was used in a BLAST search against a given database. All sequences retrieved as possible homologues, given a generous cut-off value for significance, were then reciprocally used as queries for a BLAST search and only those that retrieved the query sequence (and other defined orthologues of it) were noted as true homologues. This procedure struck a balance between allowing for divergent sequence in distantly related taxa (i.e. weak but real BLAST hits) and caution in assigning homology. In cases where the retrieved sequence had already been classified as a homologue of the query, reciprocal BLAST analysis was not performed.

Bio-informatic surveys of diverse genomes were performed using homologues of the above components as queries, to examine the origin and evolution of the vesicular-transport machinery.

Materials and Methods

Search queries

Animal or fungal representatives of each protein family identified in the Introduction were retrieved from Genbank and used as queries for the BLAST analyses. Table 2.1 lists the queries, along with their Genbank accession numbers. These proteins were used as queries since the functional characterization of the protein families that lead to their being chosen as key vesicular-transport proteins occurred in these model systems.

Search Methods

Keyword searching was performed, at all databases that supported this option, to retrieve identified homologues. BLAST analysis was performed at the NCBI BLAST server (http://www.ncbi.nlm.nih.gov/BLAST/) using default settings. Both the BLASTp algorithm, and the PSI-BLAST algorithm when necessary, were used to search the protein databases. The tBLASTn algorithm was used when searching nucleotide databases. A cut-off value of 0.05 was used when selecting potential homologues, and each retrieved sequence was reciprocally used as a query back to the "non-redundant" database. Only sequences that retrieved the initial query sequence at E values below 0.05 were deemed legitimate homologues. Two sets of searches were performed.

Query	Gene family	Paralogue	Taxon	Accession #
Arf	ADP ribosylating factor	ARF1	*Homo*	P32889
Sar1	Secretion-Associated, Ras-related	Sar1	*Saccharomyces*	NP_015106
ArfGEF	Arf-GTP exchange factor	GEA1	*Saccharomyces*	P47102
AP	Adaptin	AP2 alpha subunit	*Homo*	NP_055018
COPII	COP II vesicle coat	Sec31	*Homo*	NP_055748
COPI	Coatomer alpha	Alpha-COP	*Saccharomyces*	P53622
Clathrin	Clathrin	Chc1	*Saccharomyces*	NP_011309
ArfGAP	Arf-GTP activating factor	Gcs1	*Saccharomyces*	NP_010055
v-SNARE	Synaptobrevin	Ykt6	*Saccharomyces*	NP_012725
Syntaxin	Syntaxin	Sso1	*Saccharomyces*	NP_015092
Sec1	Sec1	Syntaxin-binding protein 2	*Homo*	XP_008937
Rab	Rab	Ypt52	*Saccharomyces*	P36018
NSF	N-ethylmaleimide-sensitive factor	NSF	*Homo*	XP_032173
p97	Transitional ER ATPase	TERA	*Homo*	P55072

Table 2.1: **Genes used as queries for comparative genomic survey.** Human and yeast sequences were used as queries since these were the functionally characterized genes. The exact name of the protein used as a query is given, followed by the Genbank accession number. Homo = *Homo sapiens*, Saccharomyces = *Saccharomyces cereviseae*.

In September 2001, searches were performed for a subset of the vesicular transport proteins (Dacks and Doolittle 2001). A second search was performed in September 2002 for an expanded set of vesicular-transport proteins and to search for prokaryotic homologues of the queries listed in Table 2.1. Therefore the homologues identified as B in Table 2.3 were identified in the September 2001 search, and all results are current as of September 2002.

Databases

The nr database at Genbank was the only database searched when attempting to find prokaryotic homologues. The search for eukaryotic orthologues was also primarily performed in the nr database. However, the "others ESTs" database, "HTGS" and the "GSS" databases were also searched. As well, searches were performed at a number of genome-project websites, including the *Dictyostelium* cDNA project (http://www.csm.biol.tsukuba.ac.jp/cDNAproject.html), the *Giardia* genome project (http://jbpc.mbl.edu/Giardia-HTML/index2.html), the *Phytophthora* Genome Consortium (https://xgi.ncgr.org/pgc/) as well as the *Chlamydomonas*, and *Porphyra* genome projects (http://www.kazusa.or.jp/en/plant/database.html).

Results

Prokaryotic homologues of vesicular-transport components

While the endomembrane system is unique to eukaryotes, many of the individual components of that system most likely came directly from our prokaryotic ancestors. This possibility was examined by asking the specific

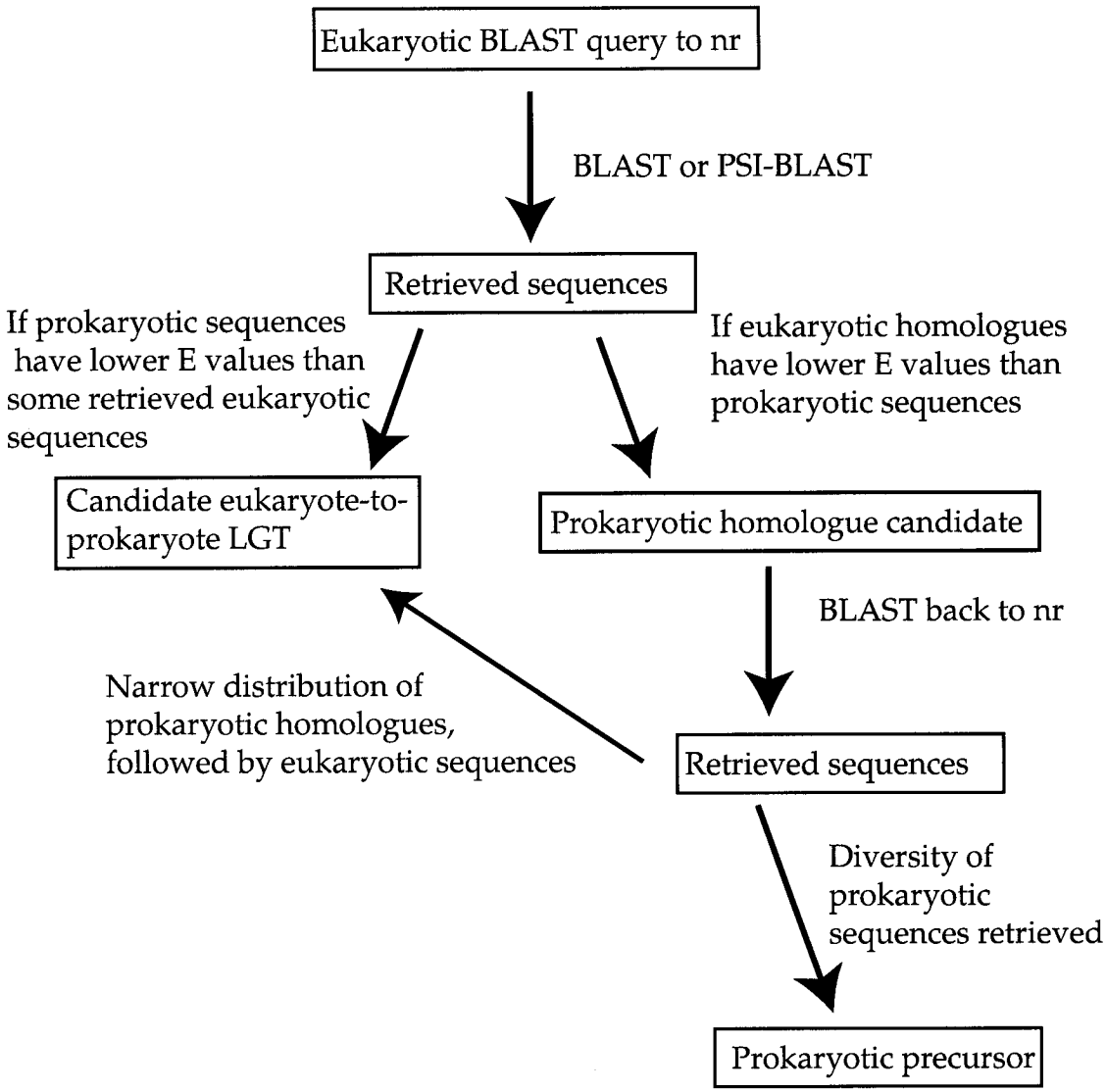


Figure 2.1: **LGT *versus* prokaryotic precursor decision diagram.** This figure illustrates the series of BLAST analyses and criteria used to classify the prokaryotic proteins retrieved from a eukaryotic query.

question, " What protein-coding genes, present in prokaryotic genomes, are homologous to components of the vesicular-transport machinery?"

Eukaryote-to-prokaryote LGTs

Rather than being a precursor, a protein might be homologous to a eukaryotic specific query due to a recent LGT event of its gene from eukaryotes into a prokaryotic genome. Figure 2.1 illustrates the decision flowchart used to determine if a retrieved protein represented a prokaryotic precursor or a putative LGT. Such recently transferred proteins may not have had much time to accumulate random mutations and so may appear as highly significant matches in homology searches. In particular, if the prokaryotic sequence has more similarity to a eukaryotic query than do some other eukaryotic homologues, this similarity may indicate a recent transfer. Prokaryotic proteins derived from LGTs should also be narrowly distributed among prokaryotes (Katz 2002). If, when the proposed prokaryotic homologue is reciprocally used as a query for a BLAST search, eukaryotic proteins rather than prokaryotic ones are retrieved, then a LGT may be suspected.

Two of the prokaryotic proteins identified as homologues of vesicular-transport components, appear to be the result of recent lateral gene transfers from eukaryotes. When using ArfGEF as a query in a BLAST search, the RalF protein from *Legionella* is returned with an expectation value of 5e-19. This protein has been shown to have Arf-modulating activity *in vivo* (Nagai, Kagan et al. 2002). Another protein, from *Rickettsia*, identified as a Sec7 (ArfGEF) homologue, is also returned with an expectation value of 2e-15. Upon reciprocal BLAST analysis, both return each other (E=e-78) and eukaryotic ArfGEFs (E=e-

30) as homologous proteins. However, the RalF protein does not seem to have a wide distribution among prokaryotes. This situation likely represents a lateral transfer to either *Legionella* (gamma proteobacteria) or *Rickettsia* (alpha proteobacteria) and subsequent transfer to the other (Nagai, Kagan et al. 2002). The RecO protein from *Deinococcus* has an identifiable GAP domain at its C-terminal end (E value of 0.051), but other RecO homologues do not. This may be a case of LGT and domain fusion specifically in this taxon.

Prokaryotic endomembrane-system precursors

The criterion used in this study for identifying a protein as a potential prokaryotic precursor was that it was retrieved with a significant BLAST score but with less significance than was observed for orthologues of the query in other eukaryotic taxa. The identified prokaryotic protein also had to be present in a wide taxonomic range (Figure 2.1). Preferably homologues were present in both the Bacteria and the Archea but if not, diversity within one domain was a minimal requirement.

Of the 14 eukaryotic-specific queries from the vesicular-transport machinery, seven belong to larger protein families for which prokaryotic homologues can be identified (Table 2.2).

Diverse GTPase proteins play important roles in the endomembrane system (Springer, Spang et al. 1999; Zerial and McBride 2001). When any one of Arf, Sar1, or Rab was used as a query sequence in a BLAST search, eukaryotic orthologues of the query sequence was retrieved, followed by eukaryotic orthologues of each other and finally other small eukaryotic GTPases (Ras and Rho). PSI-BLAST searches with Arf, Sar1 or Rab query sequences retrieved

Component	Eukaryotic E value	Prokaryotic homologue	Prokaryotic E value
Rab/Sar/Arf	e-05 to e-98	Putative GTPases	Psi-BLAST I2=e-06 to e-11
α-COP	e-26 to e-101	WD-40 proteins	e-20
Sec31	e-100	WD-40 proteins	e-40 to e-79
p97	e-130 to 0.0	cdc48 homologues	e-180
NSF	e-50 to 0.0	cdc48 homologues	e-50

Table 2.2: **Comparison of eukaryotic *versus* prokaryotic endomembrane-component homologues.** The Component column lists the protein family, or families, used as queries with specific queries matching their family designation in Table 1. The Eukaryotic E value column lists the range of expectation-values seen for retrieved eukaryotic homologues.
The Prokaryotic homologue column lists the general assignment of prokaryotic sequences assigned as putative homologues. The Prokaryotic E value column lists the range of expectation value scores seen for putative prokaryotic homologues. In the case of the Arf/Sar1/Rab searches, two iterations (I2) of Psi-BLAST were done before a significant prokaryotic homologue was retrieved. When a single E value is given, this represents the best matching score. When a range is given, then the E values fall within that range.

several prokaryotic GTPases with moderate taxonomic distribution among prokaryotes. When these were used for a reciprocal BLAST search, they would retrieve other prokaryotic GTPases and eventually eukaryotic GTPases (E=5e-13) and translation elongation factors (E=4e-04). Most likely, then, an ancestral GTPase gave rise to the small eukaryotic GTPases, but there is no one clear prokaryotic homologue that can be said to have given rise to the endomembrane-system GTPases. Based on BLAST analysis, the small GTPases (Rab, Sar1, Arf, Ras, Rho) seem to all be more closely related to each other (with E values ranging from e-05 to e-98) than to any given prokaryotic homologue, which require multiple iterations of PSI-BLAST to obtain a significant homologue. However, to verify this relationship and to establish the relationship between the GTPases, phylogenetic analysis will be required.

The proteins chosen as representatives of the coat-forming complexes of both COPI and COPII vesicles, α-COP and Sec31, respectively, both possess WD-40 domains (Schroder-Kohne, Letourneur et al. 1998). This protein domain is present in a wide variety of functionally unrelated proteins but has been implicated as a scaffolding domain, facilitating protein-protein interactions (Li and Roberts 2001). Proteins found in both bacterial and archael genomes also possess very clear WD-40 domains. Using α-COP as a query in BLAST analysis retrieved eukaryotic sequences from various taxa (E=e-26 to e-101) while prokaryotic sequences were obtained with expectation values of approximately E=e-20. Using Sec31 query in a BLAST search retrieved eukaryotic homologues with expectation values scoring at values near of E=e-100. Multiple cyanobacterial sequences were retrieved as potential homologues and, upon

being used as queries in reciprocal BLAST analysis, they retrieved diverse prokaryotic sequences from Bacteria and Archaea (E=e-40 to e-77). It is possible that Sec31 and the alpha subunit of the coatomer complex could have arisen from one or more ancestral proteins containing such a domain. However, since many of the putative prokaryotic homologues are simply assigned as WD-40 proteins without further functional prediction, it is difficult to deduce a specific function for the prokaryotic precursor prior to its co-opting into the endomembrane system.

The AAA-type ATPase family is a well-defined group of proteins associated with wide variety of cellular functions (Ye, Meyer et al. 2001). One member of this family, p97, has been shown to be involved in homotypic membrane fusion events such as post-mitotic reassembly of ER (Latterich, Frohlich et al. 1995) and Golgi (Rabouille, Levine et al. 1995). It also been implicated in a number of additional functional processes, including ubiquitin-dependent protein degradation (Ghislain, Dohmen et al. 1996) and the cell cycle (Moir, Stewart et al. 1982). A second AAA-type ATPase paralogue, NSF, on the other hand, is only known to be involved in SNARE-complex disassembly and recycling (Edwardson 1998). Clear homologues of AAA-type ATPases can be found in both Bacteria and Archaea (Pamnani, Tamura et al. 1997). A BLAST search with p97 as the query sequence yields eukaryotic homologues with expectation values ranging from e-130 to 0.0 and prokaryotic homologues with scores of approximately E=e-150. BLAST analysis of NSF retrieves eukaryotic NSF homologues (E= e-50 to 0.0) and prokaryotic sequences in the E=e-50 range as well. A number of indications point to the possibility that p97 may be the ancestral and pleisiomorphic form of the protein (Zhang, Shaw et al. 2000).

BLAST values for prokaryotic homologues are higher when using the p97 version than with the NSF query. As well, the broad spectrum of cellular processes with which p97 is involved suggests that NSF may have been a specialized offshoot. However, since BLAST values may be affected by evolutionary rate, this issue should be examined by phylogenetic analysis.

No clear prokaryotic homologues were identifiable for syntaxins, v-SNAREs, adaptins, Sec1, and clathrin *via* BLAST analysis.

Reconstructing the minimal vesicular-transport machinery in the "Last Common Eukaryotic Ancestor"

Prokaryotic homologues may indicate from where the building blocks of the endomembrane system might have come. This provides hints as to the origins of the system, but the leap from prokaryotic ancestor/proto-eukaryote to Last Common Eukaryotic Ancestor is a huge one.

Deducing the vesicular-transport system protein compliment at the LCEA would require a fully resolved, broadly sampled and rooted eukaryotic phylogeny. Unfortunately no such phylogeny currently exists, and its discovery does not seem imminent. Nonetheless, reconstructing the ancestral vesicular-transport machinery is still possible. Rather than looking at a designated deepest taxon, diversity may be used to approximate the LCEA. Using comparative genomics to search among diverse taxa for proteins known to be functionally important in the vesicular transport system allows the estimation of a minimal protein machinery present in the last common ancestor of those taxa sampled. The wider the diversity of sampling, the better their last common ancestor approximates the LCEA.

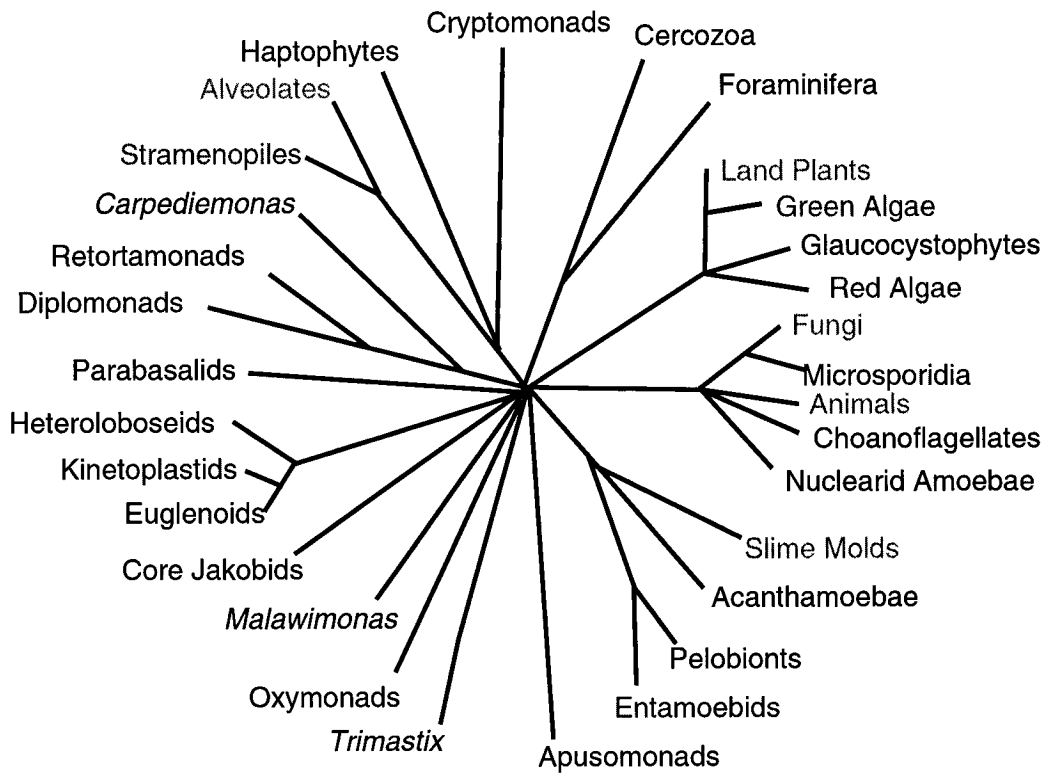


Figure 2.2: **Diversity of genome initiatives as of September 2002.** This schematic of proposed eukaryotic relationships circa 2000 is identical to Figure 1.2, but with publicly available genome projects overlaid on it. Taxa with publicly available EST projects are shown in Red. Those with GSS or genome projects are in Blue. Taxa with both are given in Purple.

The schematic diagram of eukaryotic relationships shown in Figure 1.2 has been reproduced in Figure 2.2. Colour-coded, in this version of the figure, are the major lineages with publicly accessible genomics initiatives as of September 2002. While there are certainly areas of the eukaryotic tree that are underrepresented, the current sampling of genome initiatives provides a crude but reasonable approximation of the LCEA.

The minimal protein compliment of vesicular-transport machinery present in an approximation of the LCEA was addressed by asking, "Are representatives of the major vesicular-transport protein components present in the publicly-available genome initiative databases?".

Several queries were not identified in the *Paramecium* (ciliate) genome, but this project was in its earliest stages at the time of the survey (Dessen, Zagulski et al. 2001). Given that the apicomplexans (sisters to the ciliates in the alveolate super-group (Gajadhar, Marquardt et al. 1991; Patterson 1999) possess all of the components, it is likely that the scarcity of *Paramecium* components is due to sampling rather than true lack. Most genomes examined have, at least, one member of the protein families identified as important components of the vesicular-transport machinery (Table 2.3). Similarly, all genomes had homologues of the representative queries for the three major types of vesicle coats, Sec31, Clathrin and α- COP.

Sar1 is the GTPase responsible for formation of COPII vesicles (Kaiser and Ferro-Novick 1998; Springer, Spang et al. 1999). The Arf protein family is composed of several paralogous sub-families, each of which plays a similar role in the formation of clathrin and COPI vesicles as Sar1 does for COPII

TABLE 2.3A

Higher taxon	organism	Arf	Sar1	ArfGEF	AP	COPII	COPI	Clathrin	ArfGAP
Fungi	*Sacharomyces*	A	A	A	A	A	A	A	A
Land Plants	*Arabidopsis*	A	C	A	A	C	A	C	A
Animal	*Homo*	A	A	A	A	A	A	A	A
Diplomonad	*Giardia*	A	D	D	C	E	B	D	D
Kinetoplastid	*Trypanosoma*	C	E	E	C	E	B	C	E
Apicomplexa	*Plasmodium*	A	E	E	E	C	B	E	C
Slime molds	*Dictyostelium*	C	E	E	D	NI	D	A	C
Entamoebae	*Entamoeba*	C	E	E	E	E	B	E	E
Red Algae	*Porphyra*	D	D	D	D	E	B	D	NI
Stramenopiles	*Phytophthora*	D	E	D	D	D	B	D	D
Green algae	*Chlamydomonas*	A	E	NI	D	NI	B	D	A
Ciliates	*Paramecium*	NI	NI	A	NI	NI	C	C	NI

TABLE 2.3B

Higher taxon	organism	v-SNARE	Syntaxin	Sec1	Rab	NSF	p97
Fungi	*Sacharomyces*	A	A	A	A	A	A
Land Plants	*Arabidopsis*	A	A	A	A	C	C
Animal	*Homo*	A	A	A	A	A	A
Diplomonad	*Giardia*	B	A	B	A	D	D
Kinetoplastid	*Trypanosoma*	B	A	B	A	C	C
Apicomplexa	*Plasmodium*	B	C	B	B	C	C
Slime molds	*Dictyostelium*	B	A	B	A	A	C
Entamoebae	*Entamoeba*	B	B	B	A	E	C
Red Algae	*Porphyra*	B	A	NI	B	NI	D
Stramenopiles	*Phytophthora*	B	A	B	B	E	D
Green algae	*Chlamydomonas*	B	A	B	A	E	D
Ciliates	*Paramecium*	NI-a	NI	NI	C	C	NI

Table 2.3: **Comparative genomic survey of vesicular-transport proteins in diverse eukaryotic genomes.** Table 2.3A shows proteins involved in vesicle formation and movement, while Table 2.3B shows those in vesicle fusion. A = homologues published in separate analyses. B = homologues identified in the September 2001 search (Dacks and Doolittle 2001). C = genes not yet published but found in Genbank. D = genes listed on the respective genome initiative website. E shows when a homologue was found by reciprocal BLAST analysis. NI = a clear homologue was not reliably identified by any of the above criteria. In the case of NI-a, an *Euplotes* (ciliate) homologue has been identified. This table was last updated as of September, 2002.

(Springer, Spang et al. 1999). As can be seen in Table 2.3, the majority of taxa examined have at least one homologue of both Arf and Sar1 present in their genomes. The duplication that gave rise to these two proteins is therefore likely to have occurred prior to the divergence of the taxa examined.

The situation with NSF versus p97 is slightly more complicated. Both proteins are members of a larger AAA-type ATPase family (Ye, Meyer et al. 2001). Most taxa examined seem to have at least one copy of both genes, but both proteins also retrieved cdc48 homologues in eukaryotes, as well as a number of uncharacterized "cdc48-like" eukaryotic Open Reading Frames (ORFs), with significant BLAST scores. This makes it quite difficult to distinguish the presence of p97 versus NSF. As well, while the biological function of NSF is well established, p97 seems to have multiple roles in the cell, membrane fusion being only one of them (Ye, Meyer et al. 2001). As such the biological significance of the duplication is difficult to assess. Although the story will likely be much more complicated, it is possible to deduce, at a minimum, that the duplication which gave rise to p97 and NSF occurred prior to the last common ancestor of those taxa tested.

Conclusions

Origin and evolution of the vesicular-transport machinery

From this comparative genomics survey, it is clear that prokaryotic homologues of supposedly unique vesicular-transport components do exist.

The majority of the vesicular transport protein machinery that is well characterized in model systems also seems to be present in most eukaryotes. This indicates that the entire system is relatively conserved across eukaryotes and that

the model of vesicular transport is broadly applicable to eukaryotes beyond yeast and humans. The mere presence of a homologue does not necessarily imply the same function, but the presence of multiple, interacting, components makes the conservation of mechanism the most parsimonious working hypothesis. The mechanism of vesicular transport needs to be tested *in vivo* in diverse eukaryotes, however, and it will be the difference in function that will tell us exactly how the overall model must be modified to be universally applicable.

It also appears that the last common ancestor of those taxa examined had a complex endomembrane system and that complexification of the various protein families is likely to have begun by the time those taxa examined had diverged.

Limitations of the bio-informatic approach

Although a comparative genomic survey by BLAST analysis can be used to address broad evolutionary questions, there are some severe limitations imposed both by the nature of the algorithm and by the data used.

Many questions of detailed evolutionary history require the identification of a gene sequence at the paralogue subfamily level within a larger gene family. The reliability of such an assignment by BLAST may be compromised since the algorithm does not take into account evolutionary rate, causing the mis-identification of a rapidly evolving sequence. As well, many of the databases provide only partial or poor-quality gene sequence; either end reads of cDNAs or single-pass reads of genomic fragments. These might provide enough conserved sequence for a broad gene family assignment but a sub-family identification may be beyond the boundaries of reliability. Other detailed questions of intra-gene

family evolution may involve establishing the relationship of paralogues, and the timing of their expansion relative to various lineage divergences, or the relationship of various paralogues relative to an outgroup. All such questions are better addressed by further molecular biological characterization of the genes of interest by cloning and sequencing, followed by phylogenetic analysis, as described in Chapter 3.

Chapter 3: Syntaxin protein family evolution

As mentioned in the previous section, questions of diversity and functional evolution within a protein family require molecular biology and phylogenetic analysis. I chose to address such questions for the syntaxin protein family, a member of the SNARE superfamily of vesicular transport components.

The Soluble N-ethylmaleimide-sensitive fusion protein Attachment protein REceptors, or SNAREs, have been high-profile players in the story of endomembrane biology since 1993. These proteins have been implicated in a variety of processes including vesicle tethering (Ungermann, Price et al. 2000), docking (Ungermann, Sato et al. 1998), and fusion (Nickel, Weber et al. 1999). The "SNARE hypothesis" even suggested that SNAREs alone constitute the core of the minimal fusion machinery (Sollner, Whiteheart et al. 1993) and are responsible for the specificity of vesicular transport in the eukaryotic cell (McNew, Parlati et al. 2000). Other workers downplay the importance of SNAREs, particularly in fusion (Peters, Bayer et al. 2001) and as the minimal machinery (Wickner and Haas 2000), and the status of SNAREs as the pivotal piece of the fusion machinery remains controversial. The mechanism of SNARE-SNARE interaction is, in any event, largely understood. SNARE proteins on the incoming vesicle (v-SNAREs) and on the target membrane (t- SNAREs) interact *via* central coiled-coil-forming domains to form four-helix bundles (Misura, Scheller et al. 2000). This brings the membranes in close proximity and leads to vesicle fusion.

One class of t-SNAREs, the syntaxins, forms a clearly delineated protein family based on primary and secondary structure (Bennett, Garcia-Arraras et al.

1993). In addition to the homologous SNARE motif shared by all SNAREs (Fasshauer, Sutton et al. 1998) and forming the coiled coil, syntaxins have 3 N-terminal helices, interspaced with linker regions of variable size (Figure 3.1). These helices, denoted HA, HB and HC have been assigned a regulatory role (Parlati, Weber et al. 1999), as has the linker region between the helices and the SNARE motif (McNew, Weber et al. 1999). Syntaxins generally end in a membrane-spanning domain, although some lack this anchor (Low, Miura et al. 2000).

Syntaxin proteins can themselves be classified into various paralogue families (Bennett, Garcia-Arraras et al. 1993). Each family is involved in either a step in the transport pathway or an intracellular location. There are a number of plasma membrane (PM) localized syntaxins from animals (syntaxins 1-4 and 11), plants (Knolle, Syr1 proteins and others encoded in the *Arabidopsis* genome), and fungi (Sso1 and Sso 2). These will collectively be referred to syntaxin PM homologues. Syntaxin 5 has been implicated in ER-to-Golgi transport complex as well as transport within the Golgi complex (Banfield 1995). Syntaxin 18 (Ufe1 is the yeast homologue), on the other hand, is involved in Golgi-to-ER retrograde transport (Lewis, Rayner et al. 1997; Hatsuzawa, Hirose et al. 2000). Within the highly complicated endocyctic pathway a number of different syntaxins appear to be involved. Syntaxins 6 and 16 are both found doubly localized to the endosomal system and the TGN and both are involved in retrograde transport from the endosome back to the Golgi complex (Abeliovich, Grote et al. 1998; Holthuis, Nichols et al. 1998; Mallard, Tang et al. 2002). Several syntaxin homologues that are exclusively localized to the endocytic organelles are here referred to collectively as 'endosomal syntaxins'. These include Pep12, which is

Figure 3.1: **Syntaxin secondary structure.** Cartoon of the secondary structure demonstrated for several syntaxin molecules, and predicted for all identified syntaxin sequences. The three regulatory helices (HA, B and C) and depicted as yellow barrels, the linker regions between helices are shown as narrow red barrels, while the pink narrow barrels represent the variably present N- and C-terminal sequence. The blue barrel represents the transmembrane helix (TMD).

localized at the prevacuolar compartment in both plants (Bassham and Raikhel 1999) and yeast (Becherer, Rieder et al. 1996) , and Vam3, found at the vacuole in these taxa (Sato, Nakamura et al. 1997; Wada, Nakamura et al. 1997). In metazoan cells, syntaxin 7 (Wang, Frelin et al. 1997; Mullock, Smith et al. 2000; Nakamura, Yamamoto et al. 2000) is localized to the lysosome (the vacuole equivalent). Syntaxin 13 (synonymously named syntaxin 12) is localized to the early endosome and is involved in recycling plasma-membrane markers back to the cell surface (Prekeris, Klumperman et al. 1998).

Another critical function of syntaxin biology is their role in post-mitotic organellar reassembly. Also a vesicular-fusion process, it uses similar but slightly different machinery, with p97 being involved rather than NSF (Linstedt 1999). Some of the syntaxins described above are also involved in the reassembly of the organelles to which they are localized. Syntaxin 5 has been shown to be essential for reassembly of Golgi cisternae and transitional ER (Rabouille, Kondo et al. 1998; Roy, Bergeron et al. 2000). Syntaxin 18 is involved in ER reassembly (Patel, Indig et al. 1998), while Vam3 is critical for the homotypic fusion events in vacuolar reassembly (Wickner and Haas 2000). This indicates a role, at least for these syntaxins, in maintaining the identity and describing the boundaries of a given organelle.

The majority of functional work on SNAREs has been performed using animal and fungal models. Syntaxins are particularly important in neurophysiology, being implicated in neurotransmitter release (Bennett, Calakos et al. 1992) and as the target of botulism toxin (Foran, Lawrence et al. 1996). Secretion mutants of *Saccharomyces cerevisiae* have also been instrumental in understanding the functional biology of syntaxins (Sollner, Whiteheart et al.

1993; Banfield, Lewis et al. 1995). However, prior to this thesis there had been little investigation of the diversity of syntaxins among eukaryotes, with the exception of some studies from plants (Lauber, Waizenegger et al. 1997; Leyman, Geelen et al. 1999) and of a single syntaxin from *Dictyostelium discoideum* (Bogdanovic, Bruckert et al. 2000). A more phylogenetically diverse sampling would help to determine the pattern of duplications giving rise to the syntaxin families *versus* the timing of divergence of various eukaryotic lineages. This can facilitate the deduction of when syntaxins arose and what possible role they may have played in the early evolution of the endomembrane system. Consideration of a diversity of syntaxins also allows a more effective deduction of functional constraints on syntaxin amino acid sequence from patterns of evolutionary conservation.

I have undertaken to expand the available taxonomic diversity of syntaxin sequences using a combined bio-informatic and molecular biology approach. In total, 15 syntaxin genes, from a diverse taxonomic sampling of protists, were sequenced and phylogenetically analyzed.

Materials and Methods

Syntaxin homologue identification

Putative syntaxin genes were identified using the reciprocal BLAST method to search the databases described in the Methods section in Chapter 2. Additional syntaxin searching was performed for *Giardia intestinalis* by downloading the *G. intestinalis* incomplete genome and performing BLAST

searches locally. Two sequences from each major syntaxin family were used as queries to maximize taxonomic and paralogue diversity.

Amplification of syntaxin gene sequences from genomic DNA

Genomic DNA from *Giardia intestinalis*, *Entamoeba histolytica* and *Trypanosoma brucei* was obtained from Andrew Roger, Paul Hoffman and Steven Hadjuk, respectively.

A number of the syntaxin genes were identified from single-pass genomic reads and therefore had to be re-amplified from genomic DNA and sequenced. The identified genes and the sequences of the relevant primers can be found in Table 3.1.

All syntaxin genes were amplified by Polymerase Chain Reaction (PCR) using Taq polymerase and 1/10 volume Pfu polymerase to improve the length of the product and reduce PCR induced errors (Barnes 1994). Amplifications began with 1 minute of melting at 95°C. This was followed by 39 replicates of 1 minute each of melting at 94°C, annealing at 50°C and extension at 72°C. The cycling concluded with 1 minute at 94°C, 1 minute at 50°C and 5 minutes at 72°C. Due to the melting temperatures of the primers, the *Entamoeba* syntaxin 5 and PM genes were amplified with 30 s of melting and annealing temperatures of 45°C and 49°C, respectively.

A number of syntaxin sequences were obtained from *Giardia intestinalis*. *Giardia intestinalis* syntaxin PM was amplified using primers GsynAXF1 and GsynXXR2. *Giardia intestinalis* syntaxin 16 was amplified using primers Gsyn7X1F and Gsyn7X2R. The syntaxin 18 sequence from *Giardia* was amplified

Name	Sequence 5′-3′	Organism	Gene
GsynAXF1	TCATCGCTCCTAGCTACG	*Giardia intestinalis*	Syntaxin PM
GsynXXR2	GTACAGTGCAGCATTTGGCG	*Giardia intestinalis*	Syntaxin PM
GsynPM2XF1	TTAGAAGAGGCGGTCCAAGC	*Giardia intestinalis*	Syntaxin PM2
GsynPM2XR1	CGGTAATTGGCATTGCTCACC	*Giardia intestinalis*	Syntaxin PM2
Gsyn7X1F	GCTCAAACTTGTCGAAGG	*Giardia intestinalis*	Syntaxin 16
Gsyn7X2R	TAAGCACAGCTCATTGCC	*Giardia intestinalis*	Syntaxin 16
Gsyn18XF1	CGGATTCGATTGGTCTTC	*Giardia intestinalis*	Syntaxin 18
Gsyn18XR2	GAAGAGATGACCATCAATC	*Giardia intestinalis*	Syntaxin 18
TBS5X1F	CTCCAACTATGGTTGTAGAGC	*Trypanosoma brucei*	Syntaxin 5
TBS5X2R	ATTTCATTGCCTTGAGACGGC	*Trypanosoma brucei*	Syntaxin 5
TBSyn16XF2	CTGCACTTGAGCGAACTG	*Trypanosoma brucei*	Syntaxin 16
TBSyn16XR2	AGAGAGTGGTAACGATAC	*Trypanosoma brucei*	Syntaxin 16
EHSyn5XF1	TTAATGCCAATTCATCATGG	*Entamoeba histolytica*	Syntaxin 5
EHSyn5XR2	TAATATGACAGACTCATCTG	*Entamoeba histolytica*	Syntaxin 5
EHSynMEMXF2	GGATCCATTCTGTCAGAC	*Entamoeba histolytica*	Syntaxin PM
EHSynMEMXR2	AAGTACAAGTTCAACCCAC	*Entamoeba histolytica*	Syntaxin PM

Table 3.1: **Primers used for exact-match amplification of syntaxin genes.** Primers are listed by name, 5 to 3 sequence, organism from which the gene was amplified and the gene name.

using primers Gsyn18XF1 and Gsyn18XR2. The *Giardia intestinalis* syntaxin PM2 gene was amplified using primers GsynPM2XF1 and GsynPM2XR1. 5′ and 3′ fragments of a syntaxin 5 homologue were detected in the *Trypanosoma brucei* GSS database. The complete ORF was obtained (including the missing internal portion) by amplifying the gene from *T. brucei* genomic DNA using exact-match primers TBS5X1F and TBS5X2R designed to the respective GSS fragments. As well, an ORF corresponding to the 3′ end of a syntaxin 16 homologue was identified in the *T. brucei* GSS database. This ORF was amplified using primers TBSyn16XF2 and TBSyn16XR2.

The *Entamoeba histolytica* syntaxin 5 and PM sequences were also assembled from GSS reads and therefore had to be amplified by exact-match primers. Primers EHSyn5XF1 and EHSyn5XR were used to amplify the *Entamoeba histolytica* syntaxin 5 gene. EHSynMEMXF2 and EHSynMEMXR2 were used to amplify the *Entamoeba histolytica* syntaxin PM gene.

The remaining syntaxin gene sequences were identified as partially sequenced EST or GSS clones, which were then generously provided by the relevant genome projects. In the case of MY-F08, the clone was supplied in transformed *E. coli* cells. All other clones were provided as purified plasmid, which was then transformed into *E. coli* Top10 cells. A list of the genes obtained and their corresponding genome-project clone names is found in Table 3.2.

Cloning and Sequencing

All amplified fragments were cloned into pCR Topo 2.1 vector (Invitrogen, Carlsbad VA) and transformed into *E. coli* Top 10 cells. Sequencing.

Taxon	Assignment	Accession #	Orf size	Clone name	Clone acc #	BLAST	Top Hit
T. brucei	Syntaxin 5	AF404745	327 AA	N/A	N/A	2.00e-11	STX3-*C.elegans*
P. sojae	Syntaxin 5	AF404748	321 AA	5-9d-MY.seq	Pgi:S:2018	1.00e-31	Sed5-*O.Sativa*
E. histolytica	Syntaxin 5	N/A	292 AA	N/A	N/A	2.00e-16	SD07852p- *D.melanogaster*
P. infestans	Syntaxin 6	AF404749	248 AA	MY-38-F-08	Pgi:S:5087	1.00e-16	Putative protein
C. reinhardii	Syntaxin 6	AF404746	225 AA	CM011a08_r	AV386929	4.00e-21	Syntaxin of Plants 61
P. sojae	Syntaxin 7	N/A	224 AA	PSZS006XB18F	PGC:S:1552	2.00e-09	*putsyn7-M.musculus*
P. sojae	Syntaxin 7	N/A	224 AA	PS5007XI03F	PGC:S:2585	2.00e-09	*putsyn7-M.musculus*
T. cruzi	Syntaxin 7	N/A	170 AA	G33N10	N/A	1.00e-04	*Syntaxin7-Rattus norvegicus*
T. brucei	Syntaxin 16	N/A	224 AA	N/A	N/A	7.00e-13	*Syn16A-H.sapiens*
T. vaginalis	Syntaxin 16	N/A	279 AA	Tv1956	N/A	7.00e-21	put.Synprot.-*O.sativa*
G. intestinalis	Syntaxin 16	AF404743	271 AA	N/A	N/A	1.00e-11	U00064_6- *C.elegans*
G. intestinalis	Syntaxin 18	N/A	345 AA	N/A	N/A	2.00e-06	Sim Syn 18- *M.musculus*
G. intestinalis	Syntaxin PM	AF404744	307 AA	N/A	N/A	7.00e-13	SyntaxinA-*C.elegans*
P. yezoensis	Syntaxin PM	AF404747	346 AA	PM059d08_r	AV434474	2.00e-09	Syntaxin-*S.puporea*
E. histolytica	Syntaxin PM	N/A	266 AA	N/A	N/A	3.00e-12	*Syntaxin11-H.sapiens*
G. intestinalis	Syntaxin PM	AF404744	307 AA	N/A	N/A	7.00e-13	SyntaxinA-*C.elegans*
G. intestinalis	Syntaxin PM2	N/A	293 AA	N/A	AF293409	1.00e-104	syn-like *G.intestinalis*

Table 3.2: **Syntaxin genes obtained in this analysis.** This table lists all syntaxin genes obtained in this analysis by organismal source (Taxon), gene assignment, Accession number (if they have already been submitted to Genbank), size of the conceptually translated ORF, and BLAST score for the highest scoring sequence retrieved in a BLASTp search. If the sequence was derived from a clone contributed by a genome initiative, then the clone name and its Genbank accession number is listed.

was performed on an ABI 377 sequencer with 2 clones of each syntaxin ORF sequenced fully in both directions

Structure prediction

Secondary-structure prediction was performed on each obtained syntaxin sequence using the secondary-structure prediction software at the EMBL site (www.embl-heidelberg.de/Services/index.html).

Alignment

Because of the low conservation of syntaxins across the various paralogue families and across eukaryotic diversity, a multi-step approach was taken in creating an alignment that would be robustly homologous. Paralogue-specific alignments were made by aligning regions of amino acid conservation and secondary structure. These alignments were trimmed such that only reliable blocks of sequence were retained, corresponding to the structural helices shown in Figure 3.1. Clustal X (Thompson, Gibson et al. 1997), followed by manual adjustment, was then used to further align the regions. The blocks of sequence corresponding to the helices were then aligned across the paralogue families. While reliable alignment was possible within paralogues across the length of the protein, between paralogues it was only possible to reliably confirm a small region of helix 1 and the SNARE domain as homologous. Several alignments were created. For large-scale analysis, a 65-taxon, 89-amino acid position alignment, and a 50-taxon, 78-amino acid position alignment were constructed. Sub-alignments of all pairwise combinations of paralogue clades were also constructed from the 65 taxon dataset.

Paralogue specific datasets were also constructed for the plasma-membrane syntaxins. These included several pairwise datasets against the syntaxin 5,7 and 16 with the same taxa removed as were eliminated in the above 50-taxon global dataset, and one against the syntaxin 5 clade with all long-branch taxa except the *Entamoeba histolytica* syntaxin PM sequence. Additionally, datasets of 65 and 54 taxa (both with 128 sites) representing the diversity of PM-specific syntaxins and an animal-specific dataset of 30 taxa and 170 sites were constructed.

For the endosomal, syntaxins two datasets were constructed. A dataset containing all endosomal syntaxins (25 taxa) and one with long branches removed (19 taxa) were made, each containing 170 aligned positions.

Phylogeny

Molecular phylogeny, either of DNA or protein sequences, is the primary analysis tool used in this thesis. As described in Chapter 1, there are a number of artifacts that can obscure relationships in a phylogenetic analysis. These are primarily due to the presence of sequences in the dataset that violate assumptions made by the models of evolution or the reconstruction algorithms. In all cases in this thesis, the phylogenetic analysis used the most sophisticated and rigorous phylogenetic methods possible. As well, a variety of phylogenetic methods were used for the analysis of each dataset to avoid results that were due to artifact in a particular method. The presence of sequences that adversely effect the analysis was detected, either by their failure to conform to the amino acid distribution as assessed by Tree-Puzzle 4.0 (Strimmer and von Haeseler 1997), or by their presenting long branches in the best tree topology. In the case of the

DNA analyses in Chapter 4, the RASA program was used (Lyons-Weiler, Hoelzer et al. 1996; Lyons-Weiler and Hoelzer 1999). Whenever long-branch sequences were not crucial to the question at hand they were removed from the dataset as suggested by Hillis et al. (Hillis, Moritz et al. 1996).

Protein Maximum Likelihood (ML) analysis was done using Tree-Puzzle 4.0 (Strimmer and von Haeseler 1997) with a gamma plus invariable sites model of among-site rate variation (8 plus 1 rate categories) with a shape perameter (α) and Pinv estimated from the dataset. Support values from this method are denoted QP in subsequent phylogenies. ProtML 2.2 (Adachi and Hasegawa 1996) with a q1000 search for each dataset was also used. Resampling Estimated Log Likelihood values (RELLs) were calculated using ProtML 2.2 in conjunction with Mol2con (A. Stoltzfus, personal communication) and are given under the ML column in phylogenies. The tree topology shown is the best ML distance tree found by the Fitch-Margoliash method (Felsenstein 1995). ML distance analyses were performed using Tree-Puzzle 4.0 to calculate ML distance matrices in coordination with Puzzleboot (A. Roger and Mike Holder; http://members.tripod.de/korbi/puzzle/). These matrices were then analyzed using Neighbor from the Phylip package (Felsenstein 1995) with jumbling or Fitch with global rearrangements and 10 times jumbling, if dataset size permitted. All bootstrap support values are based on 100 replicates. ML distance bootstrap values are given as MLD in phylogenetic analysis figures.

For the pairwise phylogenetic analyses, sequences were assigned to a particular syntaxin family based, either on their inclusion in that clade during the global phylogenies, or on their BLAST assignment.

Intron detection

Putative introns were detected as regions of nucleotide sequence that disrupt the reading frame of the sequence, contain in-frame stop codons, and fall in between detected regions of homology in BLAST searches. These were conceptually spliced by deleting regions of sequence bounded by GT-AG, and which yielded a contiguous reading frame consistent regions of homology identified in the BLAST algorithm.

Results

BLAST identification of clones

In order to identify putative syntaxin sequences from as wide a range of eukaryotes as possible, various EST, GSS or eukaryotic genome databases were searched. Partial syntaxin clones or sequences identified in this way were completed either by Polymerase Chain Reaction (PCR) amplification of the gene or cloning of the genomic fragment or cDNA. The full sequence was determined by double-strand sequencing of the whole ORF for each syntaxin sequence in Table 3.2. Once the entire ORF was obtained for each putative syntaxin, BLAST analysis of the full-length sequences was used to confirm that they are indeed syntaxin homologues and to give preliminary sub-family assignments. BLAST scores for the syntaxin sequences along with its top BLAST hit are listed in Table 3.2.

Although the homologue retrieved with the highest BLAST score, when the *Phytophthora infestans* syntaxin 6 homologue is used as a query, is a hypothetical ORF, the next 7 retrieved sequences were either syntaxin 6 or their closely related paralogue syntaxin 10 (E=e-16 to e-12). Similarly, when used as a query in a BLAST search, the *Chlamydomonas reinhartdii* syntaxin 6 sequence retrieved Syntaxin of Plants 61 (E=4e-21) but then retrieved syntaxin 6 and 10 homologues with E values as low as to 4e-10. The *Entamoeba histolytica* syntaxin 5 sequence produced a *Drosophila* protein (SD07852p) as its most significant BLAST hit; however, the next 16 retrieved sequences were all syntaxin 5 homologues with values ranging from 4e-16 to 7e-07. When used as a BLAST query, SD07852p retrieves syntaxin 5 homologues with E values of e-70. The *Trichomonas vaginalis* syntaxin 16, despite retrieving a putative syntaxin as its top hit, retrieved syntaxin 16 homologues as its next most significant match (E=1e-18). The *Giardia intestinalis* sequence retrieved a *Caenorhabditis elegans* protein (U00064) as its top BLAST hit. However, it also retrieves syntaxin 16 sequences with E values of e-08. U00064 retrieves syntaxin 16 homologues with E values ranging from e-26 to e-10. In each case then, the syntaxin sequence was assignable to a paralogue family by BLAST analysis.

As seen in Table 3.2, the other 10 putative syntaxin sequences all retrieved specific syntaxin paralogue families when used as queries in BLAST searches. This provided both confidence of their identity as syntaxins and a preliminary guide as to their phylogenetic affinity.

Physical attributes of syntaxin ORFs

The full-length conceptual translations of the obtained sequences fell well within the normal size range for syntaxin proteins and were predicted to share the conserved secondary structure of syntaxins (Figure 3.1). The *Porphyra yezoensis* syntaxin PM gene encodes a C-terminal extension of 54 amino acids after the end of its transmembrane domain. Given the previous data on syntaxin membrane insertion (Bennett, Garcia-Arraras et al. 1993), this region presumably forms a luminal/extracellular extension.

Several of the syntaxin sequences that I obtained from genomic DNA had the reading frame interrupted by in-frame stop codons. The reading frame was restored, however, by conceptually splicing out several GT-AG introns. The syntaxin 5 sequence from *Entamoeba histolytica* contains two putative introns of 51 and 25 nucleotides (nts), respectively. The *Entamoeba histolytica* syntaxin PM homologue also contains 2 introns, this time of 15 nts and 42 nts. The *Trypanosoma brucei* syntaxin 16 sequence also appeared to have an in-frame stop codon near to what corresponds, in other syntaxin 16 orthologues, to the C-terminus of the protein. This may correspond to a truncated version of the protein, in this case. No obvious splice junctions leading to the restoration of a strongly homologous down-stream exon could be identified, and *cis*-spliced introns are rare in *Trypanosoma*(Mair, Shi et al. 2000).

Global Phylogeny

In order to determine the evolutionary affinities of the syntaxin genes obtained, a global dataset was assembled of syntaxins from a broad taxonomic range, using predicted secondary structure of the deduced protein sequences as a

Figure 3.2: **Global syntaxin family phylogeny**. This figure illustrates the best ML distance topology found for the 65-taxon large-scale syntaxin analysis. For this, and all subsequent phylogenies, clades of interest are shaded in gray. Values for nodes of importance are listed in the inset table as Quartet Puzzling/ML distance. For all other nodes an "*" denotes support at better than 70% In ML distance, while "#" denotes better than 90%. In phylogenies where more than two methods were used, these symbols represent support by two of the three methods.

	QP	MLD
A	45	20
B	53	29
C	-	87
D	70	98

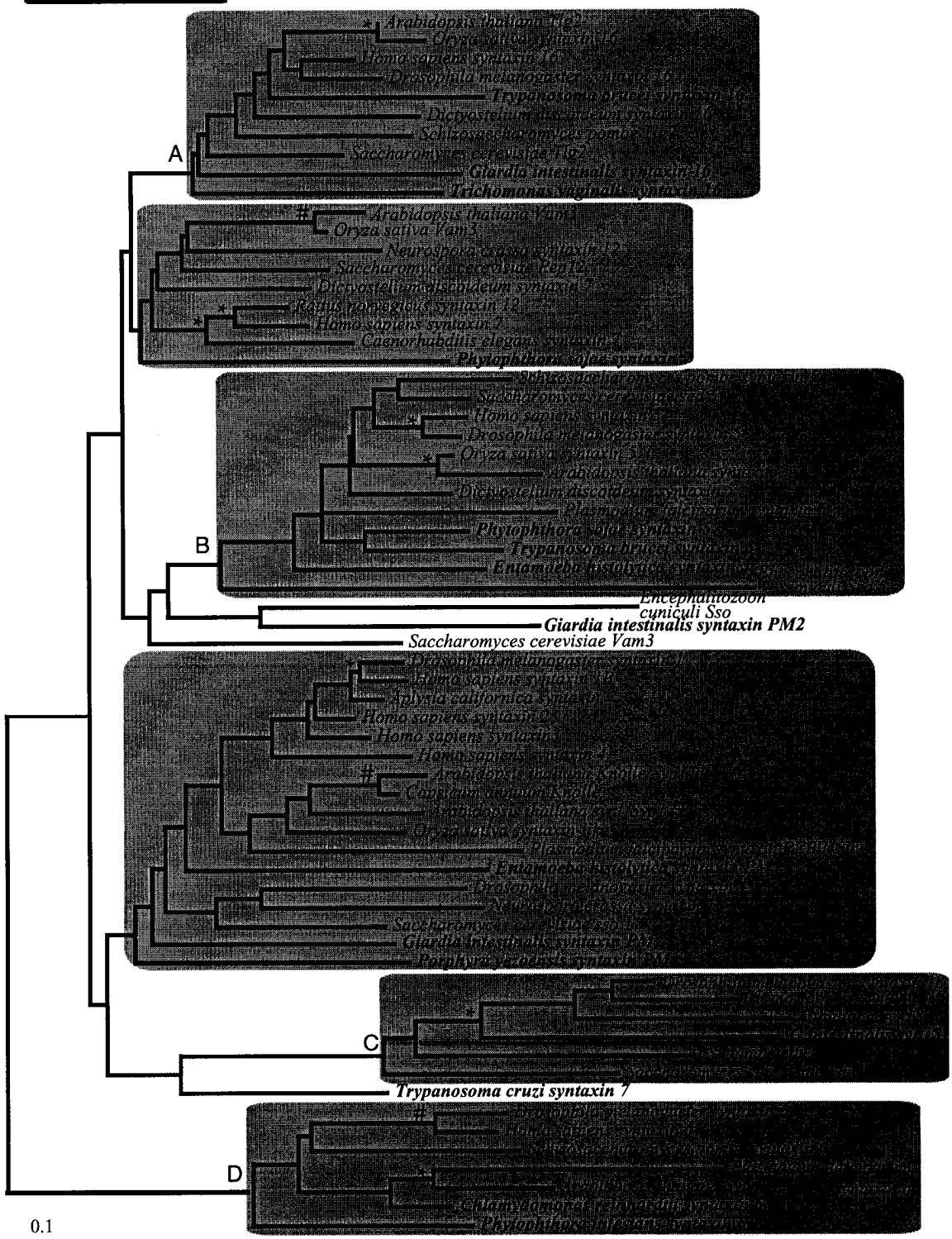
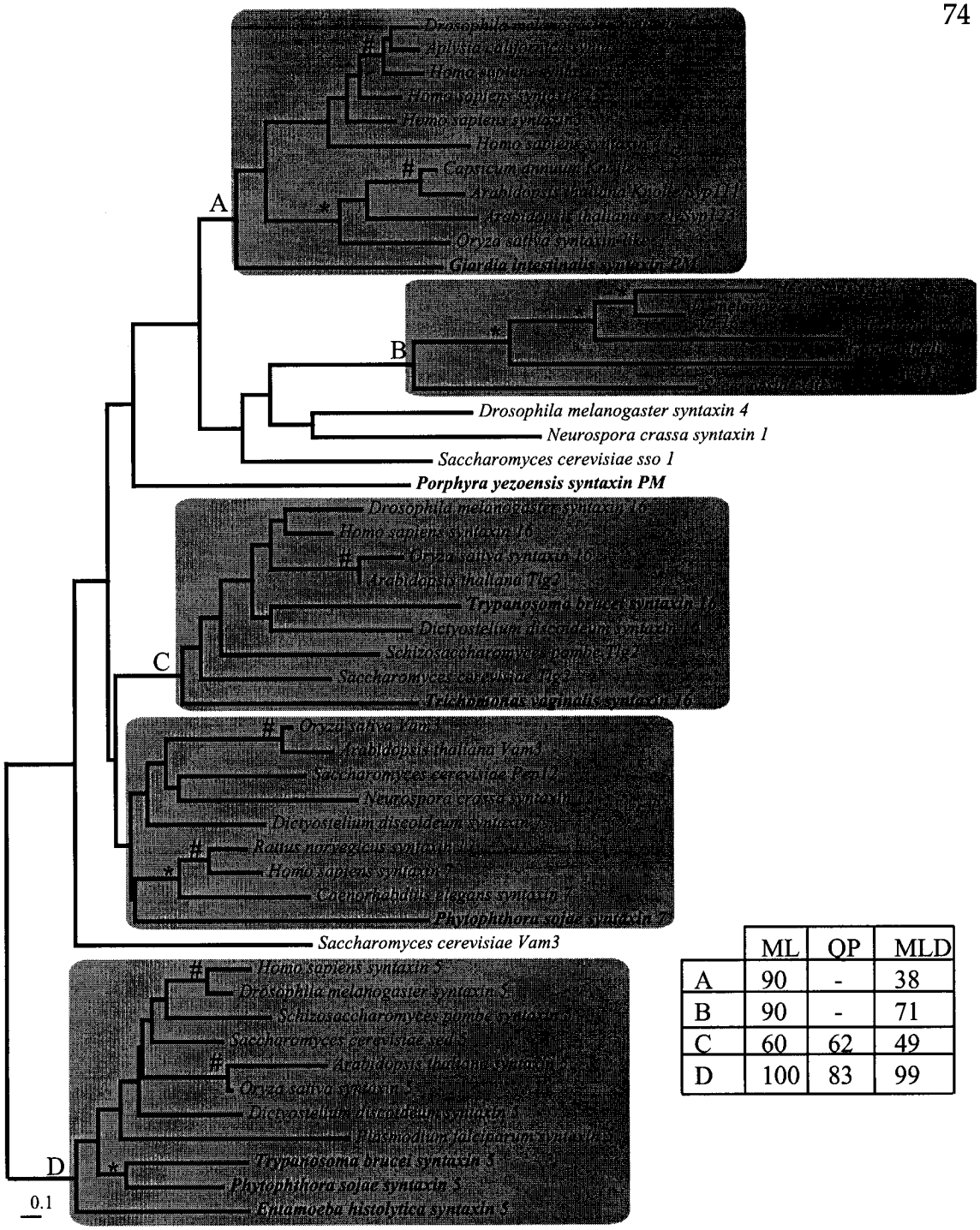


Figure 3.2: **Global syntaxin family phylogeny**

guide for sequence alignment. To be sure of the homology of the regions analyzed, only the unambiguously alignable regions were used, corresponding to a portion of helix 1 and the coiled-coil region/transmembrane domains. This limitation yielded a dataset of 65 taxa and 89 aligned sites which was analyzed using Maximum-Likelihood (ML) and ML distance methods.

As seen in Figure 3.2, the syntaxin 6 family was robustly delineated and separated from the other syntaxin genes. The syntaxin 18 clade was reconstructed in the best ML distance topology, and was strongly supported in ML distance analysis but not by Quartet Puzzling values. In addition to these there were several reconstructed clades in the best tree topology that were composed of only one type syntaxin subfamily (i.e. syntaxin 5, Tlg2, endosomal syntaxin, and syntaxin PM clades), however, these were not supported. A number of sequences fell outside of these clades.

One well-recognized source of artifact in molecular phylogeny is Long Branch Attraction (LBA) (Felsenstein 1978; Philippe, Lopez et al. 2000). In the 65-taxon analysis, several individual, or clusters of, long branches were noted that may have been interfering with resolution in the analysis. Sequences that have amino acid compositions that differ significantly from the average composition may also be artifactually misplaced. Therefore any sequences that failed the amino acid composition test in Tree-Puzzle, or appeared to form long branches in the best ML distance tree, were eliminated. Notably the syntaxin 6 clade, and the sequences derived from the microsporidian *Encephalitozoon cuniculi*, *Plasmodium falciparum* and some of the *Giardia intestinalis* sequences represented long branches. Missing data in an alignment may also cause some phylogenetic analysis programs to miscalculate phylogenies or simply not be able to complete



	ML	QP	MLD
A	90	-	38
B	90	-	71
C	60	62	49
D	100	83	99

Figure 3.3: **Global syntaxin family phylogeny with missing data and long branches removed.** Support values are listed as ProtML RELL values/ Quartet Puzzling/ML distance. Note the improved resolution for the syntaxin 16 and the syntaxin 5 clade with the *E. cuniculi* sequence removed.

their calculations. In order to counteract this effect, the alignment was trimmed of all regions containing missing data. This dataset was then analyzable by the program ProtML, whereas the previous dataset was not. In these analyses (Figure 3.3) the support for the syntaxin 5 clade rose sharply while the support for the syntaxin 16 clade rose to a moderate level. The support for a clade containing all syntaxin 18 sequences remained moderate (71) in ML distance analyses and was quite high in ProtML (90).

Because a number of the syntaxin families were poorly supported in global phylogeny, their monophyly was tested by explicitly asking whether one paralogue family was robustly separated from another given paralogue. This was done by performing phylogenetic analysis of all pairwise combinations of syntaxin families. Also, as some of the analyses may have been obscured by long branch attraction, these pairwise sets were reanalyzed with the previously identified long branch taxa eliminated. The results of the pairwise analyses are summarized in Table 3.3. In all pairwise comparisons not involving the syntaxin PM family, the separation of the various clades was quite strong. The syntaxin PM family was not supported as delineated from the syntaxin 5, 7 and 16 clades. However, upon removal of the long-branch sequences, support for the separation of the syntaxin PM sequences from those assigned to syntaxins 5, 7, and 16 rose significantly. In no case was there significant conflict between the placement of the outgroup roots within a syntaxin family, thus discounting the possibility of strong paraphyly.

Overall the syntaxin paralogue families appear to be monophyletic. Most of the syntaxins that were obtained were assignable to specific families, despite their long-branch nature in several cases.

	Syn6	Syn7	Syn16	Syn18	SynPM	SynPM-LBA
Syn5	100/94/100	98/87/77	100/93/78	100/89/100	49/-/63	100/93/100
Syn6		100/94/100	100/93100	100#/97/100	100#/83/99	na
Syn7			84/71/75	100/89/100	12/-/4	96/65/87
Syn16				100/83/100	70/-/63	68/71/81
Syn18					97/32/97	na

Table 3.3: **Outgroup analysis testing syntaxin family robustness.**

Support values shown are ProtML RELLs, Quartet Puzzling values and ML distance bootstrap values respectively. The syntaxin 7 category encompasses the "endosomal syntaxins, as defined on page 55; all other categories correspond to the sequences enclosed by the boxes in Figure 3.2. The SynPM-LBA column shows the pairwise analyses where several long branch sequences were removed. In this figure, and in subsequent trees, a dash means that this topology was not reconstructed by the method of interest. In analyses with a #, the analysis was done with portions of the dataset removed to eliminate regions of the alignment where not all sequences were represented.

Plasma-membrane syntaxin phylogeny

In order to better assess internal relationships within the plasma-membrane syntaxins, a paralogue-specific dataset was assembled. This enabled the unambiguous alignment of further helices, bringing the total number of aligned sites in the dataset to 128. The best tree topology showed that several long-branch taxa remained within this dataset that were likely causing artifactual misplacement of branches (data not shown). These were consequently removed, yielding a final alignment of 54 taxa, 128 sites (Figure 3.4). The plant and fungal clades were robustly delineated (Figure 3.4, nodes B and C, respectively). The *Drosophila melanogaster* syntaxin 4 homologue was highly divergent, but the clade of animal syntaxins other than this sequence was quite robust (Figure 3.4, node A). There appear to be at least 2 separate cases of expansion in the syntaxin PM family, one in the metazoan (animal) line and one in the streptophytes (land plants).

Sanderfoot et al. (Sanderfoot, Assaad et al. 2000) noted that the SNARE complement of *Arabidopsis thaliana* has been expanded. This is seen particularly in the syntaxin PM family. The streptophyte syntaxins are robustly separated from the Red Algal syntaxin (node C) and form a number of internally resolved clades. From this analysis it appears likely that the expanded complement will be common to streptophytes, since the *Capsicum annum*, *Oriza sativa*, and *Nicotiana tabacum* syntaxins are robustly interspersed with the *Arabidopsis* clades. This indicates that the duplication of some plant SNAREs occurred before the separation of these plant lineages.

Figure 3.4: **Plasma-membrane syntaxin phylogeny with long-branch taxa removed.** This phylogeny shows the syntaxin PM subfamily separated into lineage-specific paralogues. Nodes of interest are denoted with letters that correspond to the support values listed in the inset table. Note in particular the support for the monophyly of the fungal, plant, and animal syntaxins (with the exception of the *Drosophila melanogaster* syntaxin 4 sequence).

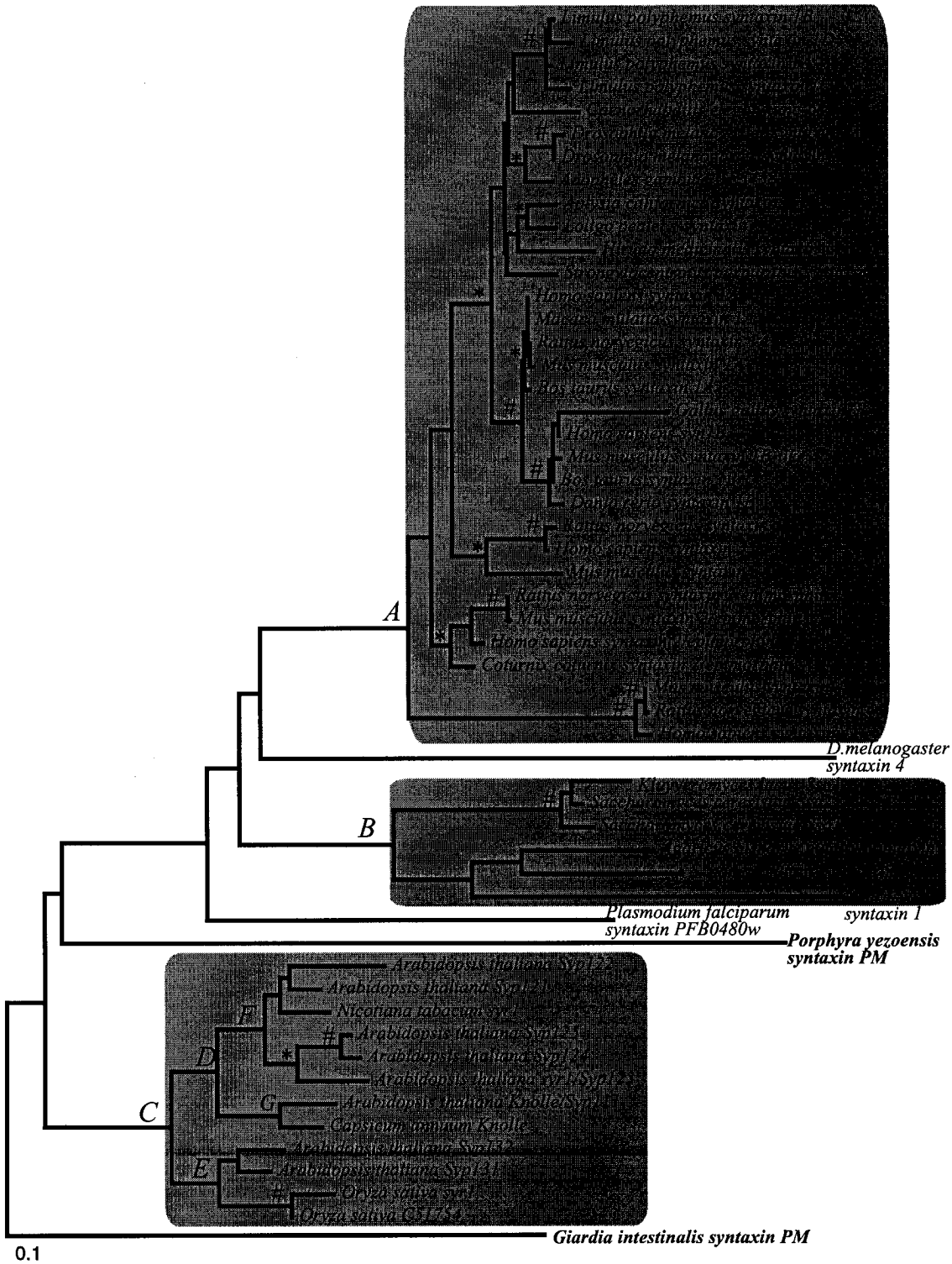


Figure 3.4: **Plasma-membrane syntaxin phylogeny with long-branch taxa removed**

	ML	QP	MLD
A	97	85	78
B	100	96	98
C	59	58	53
D	99	88	90
E	95	78	79
F	44	65	40
G	100	99	100

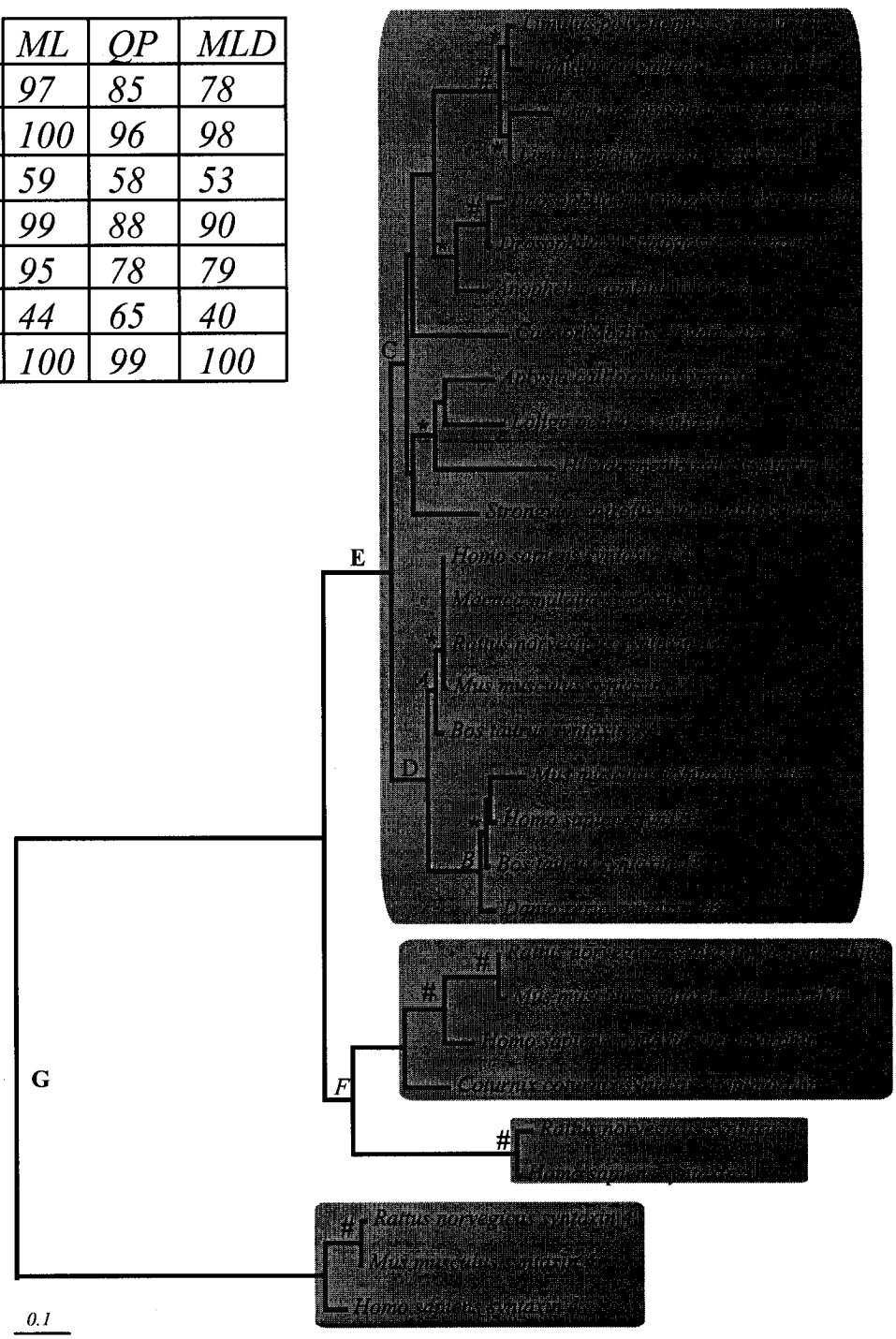


Figure 3.5: **Animal-specific syntaxin PM phylogeny.** This analysis shows the robust separation of the animal-specific paralogues. Note the syntaxin 1 (bolded node E) that includes representatives of both vertebrate and invertebrate taxa and the syntaxin 4 clades

A metazoan-specific syntaxin PM dataset was constructed to assess the relationship between the various paralogues. The syntaxin 4 clade is clearly separated from the rest of the clades (Figure 3.5, node G). There is little resolution with the branching order of the syntaxin 2 and 3 clades (Figure 3.5, node F), while the syntaxin 1 sequences form a strong clade which contains both vertebrate and invertebrate sequences (Figure 3.5, node E). The vertebrate syntaxin 1 sequences also resolve into two robust clades (Figure 3.5, nodes A and B).

Endosomal paralogous family.

Since there are multiple subfamilies of syntaxins localized to endocytic compartments, a paralogue-specific phylogeny was done on this family. The initial analysis showed a number of long-branch lineages (data not shown), leading to their removal and yielding a final dataset of 19 taxa and 170 sites. The endosomal syntaxin sequences cluster by lineage rather than by intracellular location and function (Figure 3.6). The plant sequences are quite robustly separated from the other sequences (Figure 3.6, node I), while the fungal lineage was moderately supported (Figure 3.6, node H). The *Drosophila melanogaster* syntaxin 13 sequence seemed to be highly divergent as compared to the rest of the metazoan endosomal syntaxins and the node supporting their monophyly is weak. Nonetheless, the remaining metazoan syntaxins, including representation from both vertebrates and invertebrates, are reasonably well reconstructed as a clade (node F).

	ML	QP	MLD
A	100	76	100
B	52	79	61
C	100	99	100
D	94	79	93
E	100	99	100
F	**99**	**92**	**86**
G	50	64	84
H	**100**	**66**	**84**
I	**100**	**100**	**100**

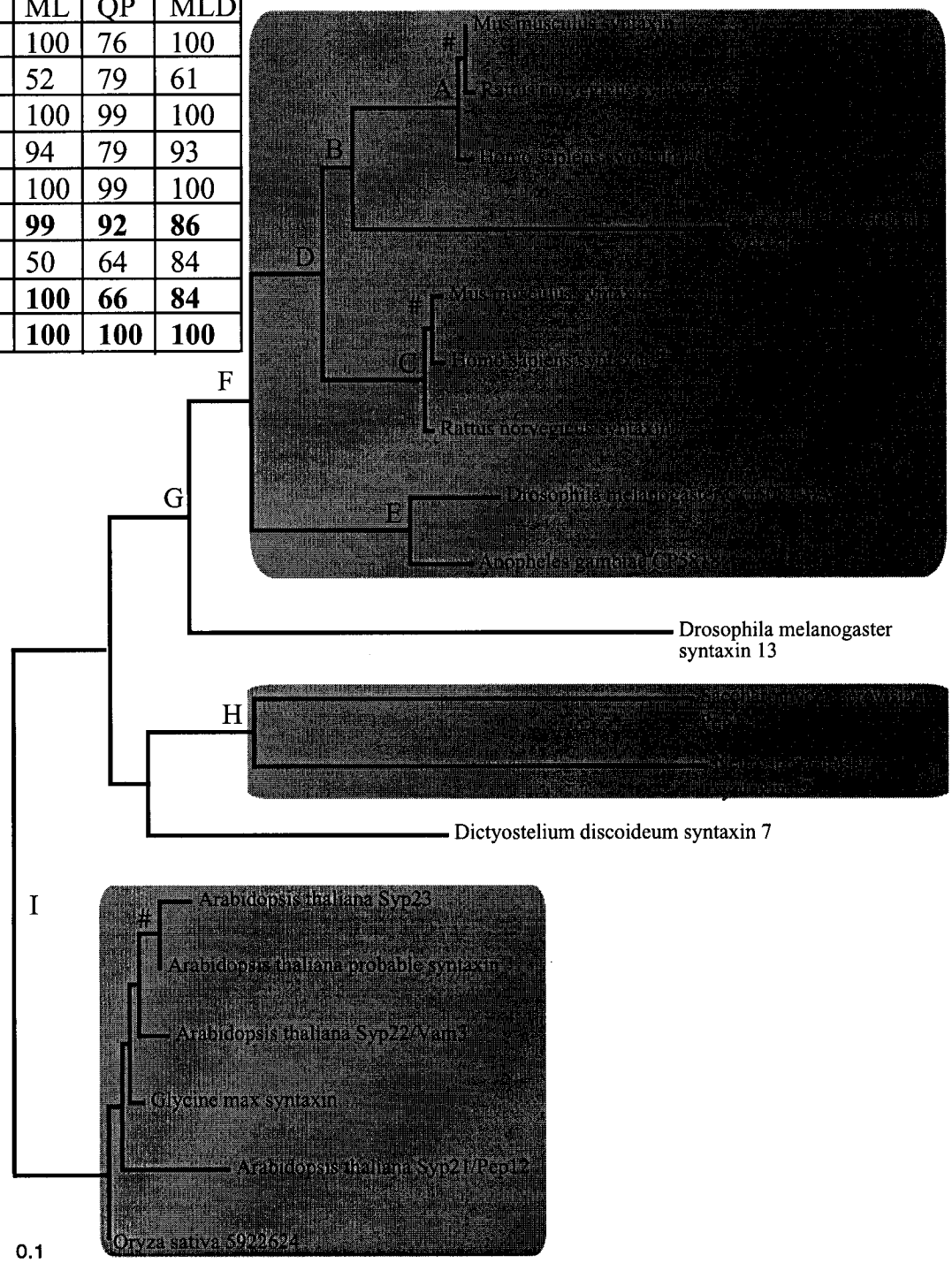


Figure 3.6: **Endosomal syntaxin phylogeny, with long-branch taxa removed.** The endosomal-specific syntaxin families are shown here separated into lineage-specific clades. The nodes supporting the monophyly of the animal, plant and fungal sequences are bolded in the inset table.

Conclusions

The analyses here have revealed a number of points regarding syntaxin evolution.

Evolution of syntaxin functional residues

SNAREs have been classified not only as v- and t-SNAREs, but also in a more evolutionary way using the presence of a functionally critical arginine (R) or glutamine (Q) residue in the coiled-coil domain (Fasshauer, Sutton et al. 1998). Syntaxins were classified as Q SNAREs based on the available syntaxin diversity at the time of that study. While the vast majority of syntaxins obtained do have the conserved Q residue, syntaxin PM homologues from *P. falciparum, E. cuniculi* and *G. intestinalis* do not (Figure 3.7, position A). This may be a result of misalignment, particularly in the case of the *Giardia intestinalis* syntaxin PM2 gene, as there is a Q residue at the adjacent position. Nonetheless, the gene sequences of all three organisms exhibit high rates of evolutionary change (Philippe, Lopez et al. 2000). *Plasmodium* also has a highly skewed genomic GC content (Gardner, Hall et al. 2002). There may well have been a substitution of the Q residue for another amino acid at this position , in these taxa. This possibility should be experimentally examined and if true, it may represent an opportunity to learn about basic syntaxin function from this deviation at an otherwise highly conserved position.

Calcium ions play an important role in membrane fusion events (reviewed in (Wickner and Haas 2000)). Bezprozvanny et al. (Bezprozvanny, Zhong et al. 2000) studied the interaction of syntaxin 1A with $alpha_{1B}$, the pore-forming unit of the N-type Ca^{2+} channels. Wild-type syntaxin 1A reduces channel gating. However, a mutant was generated with Ala 240 to Val and Val 244 to Ala

Figure 3.7: **Aligned SNARE motif for representative syntaxins.** This figure illustrates representative syntaxin homologues from evolutionarily diverse eukaryotes aligned in regions containing the SNARE motif. This corresponds to positions 205-246 of *R. norvegicus* Syn1A. Position A shows the near universally conserved Gln residue characteristic of syntaxins. Position B shows the Ile 236 residue from syntaxin 1A shown to be important for nSec1–syntaxin 1A complex formation. C and D underline the residues involved in Ca^{2+} channel interactions with Syn1.

```
R. nor-syn1A      LENSIRELHDMFMDMAMLVESQGEMIDRIEYNVEHAVDYVE
H. sap-syn1B      LETSIRELHDMFVDMAMLVESQGEMIDRIEYNVEHSVDYVE
H. sap-syn2       LETSIRELHEMFMDMAMFVETQGEMINNIERNVMNATDYVE
H. sap-syn3       LESSIKELHDMFMDIAMLVENQGEMLDNIELNVMHTVDHVE
H. sap-syn4       LERSIRELHDIFTFLATEVEMQGEMINRIEKNILSSADYVE
S. cer-sso1a      LEKSMAELTQLFNDMEELVIEQQENVDVIDKNVEDAQLDVE
A. tha-syr1       IERSLLELHQVFLDMAALVEAQGNMLNDIESNVSKASSFVM
A. tha-knolle     IEKSLLELHQVFLDMAVMVESQGEQMDEIEHHVINASHYVA
E. his-synPM      INDAIEEINGMFVSLAVLIETQGELINSIEENCNSTKEYTK
P. yez-synPM      LAGSLTELHAMFVDMGLLVNQQTELLNNIEANVEKTKVETV
G. int-synPM      IQKTAQEIHQLTMDAAMMCEQQSRLIEQIETNVLHAREAVQ
G. int-synPM2     IHRDVAEVLAMMGLMAEEVHANQETINRIEANVKAADDDVE
E. cun-synPM      IEEMVQDIVDLLNLISQEVSKRTEVVETINDQLITGEENTA
P. fal-synPM      LEKSVCDLHQTIIELSALIEMNDEIIDNIYDHVNDAQYFTE
S. cer-vam3       IHTAVQEVNAIFHQLGSLVKEQGEQVTTIDENISHLHDNMQ
H. sap-syn7       LEADIMDINEIFKDLGMMIHEQGDVIDSIEANVENAEVHVQ
A. tha-vam3       IHQQIGEVNEIFKDLAVLVNDQGVMIDDIGTHIDNSRAATS
D. dis-syn7       IEQSIVEINEIFVDLSGLVAEQGVMINTIEASLESTTINTK
A. tha-syn7       IEDQIRDVNGMFKDLALMVNHQGNIVDDISSNLDNSHAATT
T. cru-syn7       IESNMMDLRSMYQEFHDLVHHQQSNLDSMTGNVSVAKSSVE
P. soj-syn7       INHQLREVNAAFQEIDGLVQDQGEMVVEIVENTDTAKDNVE
D. dis-syn5       ISSTINQLEGIFTQLANLVSMQGEVIERIDLNVS-------
E. his-syn5       IEHMLNELLGLYNHITFLVSTQEEMVRRIDENTEEAVFNVE
H. sap-syn5       IESTIVELGSIFQQLAHMVKEQEETIQRIDENVLGAQLDVE
S. cer-syn5       IESTIQEVGNLFQQLASMVQEQGEVIQRIDANVDDIDLNIS
T. bru-syn5       IEAAVVEVGEMFNDFTRLVHEQNEIVLRIDTNVETSLRHVN
P. soj-syn5       IESHIVDIGQLFGRLSTLIHEQGDLVRRIDDNVEDSLVNVS
A. tha-syn5       VESRITELSGIFPQLATMVTQQGELAIRIDDNMDESLVNVE
A. tha-tlg2       VVESVNDLAQIMKDLSALVIDQGTIVDRIDYNIENVATTVE
G. int-syn16      ITTGIAEIANIITQMSELIYEQGTVLDRIDANVYTAVGYAE
H. sap-syn16      IVQSISDLNEIFRDLGAMIVEQGTVLDRIDYNVEQSCIKTE
T. vag-syn16      MIQSMNQLNELFADLGTLIIQQGTMLDRIDNTIVEAHEQIQ
T. bru-syn16      IVESIKELHTVFESLNSLVVDQGSALDRIDVAIQQTRTSVA
S. cer-tlg2       LARGVLEVSTIFREMQDLVVDQGTIVDRIDYNLENTVVELK
H. sap-syn6       VSGSIGVLKNMSQRIGGELEEQAVMLEDFSHELESTQSRLD
A. tha-syn6       LSKSVQRIGGVGLTIHDELVAQERIIDELDTEMDSTKNRLE
C. rei-syn6       IEQAVIRIGRQGREIGNELAEQERMLDELDQDVDTTHSRLK
P. inf-syn6       LHSDITRLHGVTVEISSEVKHQNKMLDDLTDDVDEAQERMN
S. pom-syn6       VYDTIGNIRGQAALMGEELGQQADLLDTLDNSIETTNSKLR
G. int-syn18      TEQTAHEIQSLNALFAEKVIEQSEQIDRIYAVTFETSGTLD
H. sap-syn18      IEGRVVEISRLQEIFTEKVLQQEAEIDSIHQLVVGATENIK
A. tha-syn18      TETKMVEMSALNHLMATHVLQQAQQIEFLYDQAVEATKNVE
S. cer-Ufe1       INKTILDIVNIQNELSNHLTVQSQNINLMLNNQDDIELNIK
                                      A       B       C   D
```

Figure 3.7: **Aligned SNARE motif for representative syntaxins**

mutations which abolished the wild-type syntaxin 1A effect (Figure 3.7, positions C, D).

When compared across syntaxins there appears to be little pattern to the conservation at these positions. There is variability even within the animal syntaxin 1A versus B. However, alanine and valine at positions 240 and 244 respectively are found in the *G. intestinalis* syntaxin PM sequences, some (but not all) *A. thaliana* syntaxin PM homologues, and the animal syntaxin 7 and syntaxin 5 sequences. At the other extreme, the *Dictyostelium discoideum* syntaxin 5 sequence has a portion of this region deleted entirely. Whether this indicates a functional interaction with Ca^{2+} channels in those proteins with the residues conserved, the maintenance of these residues as a historical accident, or even the drifting of these positions in other taxa due to compensatory changes elsewhere in the protein or the channel is a matter for experimental investigation.

As noted in Chapter 1, the interaction of members of the Sec1 family with syntaxins is crucial for regulation of vesicle fusions (Jahn and Sudhof 1999). Neuronal Sec1 binds to syntaxin 1A in animal nerve cells, where both over- and under-expression of Sec1 are deleterious to syntaxin function (Schulze, Littleton et al. 1994). Studies with neuronal Sec1 showed that mutating several residues in syntaxin 1A significantly decreases Sec1 binding (Misura, Scheller et al. 2000). These residues include isoleucine 236. This position is conserved as isoleucine in most paralogues with the exception of syntaxin 6 where it was replaced with leucine or phenylalanine (Figure Figure 3.7, position B), in some syntaxin 18 sequences and in the *Trypanosoma cruzi* syntaxin 7 sequence. The conservation of the isoleucine suggests conservation of function in those sequences that exhibit it, and an importance of that position for Sec1 homologue binding. The effect of a

substitution of Ile for other hydrophobic residues at this position might yield insight into a possible differential role or mechanism of syntaxin 6 and 18 compared to the other syntaxins. The other residues identified as important for Sec1 binding lie in a region that was not confidently alignable between syntaxin families. However, it was possible to align this region within the syntaxin PM proteins. Misura et al. (Misura, Scheller et al. 2000) found that leucine 165 and glutamate 166, when both mutated to alanine, abolish nSec1-Syntaxin1A complex formation. This region is reasonably well conserved among syntaxin PM orthologues. The position homologous to Leu 165 (Figure 3.8, position A) is always hydrophobic, including a leucine in the *Giardia intestinalis* syntaxin PM, a valine in the *Porphyra yezoensis* syntaxin PM and an isoleucine in the *Entamoeba histolytica* syntaxin PM sequences. The next position (Figure 3.8, position B) is slightly less well conserved, with an aspartate in the *G. intestinalis* sequence and a glutamine in the *P. yezoensis* sequence, but with the conserved glutamate in the *Entamoeba histolytica* syntaxin PM sequence. These two positions are also well conserved in the *Saccharomyces cerevisiae* Sso1 protein and in the *Arabidopsis thaliana* Syr1 and Knolle sequences, indicating the general functional importance of this region. As this region is also one of the binding sites for botulism toxin, it should be of considerable interest for studies aimed at the understanding of syntaxin function.

R. nor-syn1A	SEE**LE**DML
M. mus-syn2	DDE**LE**EML
H. sap-syn3	DEE**LE**EML
H. sap-syn4	DEE**LE**QML
G. int-synPM	DAE**L**DFVI
P. yez-synPM	EADVQAAL
P. fal-synPM	DEDISTFL
S. cer-sso1a	EDEV**E**AAI
A. tha-knolle	DEMI**E**KII
A. tha-syr1	EETV**E**KLI
E. his-synPM	DNVI**E**ESA
	AB

Figure 3.8: **Aligned botulism toxin-binding region of syntaxin PM homologues.** Botulism toxin-binding region from syntaxin PM homologues corresponds to positions 162-170 of *R. norvegicus* Syn1A. Positions A and B show the 2 residues also identified to be important for nSec1-syntaxin complex formation.

Duplications in animal and plant plasma-membrane syntaxins

The syntaxin PM family contains two sets of nested duplications, one in the animal (Figure 3.4, node A) and one in the plant lineage (node C). This expansion of the syntaxins involved in Golgi-to-plasma-membrane transport seems to have occurred twice independently and represents an interesting case of parallel evolution.

Within the plants, there are three highly supported clades, corresponding to a Knolle family (Figure 3.4, node G), a Syr1 family (node F) and another that has not yet been functionally characterized (node E). The Knolle protein is involved in cytokinesis (Lauber, Waizenegger et al. 1997) and is expressed only at specific times in the cell cycle (Lauber, Waizenegger et al. 1997; Volker, Stierhof et al. 2001). Syr1 is involved in hormonal regulation of ion concentration, having to do with the opening of the stomata, and is expressed in drought conditions (Leyman, Geelen et al. 1999). Both proteins are therefore associated with highly specific plant functions and tightly regulated, in line with their potentially later evolution in the plant syntaxin PM clade (node D). Since the red-algal (*Porphyra*) syntaxin PM homologue emerges prior to the monophyletic green-plant clade, it appears that the expansion of the syntaxin PM sub-family occurred after the red- and green-algal divergence (Moreira, Le Guyader et al. 2000). Each syntaxin PM clade within plants contains an *Arabidopsis* representative and one from another streptophyte lineage. This suggests that the expansion of paralogues had already occurred prior to the *Oryza/Arabidopsis* divergence, which is the deepest divergence among those taxa represented (Kuzoff and Gasser 2000). Having a syntaxin PM sequence from a gymnosperm, a bryophyte or a green alga would be most useful in further narrowing the time

frame for the beginning of the syntaxin PM expansion in plants (Nickrent, Parkinson et al. 2000).

In the animal lineages, syntaxins 1-4 are well characterized. All are involved in exocytosis, but while syntaxin 4 is constitutively sent to the basolateral region of epithelial cells, syntaxins 2-3 are apically associated (Low, Miura et al. 2000). Syntaxins 1-3 are robustly separated from the syntaxin 4 clade (Figure 3.5, node G) and the root of animal syntaxins is weakly placed on the branch separating these two groups (Figure 3.4). It is therefore possible that the first duplication of animal PM syntaxins was associated with the evolution of cell polarity in metazoa. Syntaxin 1 is found in the nerve synapse and is involved in neurotransmitter release (Bennett, Calakos et al. 1992). Since syntaxin 1 is clearly separated from syntaxins 2 and 3, and since syntaxin 1 is present in both vertebrate and invertebrates, the duplication giving rise to this paralogue may have been associated with the evolution of the nervous system.

The nearest outgroup to animals, among represented taxa for the syntaxins, are the fungi, which do not share this paralogue expansion. However, the expansion had already occurred by the time of the vertebrate/invertebrate split (Figure 3.5, node E). A syntaxin PM sequence from a choanoflagellate or a sponge could solidify the timing of these expansions, and particularly the prediction that syntaxin 1 evolved as a nervous-system-specific protein.

Duplications in the endosomally localized syntaxins

The parallel expansion in the syntaxin PM families is an interesting case of duplications involved in multicellularity. However, in each case the basic function of the proteins is similar: vesicular transport from the Golgi to the

plasma membrane associated with a specialized context of vesicle fusion (Bennett, Calakos et al. 1992; Lauber, Waizenegger et al. 1997; Leyman, Geelen et al. 1999; Low, Miura et al. 2000). Somewhat more surprising is the case of the endosomal syntaxins. In each of the animal, fungal and plant systems there appears to be one syntaxin associated with the vacuole/lysosome (Sato, Nakamura et al. 1997; Wada, Nakamura et al. 1997; Wang, Frelin et al. 1997; Mullock, Smith et al. 2000; Nakamura, Yamamoto et al. 2000) and one associated with an earlier compartment in the endocytic pathway (Becherer, Rieder et al. 1996; Prekeris, Klumperman et al. 1998; Bassham and Raikhel 1999). While the functional overlap is not complete in the different taxa, it is similar enough to predict that these might represent two separate but closely related syntaxin families. Instead, lineage-specific duplications are found in each of the animal, plant and fungal cases (Figure 3.6, nodes G, H, and I). It is possible that these represent independent cases of gene conversion, between Vam3 and Pep12 at the base of plants and fungi respectively and a between syntaxin 7 and 13 at the base of metazoa. This explanation is rendered less likely by the fact that, in *A. thaliana*, the Vam3 and Pep12 protein sequences are only 64.2% identical while the *H. sapiens* syntaxins 7 and 13 are only 57.4% identical, in alignable regions. To explain these results by gene conversion, one must invoke rapid divergence of the gene sequence. Also one must invoke 3 separate cases of gene conversion in the endosomal syntaxins but none in any of the other families, including the animal syntaxin PMs, which show multiple nested duplications of functionally very similar proteins (Figure 3.5, nodes A-E). The more straightforward explanation for the phylogeny seen in Figure 3.6 is that each lineage had an ancestral endosomally localized syntaxin which underwent a duplication to give

rise to the two functionally different types of syntaxins. This suggests further reason to adopt the naming convention suggested by Sanderfoot et al. (Sanderfoot, Assaad et al. 2000) , i.e. "Syntaxin of Plants #", as that does not imply homology to the fungal syntaxins at a depth that is not warranted. It also implies a striking case of functional convergence in the endosomal syntaxins, and possibly that the fusion machinery of the endocytic systems of animals, plants and fungi are not in fact homologous. Determining whether this phylogenetic pattern is conserved in other components of the endocytic system might distinguish between the two explanations.

Paralogue duplications and the ancient nature of the syntaxin system

The phylogeny of the syntaxin families, and in particular the robust nature of the syntaxin 5, 6 and 18 clades, allows me to deduce the timing of divergence *versus* duplication events in the evolution of the syntaxin superfamily. If a taxon contains at least 2 syntaxins from different paralogue families, then the duplication events giving rise to those families must have occurred prior to the divergence of that taxon. Since syntaxin sequences have been characterized from *Trypanosoma brucei* for both syntaxin 5 and 16, this indicates that the duplication giving rise to these two syntaxins occurred prior to the speciation of *Trypanosoma* and, by extension, the rest of the kinetoplastids. The same argument can be made, although less strongly in accordance with the lower bootstrap support, for the following lineages: *Phythophthora* (syntaxins 5,6, and 7); *Giardia intestinalis* (syntaxin 16, 18 and PM); and *Dictyostelium discoideum* (syntaxins 5, 7 and 16). *Entamoeba histolytica* possesses both syntaxin 5 and PM paralogues since, in pairwise analyses where all long branches but the *Entamoeba* syntaxins were

removed, there was a strong partition between these two clades (100/91/100 in ProtML, QP and ML distance analyses respectively; data not shown). Finally, while there is only one fully sequenced and analyzed syntaxin each from *Chlamydomonas reinhartdii and Trichomonas vaginalis*, these are assigned with reasonable support to syntaxins 6 and 16 respectively, indicating that the duplications that gave rise to these paralogues occurred prior to the divergence of these taxa away from the rest of eukaryotes. The animal, plant and fungal lineages each have representation from all of the syntaxin families.

In the previous chapter I established that at least one member of the syntaxin gene family was likely present in the last common ancestor of organisms whose genomes were sampled. Here the search has been expanded and it is clear that many paralogue duplications had already occurred prior to the divergence of a number of major eukaryotic lineages. This suggests that the complexification of the syntaxin gene family had already begun at an early stage of eukaryotic evolution.

Section 2- Evolution of the Golgi apparatus in eukaryotes

A central tenet of evolutionary biology is the parsimonious argument that complicated features must arise from simpler forms. It was this principle that guided the search in Chapter 2 for prokaryotic homologues of the vesicular-transport system and that drives the search for transitional forms or missing links. It was this logic, too, that prompted Tom Cavalier-Smith to propose the Archezoa Hypothesis (Cavalier-Smith 1983; Cavalier-Smith 1987). This hypothesis proposed that the earliest eukaryotes lacked key features (such as mitochondria, introns and Golgi) and therefore must have left descendants that evolved away from the rest of eukaryotes prior to the evolution of those innovations. While parsimony is a sound guiding principle, it may not hold in a given situation. The Archezoa Hypothesis has been disproven with respect to which taxa were predicted as ancient and for the primary lack of mitochondria (Roger and Silberman 2002) and introns (Simpson, MacQuarrie et al. 2002). Nonetheless, the principle remains a viable one. There are a number of eukaryotic lineages that lack evidence for Golgi bodies, i.e. there is no stacked membranous organelle visible microscopically for these taxa. The term 'Golgi-lacking' will be used to identify that have been proposed not to possess the organelle, based on this lack of microscopically visible stacked organelle. The term does not imply that these taxa do in fact actually lack the organelle. Several of these taxa have been proposed as ancient eukaryotes and are candidates for primitively lacking the organelle.

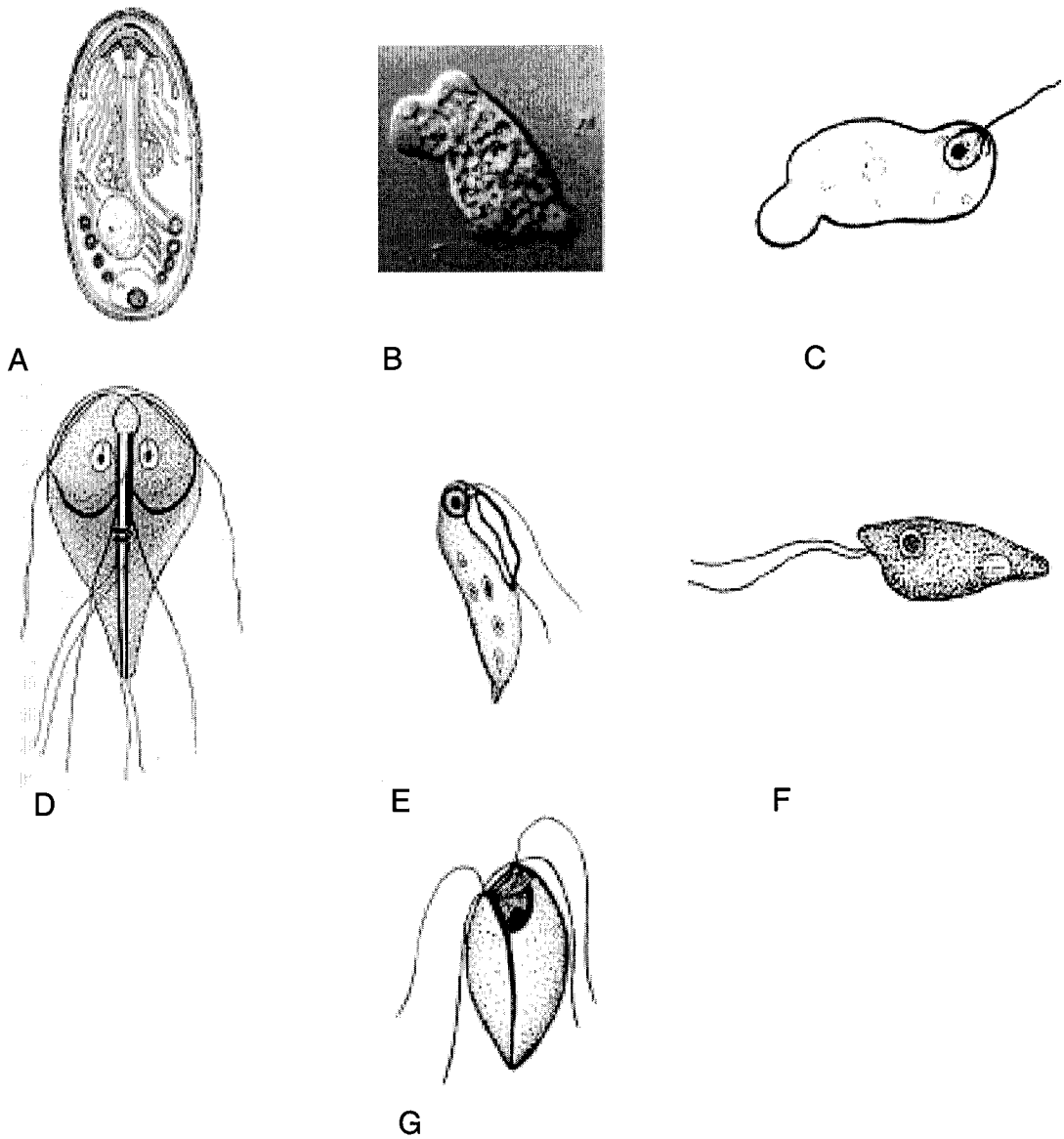


Figure S2-1: **Representative images of 7 major Golgi-lacking lineages.** Modified or taken directly, with permission from the Illustrated Guide to the Protozoa, second edition (Lee, Leedale et al. 2002). (A) Microsporidian spore. (B) *Entamoeba invadens* (entamoebid). (C) Mastigamoebid (pelobiont). (D) *Giardia intestinalis* (diplomonad). (E) *Retortamonas agilis* (retortamonad). (F) *Naegleria gruberi* (heteroloboseid). (G) *Polymastix* (oxymonad).

"Golgi-lacking" Taxa

Although some of these lineages were mentioned in Chapter 1, each taxon will be briefly described here as an introduction to the 'Golgi-lacking' groups as a collection of organisms (Figure S2-1).

The Microsporidia (Figure S2-1A) are intracellular parasites that infect humans as well as a variety of agriculturally important organisms (Keeling and McFadden 1998; Van de Peer, Ben Ali et al. 2000). Originally classified with a variety of other intracellular parasites, the Microsporidia were eventually proposed as early-branching eukaryotes, based on their apparent lack of key organelles and their seemingly prokaryote-like features (Cavalier-Smith 1981), as well as initial molecular phylogenies (Vossbrinck, Maddox et al. 1987; Sogin 1991). More recent work, however, has shown them to be highly derived fungi (Edlind, Li et al. 1996; Keeling and Doolittle 1996; Hirt, Logsdon et al. 1999; Keeling, Luker et al. 2000; Van de Peer, Ben Ali et al. 2000). The genome of the microsporidian *Encephalitozoon cuniculi* has recently been published (Katinka, Duprat et al. 2001). Genes of mitochondrial origin (Germot, Philippe et al. 1997) and the presence of a mitochondrial relic in *Trachipleistophora hominis* (Williams, Hirt et al. 2002) clearly demonstrate the secondary lack of mitochondria in this taxon. Spliceosomal introns are rare in microsporidia (Biderre, Metenier et al. 1998; Katinka, Duprat et al. 2001), but several pieces of the splicing machinery have also been identified (Fast, Roger et al. 1998; Fast, Logsdon et al. 1999).

Entamoebids (Figure S2-1B) are pathogens that infect the intestinal mucosa of animals, with potentially lethal consequences (Patterson, Simpson et al. 2002). Recent concatenated data analyses (Bapteste, Brinkmann et al. 2002) have shown entamoebids to be related to *Dictyostelium*, and *Mastigamoeba*,

although the exact branching order within them is unclear. Acanthamoebids (Baldauf, Roger et al. 2000; Dacks, Marinets et al. 2002; Forget, Ustinova et al. 2002) and *Hartmanella* (Bolivar, Fahrni et al. 2001) are also known to be part of this assemblage. As anaerobic amoebae, entamoebids lack classical mitochondria, although mitochondrial Hsp 60 gene sequences published in 1995 (Clark and Roger 1995) demonstrated the secondary lack of mitochondria in *Entamoeba*. This was followed up in 1999 with the localization of the Hsp 60 protein to a degenerate organelle, the mitosome (Tovar, Fischer et al. 1999) or crypton (Mai, Ghosh et al. 1999). Introns are not common in *Entamoeba* but are present (Wilihoeft, Campos-Gongora et al. 2001).

Members of the original Archezoa, pelobionts (Figure S2-1C) live as anaerobic soil and fresh-water dwellers (Patterson, Simpson et al. 2002). They have aspects of both amoeboid and flagellated cells, with a single flagellum in the anterior end of the cell (Patterson, Simpson et al. 2002). The giant freshwater amoeba, *Peloxyma*, has multiple nuclei and microbody-like intracellular organelles. However, *Mastigamoeba* and not *Pelomyxa* is the best-known pelobiont. This organism is also involved in a peculiar controversy in pelobiont systematics. Initial ssu rDNA sequence from *Mastigamoeba balamuthi* suggested that it was not deep branching (Hinkle, Leipe et al. 1994), but subsequent data from *Mastigamoeba invertans* disputed this claim (Stiller, Duffield et al. 1998; Stiller and Hall 1999). It was assumed that these were simply different species within the pelobionts. However, it now seems that, while *M. balamuthi* is clearly in the pelobiont clade, *M. invertans* may not be (Edgcomb, Simpson et al. 2002). This is still an area of significant controversy, and so *M. balamuthi* is generally considered the representative pelobiont. Combined data analyses as well as

single-gene phylogenies demonstrate that *M. balamuthi* clusters with the entamoebids and *Dictyostelium* in the super-group conosa (Arisue, Hashimoto et al. 2002; Bapteste, Brinkmann et al. 2002). Given the lack of a clear root of the eukaryotic tree, this does not exclude the possibility that the conosa are deep-branching, however. Electrom microscope studies have demonstrated the presence of double-membrane bound bodies resembling hydrogeneosomes in both *Peloxyma* and *Mastigamoeba*. Multiple genes from *Mastigamoeba balamuthi* have been deposited in Genbank that contain spliceosomal introns.

Perhaps the highest profile of the 'Golgi-lacking' taxa, the diplomonads (Figure S2-1D) include *Giardia intestinalis* (which, as the causative agent for giardiasis, is the leading intestinal protozoan parasite in the world) and *Spironucleus barkhanus* (an important parasite infecting salmon) (Sterud, Mo et al. 1997; Adam 2001). Their twin nuclei have provided them with their "diplo" moniker and the four flagella at their anterior end are associated with several cytoskeletal features marking them as excavate taxa (Simpson and Patterson 1999). There is a plethora of molecular data for diplomonads, primarily from *Giardia* (McArthur, Morrison et al. 2000). Phylogenetic analyses place diplomonads as branching deeply in the eukaryotic tree (Sogin 1991; Hashimoto, Nakamura et al. 1994); these same analyses, however, have been firmly stationed in the crosshairs of LBA advocates (Hirt, Logsdon et al. 1999; Stiller and Hall 1999; Philippe, Lopez et al. 2000). Nonetheless, Hsp60 (Roger, Svard et al. 1998; Horner and Embley 2001) and other mitochondrial markers have demonstrated the secondarily amitochondriate nature of diplomonads (Hashimoto, Sanchez et al. 1998; Tachezy, Sanchez et al. 2001; Arisue, Sanchez et al. 2002). Intron-splicing

machinery has been found through the *Giardia* genome project and a single intron has been published for this organism (Nixon, Wang et al. 2002).

Retortamonads (Figure S2-1E) are flagellated anaerobes which possess the characteristic feeding groove of the excavate taxa (Simpson and Patterson 2001). As anaerobic commensals of metazoa, they lack classical mitochondria (Cavalier-Smith 1981). The first molecular data available for retortamonads (ssu rDNA) has shown them to be relatives of diplomonads (as was predicted based on ultrastructural grounds) and to *Carpediemonas* (Silberman, Simpson et al. 2002; Simpson, Roger et al. 2002).

The heteroloboseids (Figure S2-1F) are amoebo-flagellates with a wide array of interesting cell-biological features. The group contains opportunistic pathogens (i.e. *Naegleria fowleri* (Carter 1970)), and both mitochondriate and amitochondriate species. While early molecular phylogenetic analyses suggested a deep placement of the heterolobosea (Cavalier-Smith 1993; Roger, Smith et al. 1996; Roger, Sandblom et al. 1999), these analyses were likely plagued by LBA artifact (Roger, Sandblom et al. 1999; Philippe, Lopez et al. 2000). Combined protein data (Baldauf, Roger et al. 2000), along with the common feature of discoid mitochondrial cristae, argue for a shared ancestry of heterolobosea with Euglenozoa (kinetoplastids, euglenoids and diplonemids) within a superphylum Discicristata (Cavalier-Smith 1983). The common presence of a ventral feeding groove in some Heterolobosea suggests an affinity also with jakobid flagellates, some diplomonads, and other newly described excavate taxa (Simpson and Patterson 2001).

The final major 'Golgi-lacking' lineage is the oxymonads. These amitochondriate protists are endocommensals of termites, cochroaches and some

mammals (Figure S2-1G), and are described in detail in Chapter 4. Until this thesis, their phylogenetic affiliation was unknown.

Logic of primary *versus* secondary absence of a feature

This section outlines my work in addressing questions of how many times the stacked Golgi morphology has been shifted to a non-canonical one, and, most importantly, whether there are any extant eukaryotic lineages that primarily lack a Golgi apparatus. The presence of a cryptic organelle (one which is present but not visible in its canonical form) can be deduced in three ways: through indirect (phylogenetic), genetic and immuno-microscopic evidence. Chapter 4 deals with the indirect type of evidence establishing the phylogenetic affiliation of the oxymonads. Chapter 5 deals with the evidence that I have collected in various 'Golgi-lacking' taxa for genes that, in yeast and mammals, are known to be involved in Golgi function.

Chapter 4: Phylogeny of oxymonads: Indirect evidence against primary Golgi lack.

Phylogenetic affiliation of a 'Golgi-lacking' taxon with one possessing a recognizable Golgi body provides indirect evidence of, either cryptic presence, or secondary loss. This relies on the assumption that the Golgi apparatus evolved only once and that all Golgi bodies are homologous. Given the widespread (Figure 4.1) and conserved nature of the stacked Golgi apparatus (Becker and Melkonian 1996), this is a well-supported assumption. As detailed above and shown in Figure 4.1, there is now evidence for the affiliation of lineages possessing Golgi bodies with all major 'Golgi-lacking' lineages but one.

The oxymonads are a group of structurally distinct, obligately symbiotic flagellates (usually with four flagella per cell), most of which are cellulose digesters found in the hindgut of termites and wood-eating cockroaches (Grassé 1952; Grassé 1952). First described by Leidy in 1877, oxymonads are best known for their distinctive cytoskeletal apparatus which features a cross-linked set of microtubules from which descends a motile axostyle running the length of most oxymonad cells (Brugerolle 1991). Additional outstanding characteristics of oxymonads include highly unusual sexual cycles (Cleveland 1956), a lack of classical mitochondria and, most importantly for this thesis, a 'lack' of Golgi dictyosomes (Brugerolle 1991). This reduced organellar complement, especially the lack of mitochondria, led to oxymonads being proposed as one of the most primitive groups of eukaryotes (Cavalier-Smith 1981).

The relationships of oxymonads with other eukaryotes are uncertain and contentious. They have generally been allied with the other amitochondriate,

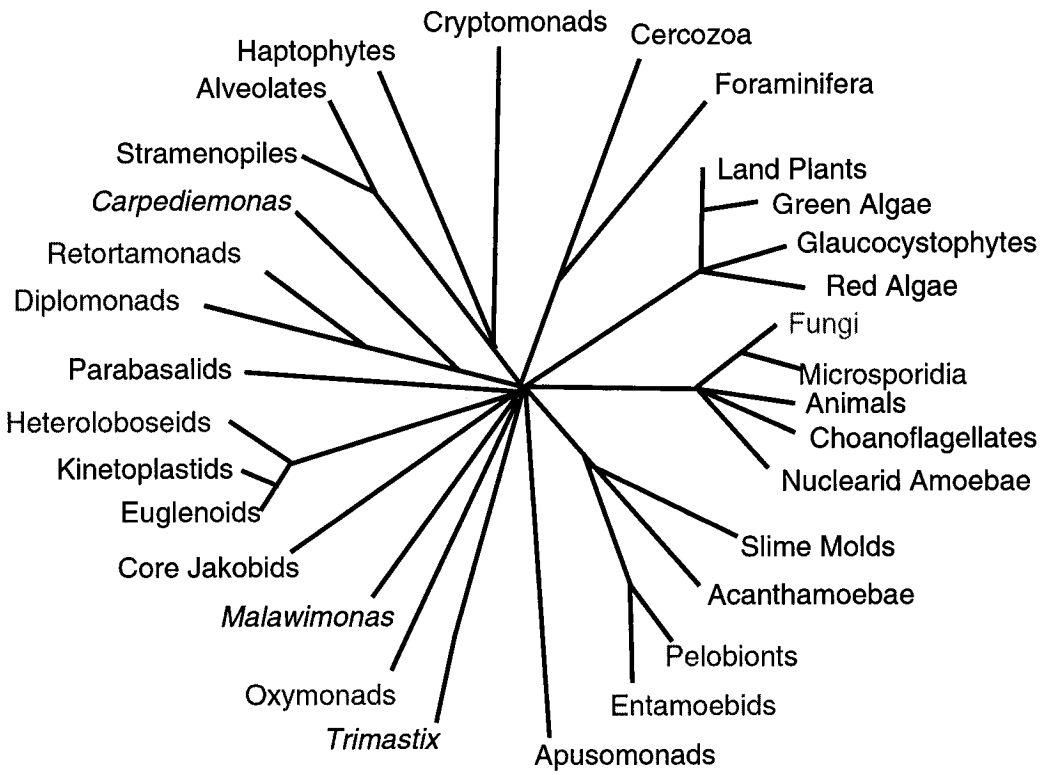


Figure 4.1: **Relationships of 'Golgi-lacking' and -possessing lineages.** Eukaryotic relationships are as shown in Figure 1.2, coloured here according to presence of stacked Golgi organelle. Taxa with stacked Golgi bodies are given in black, those without are shown in red. The fungi are shown in orange, since the majority have no stacked Golgi but there is abundant evidence for the presence of the organelle.

'Golgi-lacking protists possessing 4 anterior basal bodies, i.e., the retortamonads and diplomonads. These groups formed the widely accepted phylum Metamonada, united by their shared possession of four anterior basal bodies and apparent lack of key organelles (Cavalier-Smith 1981; Cavalier-Smith 1998). However, the distinctive presence of a motile axostyle, a cytoskeletal backbone running the length of oxymonad cells, sets the oxymonads apart from the other metamonads. In his 1991 review, Brugerolle suggested that there was "a probable long evolutionary distance between this group and the other two" (Brugerolle 1991). Recent elongation factor (EF-1 alpha) phylogenies that include the first gene sequence data from oxymonads (Moriya, Ohkuma et al. 1998; Dacks and Roger 1999) indicated that a close relationship with diplomonads is unlikely. Newer accounts of eukaryotic diversity instead place oxymonads with Heterolobosea and *Stephanopogon* in the phylum Percolozoa (Cavalier-Smith 1999; Cavalier-Smith 2000), or simply describe them as "eukaryotic taxa without known sister groups" (Patterson 1999).

As they represent a major 'Golgi-lacking' lineage, the placement of the oxymonads is key to understanding the evolution of Golgi, as well as other important eukaryotic features such as mitochondria and sex. I addressed the phylogenetic affiliation of oxymonads by sequencing the ssu rDNA genes from several pyrsonymphid oxymonads. An initial sequence was obtained, verified by *in situ* Hybridization (ISH) by a collaborator and then analyzed. An additional eight clones were subsequently sequenced to test the population diversity of my samples and address some finer-scale evolutionary issues within the oxymonads. Finally, a second analysis of the placement of oxymonads in the tree of

eukaryotes was performed to take into account taxa whose sequence data had not been available at the time of the previous analyses.

Materials and Methods

Protist Isolation and Gene Amplification

Pyrsonympha cells were obtained from specimens of the Western subterranean termite (*Reticulitermes hesperus*), a species known to harbor the oxymonads *Pyrsonympha* and *Dinenympha* (Grosovsky and Margulis 1982), collected from a natural colony near Kelowna, Canada. Termite gut contents were diluted into modified Trager's media (Buhse, Stamler et al. 1975). The largest cells with typical *Pyrsonympha* morphology were selected away from non-oxymonad flagellates by micromanipulation, washed, and reselected. Unfortunately, the cells were identifiable only as *Pyrsonympha* sp., due to the difficulty of manipulation and identification.

About 50–75 cells were pelleted by centrifugation at 3,000 rpm for 1 minute, and DNA was extracted using standard techniques (Maniatis, Fritsch et al. 1982). The 3′ region of the *Pyrsonympha* sp. ssu rDNA gene (639 nts) was amplified by PCR, using eukaryotic specific primer 5′N (TGAAACTTAAAGGAATTGACGGA) and primer B from Medlin et al. (Medlin, Elwood et al. 1988). Cycling parameters began with an initial denaturation of 95°C for 1 minute, followed by 1 minute at 45°C and 3 minutes at 72°C. This cycle was repeated an additional 29 times with the initial heating step at 94°C for 10 s, and was followed by a final cycle with extension time increased to 4 minutes to promote the complete extension of products. The resulting PCR products were cloned into a pGem-T vector (Promega BioTech, Madison, Wis.) and sequenced

on an ABI sequencer. This work was done at UBC from 1995-1998 and is included in my M. Sc. Thesis (Dacks 1998).

Once the identity of this clone was verified by *in situ* hybridization (see below), its sequence was used to design the 3′ primer 3A (ACGCGTGCGGTTCAGATT). This was used with the universal 5′ primer 5A2 (CTGGTTGATCCTGCCAG) to amplify the remaining 5′ component of the oxymonad ssu rDNA gene. The reaction was performed using Taq polymerase augmented with trace amounts of Pfu polymerase to discourage PCR-induced replication errors(Barnes 1994). Cycling parameters of 95°C for 1 minute, 52°C for 1 minute, and 72°C for 3 minutes were used for the first cycle. This was followed by 31 repetitions with the melting step at 94°C decreased to 30 s and one additional cycle with the final extension time at 72°C increased to 4 minutes. The resultant PCR products were cloned into a pCR TopoTA vector 2.1 (InVitrogen, Carlsbad, Calif.). For the sequence *Pyrsonympha* sp. JD2000, two independent 5′ ssu rDNA PCR clones, from separate PCR reactions, were sequenced on a LICOR sequencer. These 5′ ssu rDNA clones provided 1,553 new unambiguously assignednucleotide positions overlapped the previous 3′ fragment by 182 positions. The consensus sequence was assembled based on two- to four-fold coverage of all regions (not always on both strands), with any discrepancies checked against gel traces and nucleotides assigned manually.

In total from the second round of PCR amplification, eight clones ranging from 1,730 to 1,754 nucleotides were sequenced fully in both directions and ambiguous nucleotides manually called. These were assigned the names OS 1,2,6,8,13,15,17, and 19 (subject to further identification by phylogenetic analysis).

It was originally thought that the two clones (OS 6 and 17) used to construct the *Pyrsonympha* sp. JD2000 sequence were clones of the same sequence from the same organism. However, upon full sequencing of the 8 clones, the diversity of the *R. hesperus* population was realized. The sequence annotated *Pyrsonympha* JD2000 is actually a composite of clones OS6, OS17 and the 3′ sequence. The sequence itself is a chimera. However, none of the analyses of the *Pyrsonympha* JD2000 sequence in datasets Global-Eukaryotes 1 and 2 involved regions where there were discrepancies between the two clones. The results of the phylogenetics analyses should therefore not be affected.

Materials and analyses donated by collaborators

The *Trimastix* ssu rDNA gene sequences (made available by Alastair G. B. Simpson), and the RASA and phylogenetic analyses of datasets Global-Eukaryotes 1 and 2 (done in cooperation with Jeffrey Silberman and Mike Holder), were all done as part of the collaboration that lead to the publication of Dacks et al. 2001. While I was involved in analysis (phylogenetic and RASA) of datasets Global-Eukaryotes 1 and 2, the final phylogenies and values shown in Figure 4.3 and Table 4.1 were calculated by Jeffrey Silberman. The phylogeny in Figure 4.4 was calculated by Jeffrey Silberman and Mike Holder. The *Malawimonas jakobiformis, Reclinomonas americana, Jakoba libera,* and *Carpediemonas membranifera* sequences were made available prior to their publication by A. G. B. Simpson as well. The pyrsonymphid sequences derived from *Reticulitermes speratus* gut content, as well as the *Oxymonas sp.* Sequences, were made available by Shigeharu Moriya as part of a collaboration that lead to the publication of Moriya et al. 2003 (Moriya, Dacks et al. In Press).

In situ hybridization studies were performed by Shigeharu Moriya as part of collaborations which lead to the publication of Dacks et al. 2001 (Dacks, Silberman et al. 2001) and Moriya et al. 2003 (Moriya, Dacks et al. In Press).

Phylogenetic Analysis

Five distinct ssu rDNA data sets were analyzed to establish the phylogenetic affinities of the oxymonad sequences that I obtained. To assess the placement of the initial *Pyrsonympha* sp. JD2000 sequence, a dataset containing sequences from a broad diversity of the major eukaryote lineages (Global-Eukaryotes 1) was made (45 taxa and 1,303 aligned positions). A more restricted data set (Global-Eukaryotes 2) containing 31 taxa and 1,447 aligned characters was also constructed. This dataset also had representatives from diverse eukaryotic lineages but was pruned to eliminate long-branch taxa as determined by the RASA analyses (see below). To examine the diversity of oxymonad clones in the *R. hesperus* gut fauna, a dataset (Preaxostyla) was constructed with 18 pyrsonymphid sequences along with three *Trimastix* and the two *Oxymonas* sequences, for a total of 23 taxa and 1310 sites. A more restricted dataset (Pyrsonymphid) was constructed, with 18 taxa and 1621 sites, that contained only the pyrsonymphid sequences. Finally, a third broadly diverse eukaryotic dataset (Global-Eukaryotes 3) was constructed of 72 taxa and 907 sites. This alignment was based on aligned sequences from the "RDPII" database, with the excavate and oxymonad sequences added manually.

Hierarchical log likelihood ratio tests, using the program MODELTEST version 3.0b (Posada and Crandall 1998), showed that a general time-reversible model (GTR) incorporating a correction for among-site rate variation (G) and

invariable (I) sites (GTR+G+I) best described datasets Global-Eukaryotes 1, 2 and 3 as well as the Pyrsonymphid dataset, while a Tamura-Nei model with Invariable sites and among-site rate variation best fit the Preaxostyla data set. The character state rate matrix, the base composition, the gamma shape parameter (α value), and the proportion of invariable sites (I) were similarly estimated by maximum-likelihood methods. This explicit model of nucleotide evolution was used in maximum-likelihood (ML) and distance analyses. For all analyses, gaps were treated as missing data and starting trees were obtained by 100 replicates of random stepwise taxon addition. Branching order and stability were assessed by analyses of 100 or more bootstrapped data sets, except in the case of the Pyrsonymphid dataset where 2000 bootstrap replicates were performed. All phylogenetic analyses were performed using PAUP*, version 4.0b (Swofford 1998).

Kishino-Hasegawa (KH) tests (Kishino and Hasegawa 1989) using PAUP*, version 4.0b, were performed by constraining the backbone ML topology and removing the branch/clade of interest. All possible trees were then constructed by replacing the taxon/clade at each position on the constrained backbone. Significance between the difference in likelihood scores of the alternative tree topologies was tested under a GTR+G+I model of nucleotide evolution. For the Pyrsonymphid dataset, "Approximately Unbiased" (AU) tests of tree topology selection (Shimodaira 2002) were performed using the Paup 4b10 and Consel computer programs (Swofford 1998; Shimodaira and Hasegawa 2001).

Assessment of phylogenetic signal content within the datasets and identification of taxa contributing excessive phylogenetic noise (i.e., putative long-branch taxa) were done by tree independent regression and variance

analyses using the RASA computer package, version 2.3.7 (Lyons-Weiler, Hoelzer, and Tausch 1996), by implementing the analytical model for the estimation of null slope. Plotting the ratio of the variances of phylogenetic (cladistic) similarity to phenetic similarity (taxon variance ratio) identified those taxa which most contributed to branch-length heterogeneity. The phylogenetic signal content of the dataset was reassessed after systematic removal of long-branch taxa. The 31-taxon dataset was also analyzed using the permutation model for the calculation of null slope provided by RASA, version 2.5 (10 permutations) (Lyons-Weiler and Hoelzer 1999).

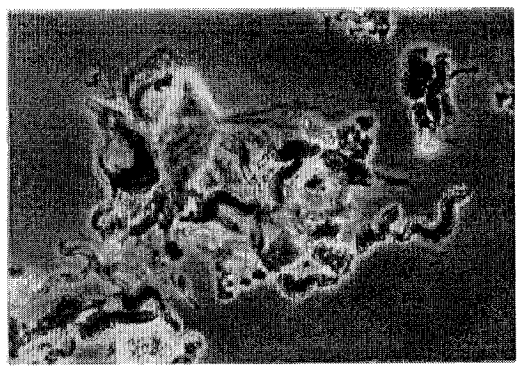
Results

Physical Attributes of ssu rDNAs

The initial sequence *Pyrsonympha* sp. JD 2000 was 2012 nucleotides long. The additional eight 5- end clones ranged in size from 1,730 to 1,754 nts. Although the OS clones were not identified by ISH, the sequences each retrieve the *Pyrsonympha* JBD2000 sequences as their top BLAST hit with E values of 0.0, this confirming that they are derived from oxymonad cells.

Origin Confirmation of *Pyrsonympha* JBD2000 by *In-situ* Hybridization

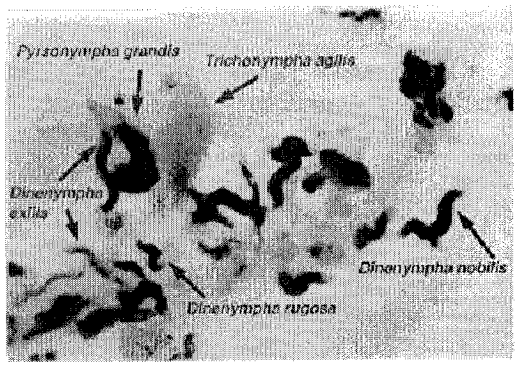
The *Pyrsonympha* sp. cells were obtained from the hindgut of the subterranean termite *Reticulitermes hesperus*, which contains a heterogeneous protist community including three species of each of the two oxymonad genera *Pyrsonympha* and *Dinenympha* (Kirby 1934). The genera are readily distinguished by morphology and size (170 mm average and 25–80 mm, respectively), but the species of each are not readily distinguished, as their sizes and morphologies



A

B

C



D

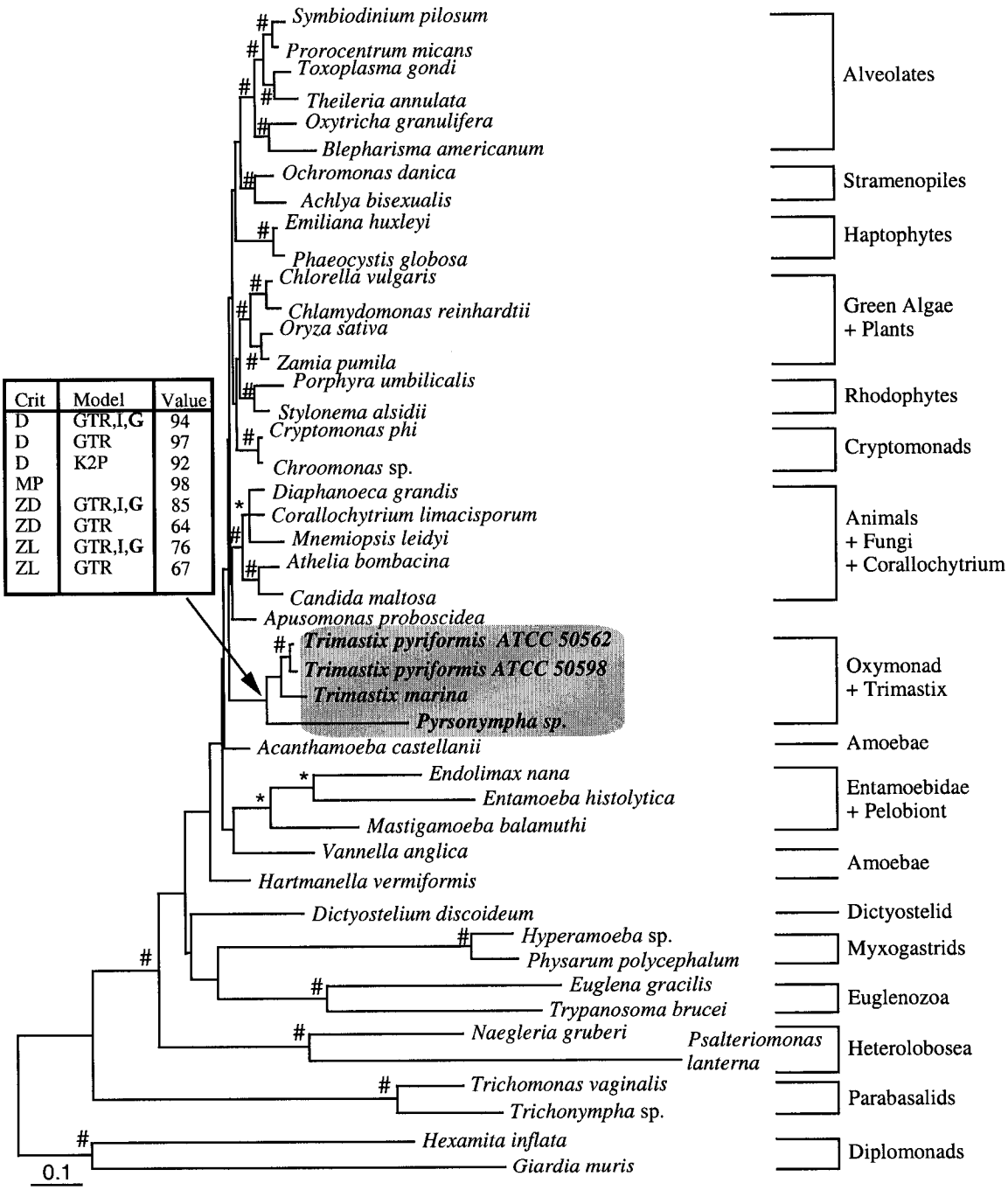
Figure 4.2: ***In situ* micrographs of *Reticulitermes speratus* gut fauna (200X).** (A) Phase contrast of termite gut protists. (B) Gut fauna stained with universal eukaryotic probe (Texas Red-Euk1379) demonstrating that all cells are capable of taking up probe. (C) Gut fauna stained with *Pyrsonympha* JBD 2000 ssu rDNA probe (FITC-Oxy1270). (D) Gut fauna stained with anti-FITC antibody. Panel A shows the same fauna as in D, while B shows the same as in C. Micrographs were taken by S. Moriya.

overlap (Grassé 1952). Because the DNA preparation was not from a pure culture, ISH studies were performed to confirm the source of the ssu rDNA sequence. For convenience, these studies used gut fauna from the closely related Japanese termite *R. speratus*, which contains the same two oxymonad genera (Fig 4.2A). The positive control for ISH experiments was a Texas Red–labeled probe complementary to all eukaryotic ssu rDNA (Euk1379); it annealed to all of the protistan inhabitants of the *R. speratus* hindgut (Fig 4.2B). The FITC-labeled *Pyrsonympha* probe (Oxy1270), which differed from the eukaryote consensus at six strongly conserved positions, hybridized strongly to all cells with *Pyrsonympha* or *Dinenympha* size and morphology but not to non-oxymonad protists (Fig 4.2C). Similar results were obtained when termite gut contents stained with Oxy1270-FITC were examined with anti-FITC antibodies (Fig 4.2D). These results confirm that an oxymonad species was the source of the ssu rDNA sequence obtained. As the ssu rDNA sequenced were amplified from the largest cells with *Pyrsonympha* morphology, we assigned the sequence to *Pyrsonympha* sp.

Phylogenetic placement of the *Pyrsonympha* sp. JD2000 sequence

To test the relationship of the *Pyrsonympha* sp. JD2000 sequence to those from other eukaryotes, an initial phylogenetic analyses was performed on the Global-Eukaryotes 1 dataset, which contains representatives of all major eukaryote groups. With this set, *Pyrsonympha sp.* and the *Trimastix* species formed a clade that was highly supported by bootstrap values under all models and methods of phylogenetic analyses (Fig 4.3) and was recovered in all optimal trees. Within this clade, the *Trimastix* sequences were monophyletic in ML and

Figure 4.3: **Placement of *Pyrsonympha* JBD 2000 amongst eukaryotes (Global-Eukaryotes 1).** The oxymonad/*Trimastix* clade is boxed in gray. The optimal ML (GTR+G+I) tree is shown, with asterisks (*) and pound signs (#) placed at nodes that are supported by better than 70% and 90% respectively with both ML and ML distance methods. For the *Pyrsonympha/Trimastix* node, support values calculated under a variety of methods and optimality criteria are listed. Crit = optimality criteria used in analysis; D = distance; K2P = Kimura 2 perameters; L = likelihood; MP = maximum parsimony; Z = quartet puzzling. The tree is arbitrarily shown as rooted on diplomonads. The phylogeny in this figure was calculated by J. Silberman.



Crit	Model	Value
D	GTR,I,G	94
D	GTR	97
D	K2P	92
MP		98
ZD	GTR,I,G	85
ZD	GTR	64
ZL	GTR,I,G	76
ZL	GTR	67

Figure 4.3: **Placement of Pyrsonympha JBD 2000 amongst Eukaryotes (Global Eukaryotes 1)**

parsimony analyses, but *Pyrsonympha sp.* and *T. marina* were sister taxa under distance methods with the best available model of nucleotide evolution. The strength of the oxymonad/*Trimastix* clade was further examined by performing a series of Kishino-Hasegawa (KH) log likelihood ratio tests under the optimum model of phylogenetic reconstruction (GTR+G+I). The best ML tree from the 45-taxon dataset was used as a backbone constraint in the absence of *Pyrsonympha sp.*, and the *Pyrsonympha sp.* branch was then grafted to each possible branching position. Log likelihood scores for each tree were then calculated. The most likely tree topology was that shown in Figure 4.3. Only five other topologies fell within the acceptable 95% confidence interval. Of these, the top four were simple permutations, with the *Pyrsonympha sp.* branch connecting to all possible nodes within the *Trimastix* clade. Interestingly, the top P value for non-rejected trees was also the optimal topology recovered in distance analyses, a specific relationship of *Pyrsonympha* sp. with *T. marina* (P = 0.2879). Other support values ranged from P = 0.23 to P = 0.07. The least likely topology that failed to be rejected was that with *Pyrsonympha* branching as the sister taxon to *Vanella anglica*, but the value was marginal (P = 0.06). The KH tests were then repeated with the *Pyrsonympha* sequence retained and the *Trimastix* sequences removed, but no other topologies for the placement of the *Trimastix* sequences fell within the 95% confidence interval. Overall, these analyses strongly supported a specific relationship between oxymonads and *Trimastix*.

RASA Analyses of the Global-Eukaryotes 1 data set

Rapidly evolving gene sequences in molecular datasets can produce enough phylogenetic noise to obscure biologically meaningful relationships

(Lyons-Weiler, Hoelzer et al. 1996; Stiller and Hall 1999). To determine whether *Pyrsonympha* and *Trimastix* constitute long-branch sequences, to assess the phylogenetic signal of our 45-taxon data set, and to aid in taxa selection for finer-scale analyses, a series of regression analyses of signal content was performed using the computer program RASA (Lyons-Weiler, Hoelzer et al. 1996). The null hypothesis was that no relationship existed between cladistic signal and phenetic similarity among the sequences tested. For the 45-taxon broad-scale dataset, this hypothesis could not be rejected, indicating that long-branch sequences may be obscuring some phylogenetic signal. Using variance analyses as a guide, ssu rDNA sequences were then removed from the dataset until a statistically significant phylogenetic signal (tRASA = 1.65) was achieved. Table 4.1 shows, that to obtain significant signal content, it was necessary to remove all diplomonads, parabasalids, heteroloboseids, euglenozoans, myxogastrids, and entamoebids and either *Mastigamoeba balamuthi* or *Dictyostelium discoideum*. Thus, any of the relationships in Figure 4.3 involving these lineages may be due to LBA rather than phylogenetic signal. Importantly, *Pyrsonympha* and *Trimastix* do not branch among these taxa (Fig 4.3). Consequently, their placement in trees is likely to be independent of long-branch artifacts.

To explore relationships involving *Pyrsonympha* and *Trimastix*, the diplomonads, parabasalids, heteroloboseids, euglenozoans, myxogastrids, and entamoebids were removed from the dataset, leaving 33 taxa and 1,416 aligned characters, and RASA analyses were repeated. This set did not give a significant phylogenetic signal (Table 4.1). However, when any one of the *M. balamuthi*, *D. discoideum*, or *Pyrsonympha* sp. sequences were also removed, significant phylogenetic signal was recovered. Thus, *M. balamuthi* and *D. discoideum*

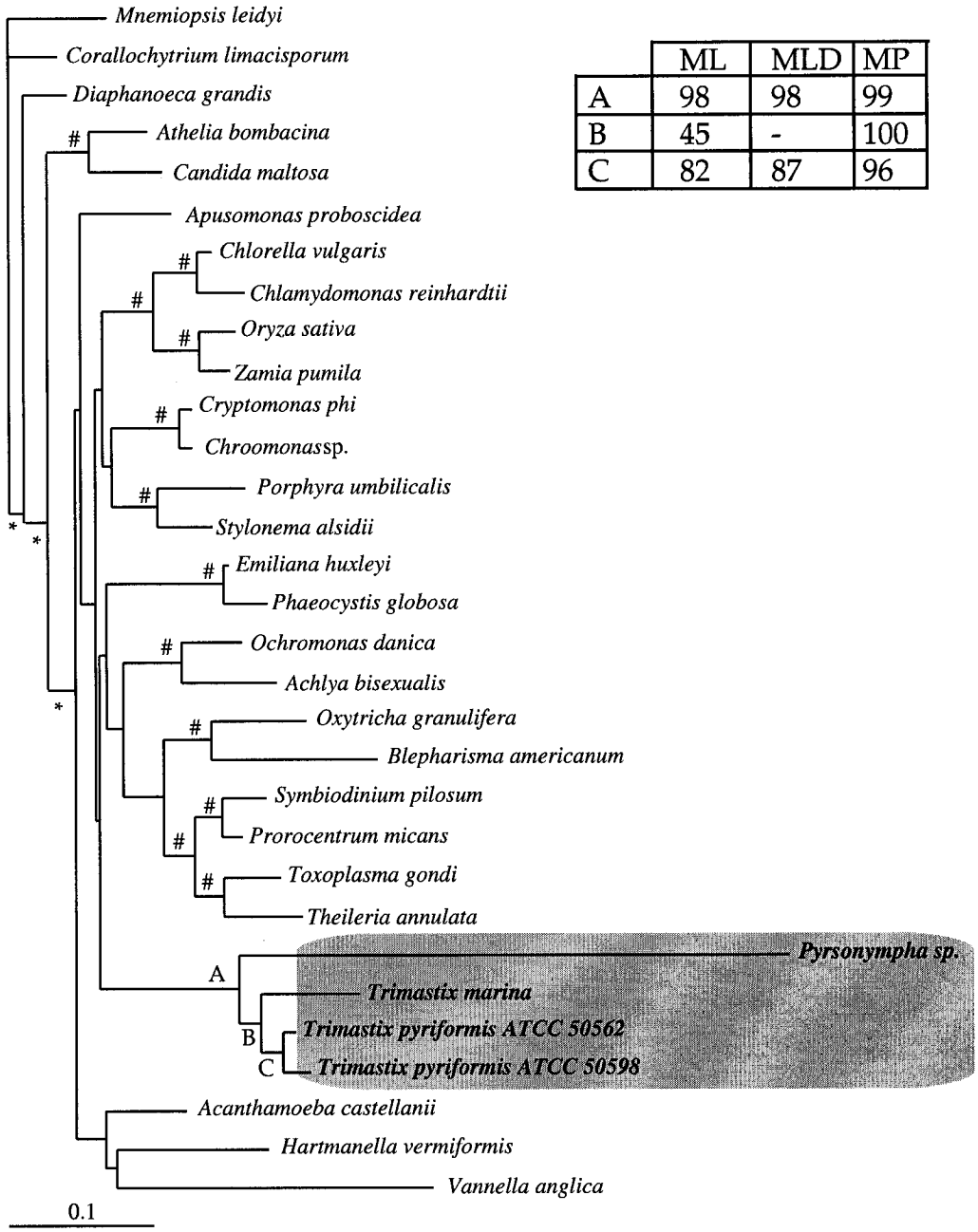
TaxaRemoved	df	tRASA
Setof45taxa		
None	942	- 2.43
A	857	- 3.11
B	857	- 2.60
C	857	- 2.00
D	857	- 2.97
E	857	- 3.25
J	857	- 2.62
A, B	779	- 2.97
A, J	776	- 3.27
A-C	699	- 1.86
A-D	626	- 1.06
A-D,I	557	- 1.40
A-D,J	557	- 1.62
A-E	557	0.19
A-D,I, J	524	- 1.85
A-E,I	524	0.27
A-E,J	524	1.32
A-F	524	0.48
A-E,I, J	461	1.85*
A-F, I	461	0.05
A-F, J	461	2.21*
A-D,F, I, J	492	- 1.66
A-F, I, J	431	3.30*
Setof33taxa		
None	492	0.96
F	461	1.92*
I	461	1.84*
L	461	2.72*
F, I	431	3.60*
F, L	431	4.84*
I, L	431	4.95*
F, I, L	402	11.04*
F, H, I, L	374	16.25*

Table 4.1: **RASA analyses of Global Eukaryotes 1 and 2.** This table depicts the tRASA scores for the Global-Eukaryotes 1 and 2 datasets with various taxa removed. Note-A= diplomonads; B = parabasalids; C = Heterolobosea; D = Euglenozoa; E = myxogastrids; F = Dictyostelium; G = *Hartmanella;* H = *Vanella;* I = *Mastigamoeba balamuthi;* J = Entamoebidae; K = *Acanthamoeba;* L = *Pyrsonympha.* Significant tRASA scores (> 1.65) are indicated by an *. The calculations in this table were done by J. Silberman.

sequences was removed, yielding a data set of 31 taxa and 1,447 characters. RASA analysis under an analytical model then confirmed that this set produced significant phylogenetic signal (df = 431, tRASA = 4.03). Two taxa in this set, *Pyrsonympha* sp. and the amoeba *Vanella anglica*, had relatively high taxon variance ratios and long branches. In fact, when the more stringent permutation model provided by RASA, version 2.5, was used for the calculation of the null slope (Lyons-Weiler and Hoelzer 1999), the presence of these two taxa in the dataset caused tRASA to be below the significance value. However, their wide separation in the phylogenetic trees in Figure 4.3 suggests that LBA between them is not a problem. The removal of both *Pyrsonympha* and *Vanella* from the data set yielded a tRASA value well above the significance value (df = 374, tRASA = 7.39).

Phylogenetic Analysis of the Global-Eukaryotes 2 Data Set

Guided by the RASA results, the number of fast-evolving sequences in the dataset was reduced to better resolve the phylogenetic relationship of *Pyrsonympha* to the different *Trimastix* species and that of the *Pyrsonympha/ Trimastix* clade to other eukaryotic lineages. Use of only these 31 taxa allowed unambiguous alignment of 1,447 nucleotide positions, increasing the power of the phylogenetic analysis. The results from this reduced data set paralleled those of the previous analyses (Fig 4.4). High bootstrap values (98/98/99) under ML, ML distance, and maximum parsimony strongly supported a *Pyrsonympha*-plus-*Trimastix* clade. *Pyrsonympha* was the earliest-diverging taxon within this clade in ML and parsimony analyses, while a weakly supported sister-taxon relationship between *Pyrsonympha sp.* and *T. marina* was observed in distance analyses



	ML	MLD	MP
A	98	98	99
B	45	-	100
C	82	87	96

Figure 4.4: **Placement of *Pyrsonympha* JBD 2000 amongst eukaryotes (Global-Eukaryotes 2).** The optimal maximum-likelihood topology is shown with maximum likelihood, minimum evolution and maximum parsimony bootstrap values shown at relevant nodes. This phylogeny was calculated by J. Silberman and M. Holder.

(bootstrap value of 43). Finally, as with previous phylogenetic analyses of ssu rDNA phylogenies, the major eukaryotic lineages were robustly monophyletic. RASA analyses of the 31-taxon dataset (which included both the *Pyrsonympha* and the *Vanella* sequences) rejected the null hypothesis of no relationship between cladistic signal and phenetic similarity under an analytical RASA model, while analyses under a permutations model did not reject the null hypothesis. For this reason, phylogenetic analyses were performed to determine whether these "long-branch taxa" were masking any other affinities of the *Trimastix* sequences. In the absence of the *Pyrsonympha* sequence, with or without *Vanella*, the *Trimastix* sequences formed a clade with 100% support under parsimony and ML distance models and showed no strong affinity for any other lineage in the data set. Returning the *Pyrsonympha* sequence to the dataset in the absence of *Vanella* produced a robust *Pyrsonympha/Trimastix* clade with 100% support in phylogenetic analyses with both parsimony and ML distance models (data not shown). Therefore, the presence of the "long-branch sequences" of *Pyrsonympha* and *Vanella* do not appear to obscure any relationships relevant to this study.

Oxymonad diversity and the *Pyrsonympha/Dinenympha* split

In order to examine the diversity of oxymonad sequences present in the environment that I sampled, the full double-strand sequence of eight clones was obtained. These showed unexpected diversity and resolved into five phylotypes (a phylotype being defined as a group of sequences which share 98% identity or better). One phylotype contained clones OS6, 8, 15 and 17, and then each of the others represented a unique phylotype. This diversity of sequences likely

corresponds to two different genera of pyrsonymphids. For a more detailed analysis of the internal topology in the oxymonad clade, an oxymonad and *Trimastix* specific dataset (23 taxa, 1,310 sites) was constructed. The *Pyrsonympha* sp. JD2000 sequence was aligned with the OS clones (clones OS1, 2, 6, 8, 13, 15, 17, 19) as well as pyrsonymphid ssu rDNA sequences from the Japanese subterranean termite *R. speratus* and *Oxymonas sp.* sequences from *N. koshunensis*. Phylogenetic analysis showed that the three identified oxymonad genera formed strongly supported clusters when *Trimastix* was used as the lone outgroup (Figure 4.5). Clones OS 6, 8, 13, 15 and 17 clones clustered with the identified *Pyrsonympha* sequences (74/82/70), while clones OS1, 2, and 19 went with *Dinenympha* ones (70/36/90), based on Maximum Likelihood, ML distance and parsimony values, respectively. The topology was further confirmed with a pyrsonymphid-specific phylogeny. To avoid unexpected long-branch effects and to obtain more resolution, the *Trimastix* and *Oxymonas* sequences were removed, yielding the Pyrsonymphid dataset. Analysis of this dataset (Figure 4.6) with 2,000 bootstrap replicates gave a strongly supported partition between the *Dinenympha* and *Pyrsonympha* clades (100/97 ML distance and parsimony values), with the same clones belonging to each group and very similar topology to the one in Figure 4.5. It should be noted that the extreme length of the branch rooting the pyrsonymphids, compared to the short internal nodes in the clade, makes finding a root difficult. A sequence which breaks up the long branch leading to the pyrsonymphid clade would allow a more accurate determination of where the root lies.

One final test was performed to confirm the internal topology of the pyrsonymphidae. I tested whether the clones cluster by genus (*Pyrsonympha* vs.

	ML	MLD	MP
A	74	82	70
B	70	36	90
C	100	100	100
D	96	98	100

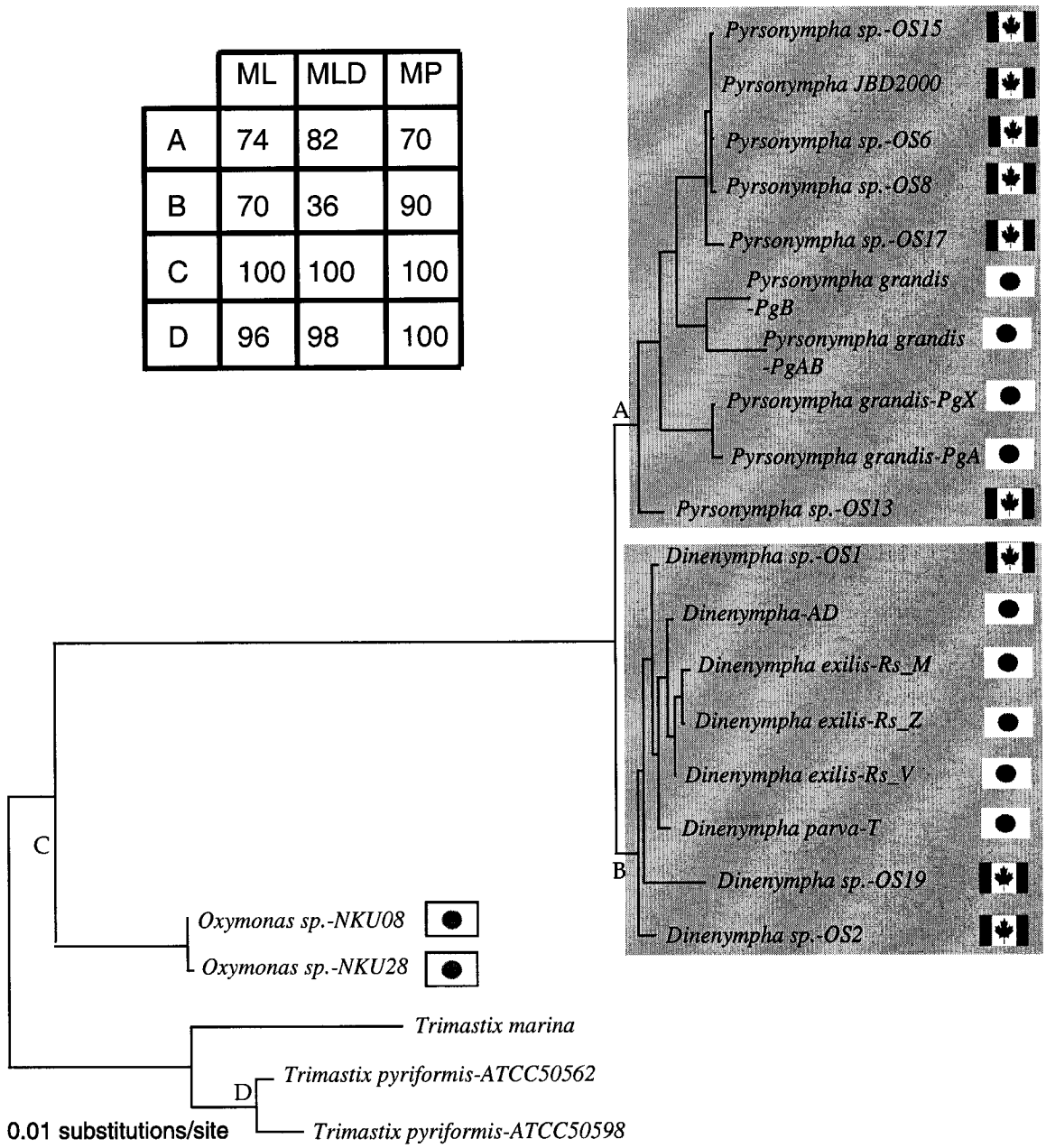
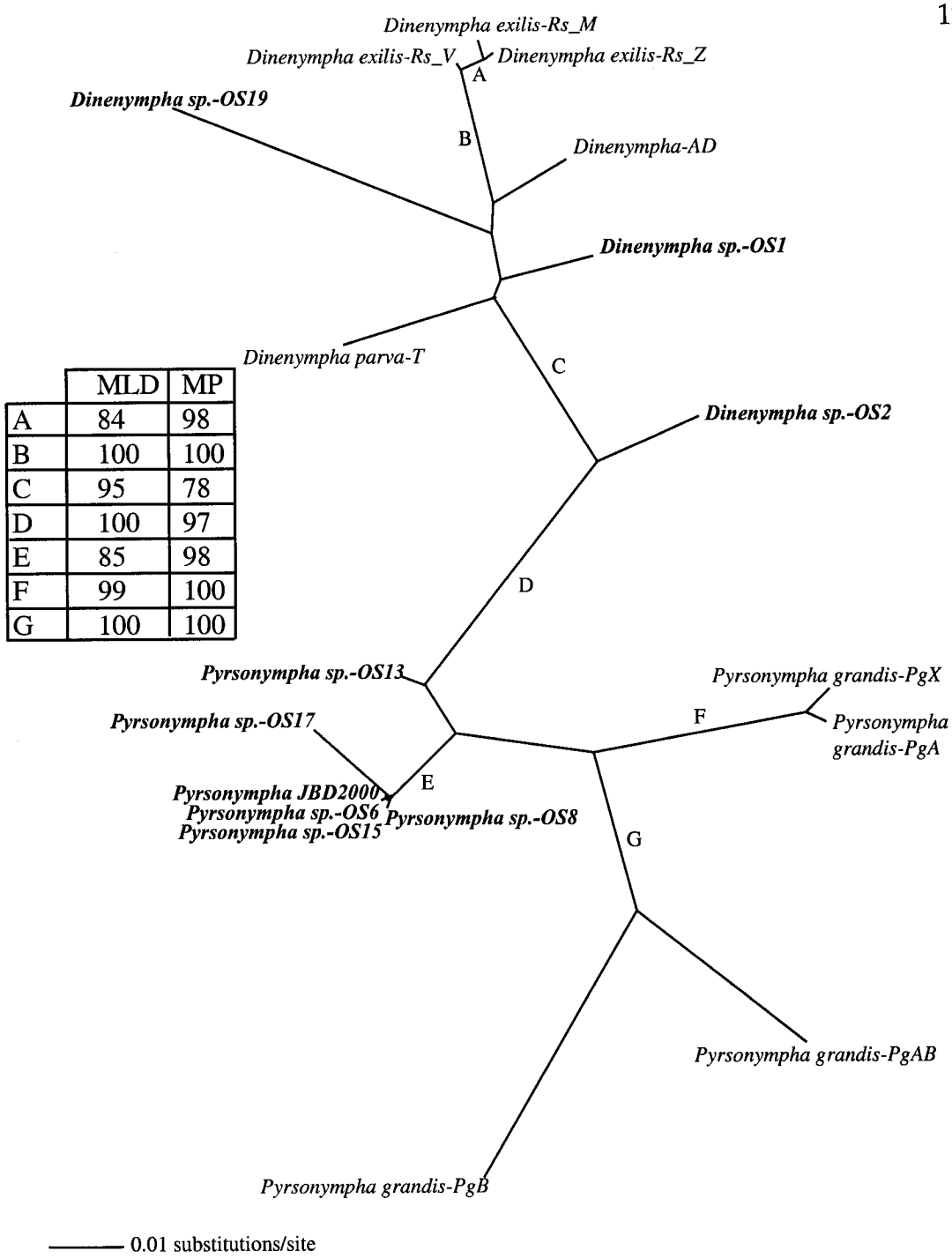


Figure 4.5: **Internal oxymonad phylogeny rooted by *Trimastix*.** The optimal maximum-likelihood (TrN+G+I) tree is shown with bootstrap values for full maximum likelihood, maximum likelihood distance and maximum parsimony analyses. Scale bar shows the number of changes per site. Shaded boxes surround the *Pyrsonympha* and *Dinenympha* clades. The sampling source for each sequence is denoted by either a Japanese or Canadian flag to its right.



	MLD	MP
A	84	98
B	100	100
C	95	78
D	100	97
E	85	98
F	99	100
G	100	100

Figure 4.6: **Pyrsonymphid-specific phylogeny.** This unrooted phylogeny shows the optimal maximum-likelihood tree with ML distance and maximum parsimony values at relevant bootstrap nodes. Note the long branch length of the internal node and the high support regardless of method at node D.

Dinenympha) or by sampling site (Canada vs. Japan). To evaluate these possibilities, we compared the two alternate tree topologies; ((Japanese *Pyrsonympha* clones, Canadian *Pyrsonympha* clones), (Japanese *Dinenympha* clones, Canadian *Dinenympha* clones)) versus ((Japanese *Pyrsonympha* clones, Japanese *Dinenympha* clones), (Canadian *Pyrsonympha* clones, Canadian *Dinenympha* clones)) by finding the best tree corresponding to the constraints above and evaluating using AU tests (Shimodaira 2002). The topology corresponding to clustering by sampling site was rejected with $p< 0.000$. Based on these analyses, clones OS 6, 8, 13, 15 and 17 were deemed to represent *Pyrsonympha* clones, while OS1, 2, and 19 were designated *Dinenympha* clones. The sequences are therefore named accordingly in Figures 4.5 and 4.6.

The *Trimastix*/oxymonad relationship in the light of the Excavate Hypothesis

Trimastix had originally been proposed, on morphological grounds, to be related to other excavate taxa including *Carpediemonas membranifera, Reclinomonas americana, Jakoba libera* and *Malawimonas jakobiformis* (Simpson and Patterson 1999; Simpson and Patterson 2001). The global phylogeny, shown in Figures 4.3 and 4.4, with only the *Pyrsonympha* sp. JD2000 sequence left open the possibility that these taxa, which had not been included in these analyses, might intervene between the oxymonads and *Trimastix*. To address this possibility a final dataset of 72 taxa, 907 sites was assembled. As seen in Figure 4.7, the oxymonads formed a monophyletic clade with 74/76% support from ML distance and parsimony, and the *Trimastix*/oxymonad clade was supported with 83/84%, excluding the other excavate taxa. Since the pyrsonymphid sequences clearly represent long branches, a further phylogeny was done using only the *Oxymonas sp.* sequence to

Figure 4.7: **Placement of oxymonads amongst eukaryotes (Global-Eukaryotes 3).** This phylogeny shows the best ML topology with ML distance and maximum parsimony bootstrap values at relevant nodes. The oxymonad-plus-*Trimastix* clade is boxed and the oxymonad sequences are bolded. Note the monophyly of the oxymonad and *Trimastix* sequences to the exclusion of the other excavate sequences (shown in courier font). Asterisks (*) and pound signs (#) are placed at nodes with better than 70% and 90% support, respectively, with both methods.

	MLD	MP
A	84	83
B	76	74
C	100	100

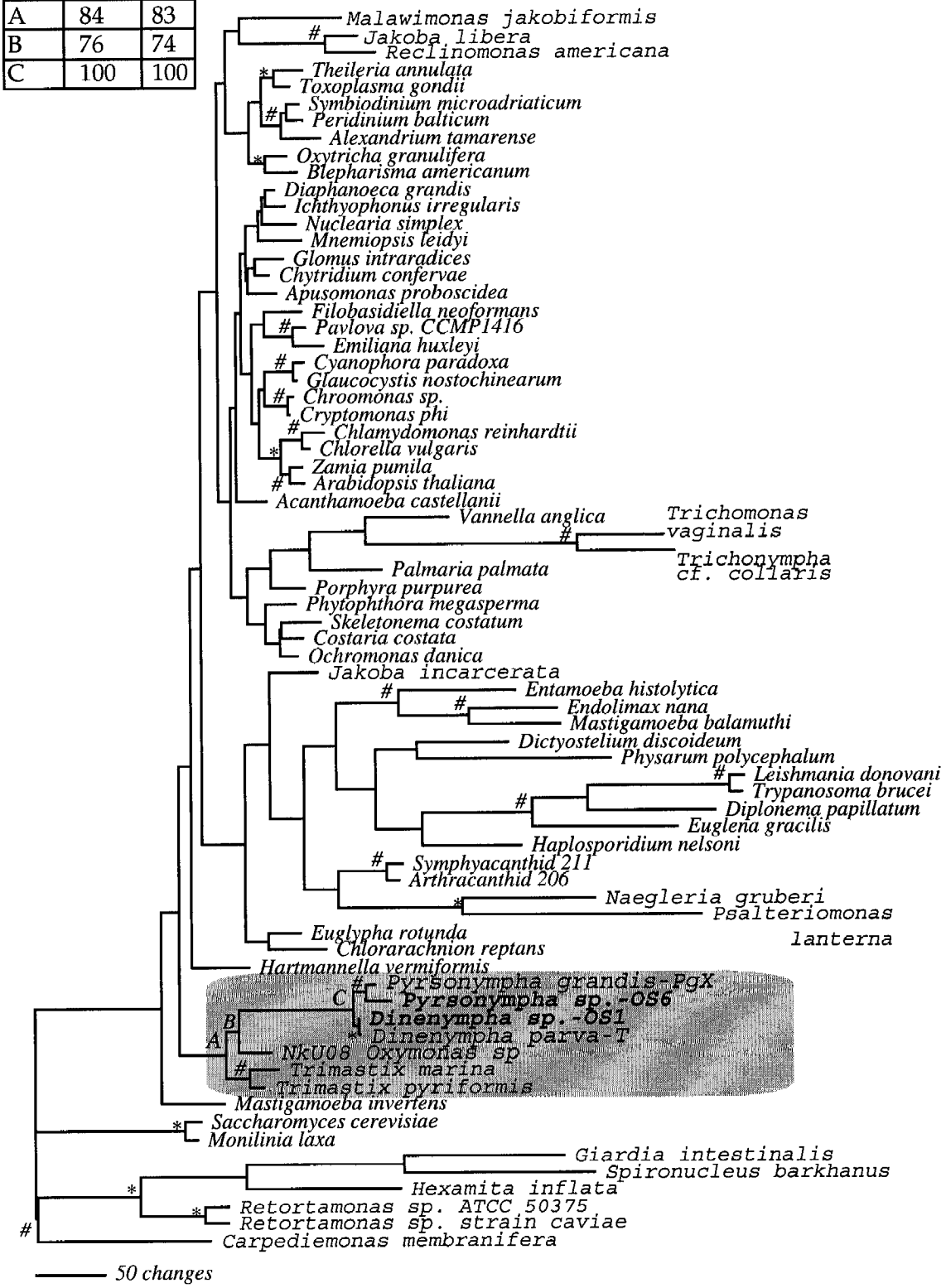


Figure 4.7: **Placement of oxymonads amongst eukaryotes**

(Global Eukaryotes 3)

represent the oxymonads (data not shown). In this case the support remained relatively constant for ML distance but jumped to 99% support in the parsimony analyses.

Discussion

In this chapter I have shown a specific relationship between the oxymonads and *Trimastix* to the exclusion of other eukaryotes. This result was robust under a variety of phylogenetic reconstruction methods and optimality criteria, and was confirmed by rigorous statistical tests. *Trimastix* are anaerobic free-living aquatic protozoa found living in benthic sediments (O'Kelly, Farmer et al. 1999). They are small (16-30 μm) and tetra-flagellated (Simpson, Bernard et al. 2000). *Trimastix* lacks a mitochondrion, although it does have a putative mitochondrial homologue (O'Kelly, Farmer et al. 1999; Simpson, Bernard et al. 2000), and at least one species of *Trimastix* has a well-recognized Golgi body (O'Kelly, Farmer et al. 1999). That oxymonads are likely to be secondarily 'Golgi-lacking' is the most salient deduction from the unification of oxymonads and *Trimastix*. However, this result has also allows the clarification of several outstanding points regarding oxymonad evolution.

Pyrsonymphid taxonomy

There is a long-standing controversy about oxymonad taxonomy dating back almost to the original description of the organisms at the end of the 19th century. Yamin, and others, treated *Pyrsonympha* and *Dinenympha* as separate genera (or subgenera in Koidzumi's case) in the family Pyrsonymphidae (Koidzumi 1921; Grassé 1952; Smith, Stamler et al. 1975; Yamin 1979). On the

other hand, Leidy, and Dubosc (based on morphological similarity) and Hollande (based on microscopic observation and DNA content measurements) considered *Dinenympha* to be a different life stage of *Pyrsonympha* (Leidy 1881; Duboscq and Grassé 1925; Hollande and Carruette-Valentin 1970). In results not shown here (Moriya, Dacks et al. In Press), fluorescence *in situ* microscopy was used to demonstrate that given ssu rDNA sequences are present in cells with a *Dinenympha* morphology, but not in *Pyrsonympha* cells, and *vice versa*. This exclusivity was also shown using sequences obtained from micro-manipulated cells, from which only one phylotype was obtained for each cell type. This finding discounted the possibility that the sequences were differentially expressed alleles. The various sequences obtained are clearly separated in my *Pyrsonympha* specific phylogeny (Figures 4.5 and 4.6) and in the AU tests. These data, in total, show that *Pyrsonympha* and *Dinenympha* are separate genera and not morphs of the same organism.

Evolution of mitochondria

In addition to the evolution of Golgi bodies, the Archezoa concept has been applied to other eukaryote-specific features, most importantly the evolution of mitochondria (Roger 1999). Oxymonads were among the prominent amitochondriate lineages named as primitively mitochondrion-lacking in the original Archezoa Hypothesis. The relationship of oxymonads with *Trimastix* now suggests that this condition is secondary, as *Trimastix* contains densely staining membrane-bound compartments that appear similar to the hydrogenosomes in other amitochondriate organisms (O'Kelly, Farmer et al.

1999; Roger 1999; Simpson, Bernard et al. 2000). The ancestor of oxymonads may have already been anaerobic and "pre-adapted" to the symbiotic lifestyle.

Implications for Broad-Scale Eukaryotic Systematics

The close relationship between oxymonads and *Trimastix* has important implications for phylum-level eukaryotic systematics. In the recent past, oxymonads have been considered to be allied with diplomonads and retortamonads in the phylum Metamonada (Cavalier-Smith 1981; Cavalier-Smith 1998) based on their common lack of mitochondria, or with Heterolobosea and Stephanopogon in the phylum Percolozoa (Cavalier-Smith 2000) based on their common lack of Golgi bodies. *Trimastix* has been allied with parabasalids in the phylum Trichozoa (Cavalier-Smith 1997) based on the shared presence of Golgi bodies and hydrogenosomes, or with jakobids in the phylum Loukozoa (Cavalier-Smith 1999) based on their common presence of the ventral feeding groove. However, the description of Metamonada and Percolozoa would not accommodate *Trimastix*, which has both mitochondrion-like organelles and Golgi dictyosomes, nor would the circumscriptions of Trichozoa and Loukozoa accommodate oxymonads, which lack Golgi dictyosomes and ventral feeding grooves. The phylogenies presented here, particularly the final large-scale phylogeny, show that oxymonads are the closest relatives of *Trimastix*, to the exclusion of other excavates. It may be most expedient to create a new taxon to encompass *Trimastix* and oxymonads. Given that each has generally been given its own class or subphylum (e.g., (Cavalier-Smith 1997; Cavalier-Smith 1998; Cavalier-Smith 1999; Cavalier-Smith 2000)), this new taxon would arguably deserve the rank of phylum. The name Preaxostyla has been suggested for the

group encompassing oxymonads and *Trimastix*, based on the similarity of the *Trimastix* I-fibre to the oxymonad pre-axostyle (A. G. B. Simpson, personal communication.

There is now evidence linking each potentially 'primitively Golgi-lacking' taxon to one possessing a Golgi apparatus: the diplomonads and retortamonads with *Carpediemonas* (Simpson, Roger et al. 2002); pelobionts and entamoebids with *Dictyostelium* (Arisue, Hashimoto et al. 2002; Bapteste, Brinkmann et al. 2002) and *Acanthamoeba* (Baldauf, Roger et al. 2000; Dacks, Marinets et al. 2002; Forget, Ustinova et al. 2002); microsporidia with fungi (Keeling and McFadden 1998); heteroloboseans with euglenoids (Baldauf, Roger et al. 2000); and oxymonads with *Trimastix*. These taxa are shown in red in Figure 4.8 and may have secondarily lost Golgi or, more likely, shifted their Golgi to an unrecognizable morphology.

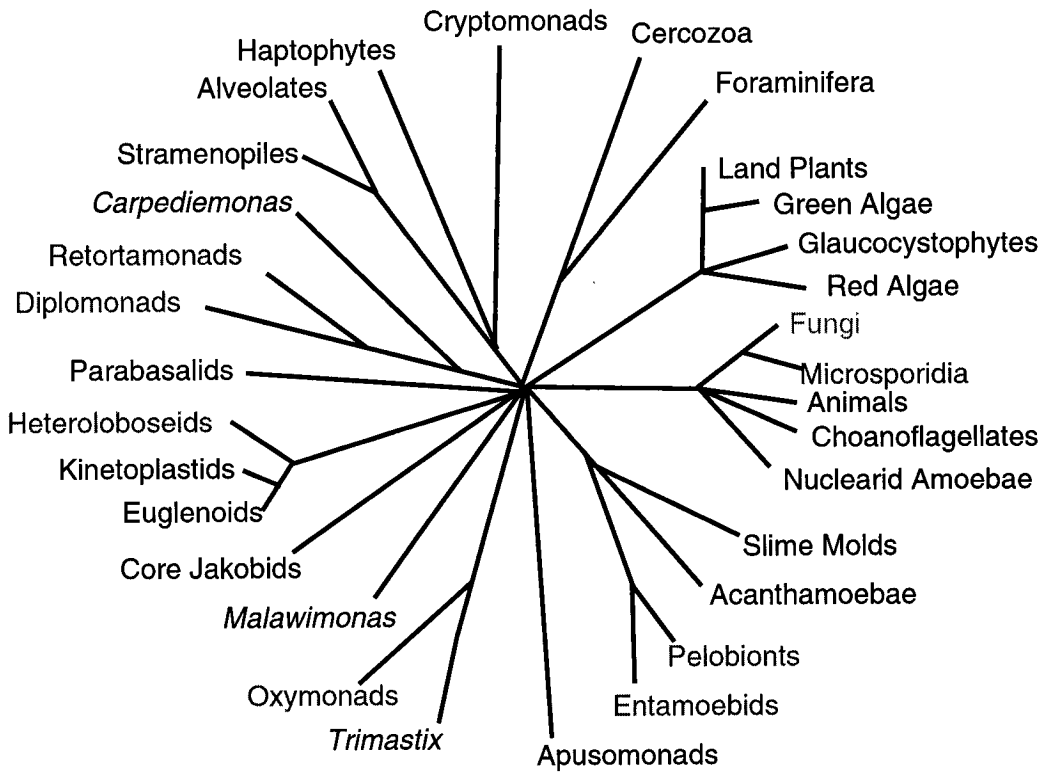


Figure 4.8: **Relationships of 'Golgi-lacking' and -possessing lineages, including oxymonads.** In previous versions of this figure the oxymonad/*Trimastix* relationship had been omitted. It is now included here. From this diagram, one can see that all major Golgi-lacking lineages have affiliations with ones possessing stacked Golgi bodies.

Chapter 5: Retromer and other genes from 'Golgi-lacking' taxa: Direct evidence of cryptic Golgi.

The conclusion that all major 'Golgi-lacking' taxa lack Golgi secondarily, as based on the indirect phylogenetic evidence presented in Chapter 4, is unfortunately vulnerable on two levels. Although the proposed relationships (Figure 5.1) seem robust, if one of them is incorrect then the resulting deduction about the secondarily 'Golgi-lacking' nature of that lineage is false as well. A more serious objection is the placement of the root of the eukaryotic tree. Even if the phylogenetic hypotheses in Figure 5.1 are all correct, rooting the tree of eukaryotes on any of the 'Golgi-lacking' lineages would parsimoniously imply its primary lack of the organelle. Several of the competing placements (Sogin 1991; Lang, Burger et al. 1997; Simpson and Patterson 1999; Keeling and Palmer 2000; Stechmann and Cavalier-Smith 2002) for the root of eukaryotes (Figure 5.1) place 'Golgi-lacking' groups as basally diverging.

In the absence of a rooted phylogeny of eukaryotes a second, more direct, kind of evidence is needed to suggest the cryptic presence or secondary loss of Golgi bodies in these lineages. This may come as biochemical evidence such as sensitivity to drugs (Brefeldin A) (Lujan, Marotta et al. 1995), or as molecular biological evidence. The presence of genes whose products are known to be strictly involved with Golgi function in model systems may be taken as evidence for the presence of the organelle in a 'Golgi-lacking' taxon. Genes of this type are at times referred to as 'Golgi-associated genes' in this thesis. This designation in no way implies that I think that the genes themselves are physically associated with the organelle. The rationale of 'Golgi-associated genes'

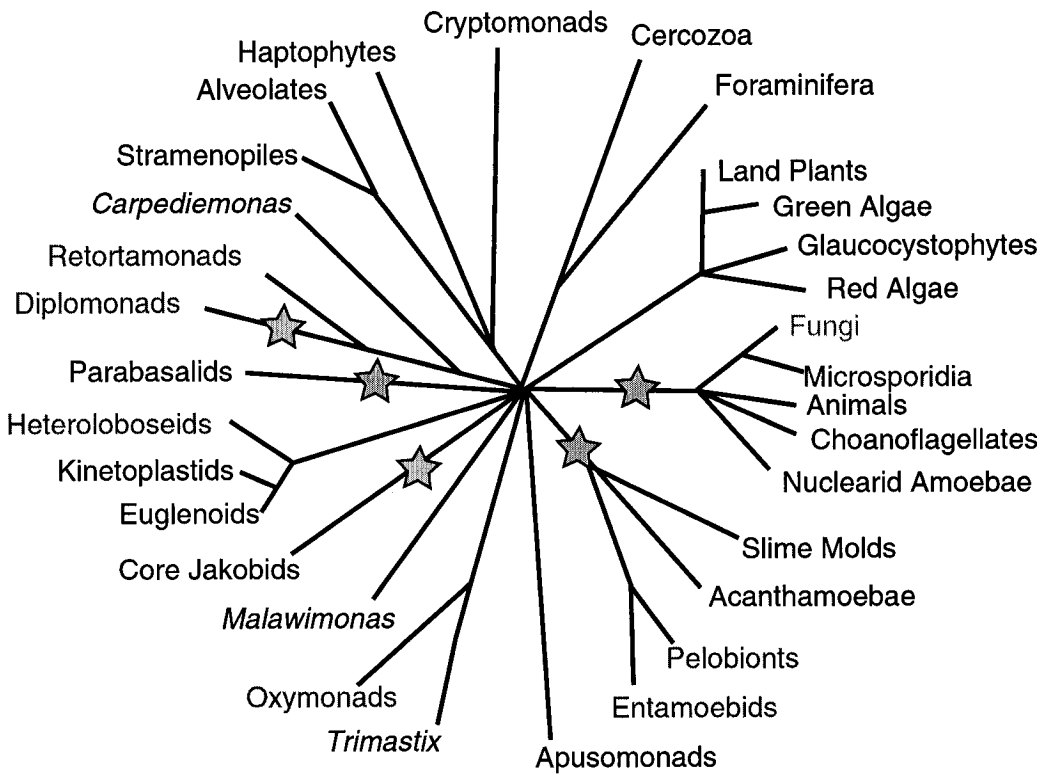


Figure 5.1: **Eukaryotic relationships with Golgi-lacking taxa and potential roots.** This illustration is identical to Figure 4.8, but now shows possible root placements denoted by yellow stars. Some of these would render the indirect inference of secondary Golgi body loss in Golgi-lacking taxa invalid.

being evidence for the presence of a Golgi body is particularly true if the product is involved in mechanistic function of the organelle (i.e. vesicular transport), since the only reason for the gene to be present in the genome is if the organelle is also present.

Many of the genes involved in Golgi function were, in fact, identified and characterized in the yeast *Saccharomyces cereviseae*, which lacks classically stacked Golgi apparatus (Jahn and Sudhof 1999). Some such genes may be components of Golgi-specific complexes such as the coatomer or retromer complex. Proving the homology of such gene sequences can be done with BLAST comparisons (Altschul, Madden et al. 1997) and looking for the presence of conserved signature motifs. Other Golgi-associated genes may be organelle-specific paralogues of gene families involved in vesicular transport such as the adaptins (Robinson and Bonifacino 2001) or syntaxins (Bennett, Garcia-Arraras et al. 1993). These require phylogenetic analysis to confirm that the sequence obtained does, in fact, belong to the "Golgi-specific" paralogue.

Biology of the retromer complex

As it is not part of the generalized model of vesicular transport, the retromer complex was not described in Chapter 1. Given the large role that it will play in this chapter, however, a small overview of the retromer complex is warranted.

Continuous anterograde movement of proteins from the Golgi body to the lysosome requires that cargo receptors for those proteins be recycled back to the Golgi apparatus for further rounds of transport. Vps10 acts as that cargo receptor in yeast, and its recycling is mediated by a complex of five proteins collectively

referred to as the retromer complex. This complex is also responsible for recycling the Kex2 protease and dipeptidyl aminopeptidase-A. Retromer components Vps29 and Vps35 are thought to select cargo at the endosome through interactions with the Vps10 (Seaman, McCaffery et al. 1998). Vps5 and Vps17 are then proposed to form the vesicle coat with Vps26 promoting interaction, at least between Vps35 and the Vps5/Vps17 coat (Seaman, McCaffery et al. 1998; Reddy and Seaman 2001). Mammalian homologues have been identified for Vps 26, Vps29 and Vps35 and have been shown to form multimeric complexes, suggesting that the retromer plays a similar role in mammalian cells as in fungal ones (Haft, de la Luz Sierra et al. 2000). Providing that the retromer is present in eukaryotic taxa other than just animals and fungi, it may be a useful marker for the presence of Golgi bodies in taxa proposed to primitively lack Golgi bodies, due to the retromer's specific role in Golgi function and its lack of paralogues that act at other membrane organelles.

In this chapter, genes encoding components of the retromer complex are characterized from diverse eukaryotes, including putatively 'Golgi-lacking' taxa. Additionally, a number of genes were obtained whose products are homologous to vesicular-transport components with well-characterized Golgi function in model systems.

Materials and Methods

DNA

Entamoeba histolytica strain HM1:IMSS, and some *Giardia intestinalis* DNA, were generous gifts from Paul Hoffman and Gary Sisson. *Giardia intestinalis*

strain WB and *Mastigamoeba balamuthi* DNA were generous gifts from Andrew Roger and Lesley Davis. *Hartmanella vermiformis* was kindly provided by Mike Gray and Amanda Lohan.

Amplification of retromer genes

Vps26

Degenerate PCR was used to obtain fragments of the Vps26 gene from various taxa. The sequences of all named primers are given in Table 5.1 Amplification of the *Hartmanella vermiformis* Vps26 gene sequence was performed with primer combinations VPS26F1-VPS26R1, VPS26F2-VPS26R1, and VPS26F3-VPS26R2. Fragments of the *Trypanosoma brucei* Vps26 gene sequence were obtained using primer combinations VPS26F3-VPS26R2 and VPS26F2-VPS26R1, while a portion of the *Mastigmoeba balamuthi* Vps26 gene sequence was also amplified using this last primer set. For all amplifications, PCR amplification began with 1 minute at 95°C, followed by 40 cycles of 94°C, 50°C and 72°C each for 1 minute. The amplification ended with a cycle of 94°C for 1 minute, 50°C for 1 minute and 72°C for 5 minutes.

Using exact-match primers designed from GSS or genome sequence, the genes encoding Vps26 were amplified from *Entamoeba histolytica* and *Giardia intestinalis*, with primer combinations EHvps26xf1-EHvps26xr2 and GIvps26xf1-GIvps26xr1, respectively. Amplification began with a denaturing step at 95°C for 1 minute followed by 40 cycles of 94°C for 30 s, 50°C for 1 minute and 72°C for 2 minutes, and concluded with 5 minutes at 72°C to complete any fragments obtained.

Name	Sequence (5'-3')	AA motif	positions
	Vps26 PCR primers-Degenerate		
VPS26F1	GGA GGT CGG CRT CGA RRA STG	MEVGIEDC	166-173
VPS26F2	GAG AAG TCG CTG CAY ATH GAR TT	EDCLHIEF	171-178
VPS26F3	AGG TAC GAG HTH ATG GAY GG	RYEIMDG	232-238
VPS26R1	TTG ACG GGG CYN CCR TCC AT	MDGAPVK	236-242
VPS26R2	AGC GTG ATC TCC TGY TGY TTR AA	FKQQEITL	287-294
	Vps35 PCR primers-Degenerate		
VPS35F3	AAG GAT CTG GTN GAK ATG TG	KDLVEMC	131-137
VPS35R3	GCT CGT CTG GRA ANA CYT G	QVFPDE	268-273
VPS35R4	CGG CAC TGG TCN GGY TTY TT	KKPDQCR	662-668

A

Name	Sequence (5'-3')	Organism
	Vps26 PCR primers-exact match	
EHvps26xf1	GGA ACT TCA AAA TAG AAG AAT GGC	*E. histolytica*
EHvps26xr2	GTT GAG GTT GAA CTT CTG G	*E. histolytica*
GIvps26xf1	CTG CAG ACA CTG TCA TTG C	*G. intestinalis*
GIvps26xr1	GAG CGG CTT CAG TTT CCA TC	*G. intestinalis*
	Vps35 PCR primers-exact match	
VPSRAXF1	GAC ATT CTC AAG GAT CTA GCC	*R. americana*
VPSRAXF2	TGT TCC TGC GCA ACT ATC TGC	*R. americana*
VPSTBXF1	GGC GAC GTA AAT ACC CAA GG	*T. brucei*
VPSTBXR2	TAT CAC GCG GAT AAT GAG ACG	*T. brucei*
VPSTBXR3	GGT GGA TGA ATA TGT GTC TCC	*T. brucei*
VPSGMXF1	AAG CGG TAC TGT TCG TGC TG	*G. intestinalis*
VPSGMXR1	CCA ACG AAA GAC GGT CTA GC	*G. intestinalis*
UGVPSGMXF1	TAT GCA CAA TAG ATG GTC TCC AGG	*G. intestinalis*
UGVPSGMXF2	ACC CTA TGT CCA CCG CAA CTA TTC TGG	*G. intestinalis*
LWGVPS351B	B-GAG ATT CTG TGA AGC AGG AAC	*G. intestinalis*
LWGVPS352	TGC AAT ACG GTA CTT CAG GAG	*G. intestinalis*
5'EHVPS35-1-XF1	GTT CTC TTT CCT TAC AAG GG	*E. histolytica*
5'EHVPS35-1-XR1	GCT GCA ATG ACA CTT CTC	*E. histolytica*
EHWVPS35XF2	TAT TCA ATA CCC AGA GAT TAT C	*E. histolytica*
3'EHVPS35-I-XR4	GCC AAT GCT TAT AAA AGA TG	*E. histolytica*
RAMVPS35FULL5X2	GAG GAT CTG TTC GAG ATT GCG	*R. americana*
RAMVPS35FULL3X1	AGC AGC TTG ACA GAG TAC TGG	*R. americana*
RAMVPS35FULL3X2	GAC AGC GTG TCG TAG TTG TCC	*R. americana*

B

Table 5.1: **Primers used to obtain various retromer component genes.**

(A) Degenerate primers for Vps26 and 35. Primers are listed by name, nucleotide sequence, amino acid sequence to which the primer was designed, and position corresponding to the primer site. (B) Exact-match primers for Vps26 and 35. Primers are listed by name, nucleotide sequence, and organism from which the sequence was derived.

Vps35

The *Hartmanella vermiformis* Vps35 gene sequence was amplified using a combination of the degenerate primers VPS35F3-VPS35R4. Amplification began with 95°C for 1 minute, 55°C for 30 s and 68°C for 5 minutes. This was followed by 40 cycles of 94°C for 20 s, 55°C for 30 s and 68°C for 5 minutes. The amplification was concluded with 5 minutes at 68°C.

A 5′ fragment (431 nts) of the *Reclinomonas americana* Vps35 gene was serendipitously obtained and generously made available by Yuji Inagaki. After sequencing of the gene fragment, exact-match primers were designed and used in combination with degenerate primers VPS35R3, and VPS35R4. Amplification used identical parameters as in the amplification of *Hartmanella vermiformis* Vps35, but with 47°C as the annealing temperature. This successfully amplified both 500-nt (VPSRAXF2-VPS35R3) and 2300-nt (VPSRAXF2-VPS35R4) fragments of the gene. Since the assembly of the gene indicated that various versions of the gene were present, a final PCR was done using primers that spanned the region obtained from the previous experiments. Primer combinations RamVps35full5x2-RamVps35full3x1 and RamVps35full5x2- RamVps35full3x2 were used with the same parameters as in the amplification of the *H. vermiformis* Vps 35 gene.

Two fragments of the *Trypanosoma brucei* Vps35 gene, corresponding to middle and 3′ regions, were identified from GSS fragments. These regions, plus the intervening sequence, were amplified with the exact-match primers VPSTBXF1- VPSTBXR3, using the parameters described for the PCR of the *H. vermiformis* Vps26 fragments. Once sequenced, the obtained fragment was used to design exact-match primer VPSTBXR2 which, in combination with primer VPS35F3, was used to amplify nearly the entire remaining coding sequence of

the gene. Identical parameters as in the amplification of the *Hartmanella vermiformis* Vps35 gene amplification were used, but with 47°C as the annealing temperature.

A 5′ fragment of the *Giardia intestinalis* Vps35 gene was identified in the single-pass reads of the *Giardia* genome. Exact-match primers (VPSGMXF1-VPSGMXR1) were used to amplify the fragment under the same PCR conditions as described in the amplification of the *H. vermiformis* Vps26 fragments. To obtain the remaining 3′ portion of the gene, uneven PCR was attempted using primers UGVPSGMXF1 and UGVPSGMXF2, which provided an additional 500 nucleotides. The rest of the ORF was finally obtained through Long Walk PCR with primers LWGVPS351B and LWGVPS352 in combination with the random primers and the cycling parameters described in the original protocol (Katz, Curtis et al. 2000).

Several fragments encoding Vps35 genes were also identified from *Entamoeba histolytica* GSS sequences. A portion of the *E. histolytica* Vps35-1 gene was amplified using primers 5'EHVPS35-1-XF1 and 5'EHVPS35-1-XR1. The same cycling parameters were used as in the amplification of the *M. balamuthi* Vps26 gene, but with annealing at 49°C. As this fragment was 1360 nucleotides in length (1031 nucleotides of the ORF) and covered the most conserved region of the gene, the entire ORF was not characterized. A second Vps35 gene was also characterized (*Entamoeba histolytica* Vps35-2). A fragment spanning nearly the entire ORF was amplified using primers EHWVPS35XF2 and 3'EHVPS35-I-XR4. This yielded a fragment 2150 nts long that clearly overlapped the 3′ end of the gene and was missing only approximately 50 conceptually translated amino acids from the N-termini of other Vps35 homologues. Amplification began with

an initial denaturation at 95°C for 2 minutes. This was followed by 40 cycles of melting at 92°C for 1.5 minutes, annealing at 50°C for 1.5 minutes and extension at 72°C for 2 minutes with an additional 6 s per cycle added to the extension time. The reaction was concluded with 7 minutes at 72°C.

Cloning

All genes obtained by PCR were ligated into pCR Topo2.1 (InVitrogen) and transformed into *E. coli* Top10 cells.

In addition, several genes were obtained through identifying relevant clones from various GSS or EST projects, obtaining these clones and fully characterizing the ORF. These included clones encoding genes for: an Adaptin subunit from *Mastigamoeba balamuthi* (Clone MA1267), a Coatomer subunit from *Naegleria gruberi* (gNgTorT7-183), Vps26 from *Spironucleus barkhanus* (gSp-00270), Vps29 from *Trichomonas vaginalis* (TV1509) and Vps35 from both *Dictyostelium discoideum* (SLJ872) and *Chlamydomonas reinhartdii* (CL57h02). Clones provided as plasmids were transformed into *E. coli* Top10 cells.

Sequencing

Double-strand sequence of most sequences was obtained for at least two clones in both directions on an ABI 377, with the exception of the following. The B-COP sequence from *S. barkhanus* was obtained through a contiguation of GSS reads generously donated by Andew Roger. The *Chlamydomonas reinhartdii* Vps35 sequence, although part of what appears to be a full-sized cDNA clone of the gene, was characterized only at the 5′ end of the gene. This is because, despite multiple technically varied attempts, I was unable to sequence through the

middle region of the gene. The GSS clone containing the *S. barkhanus* Vps26 gene was not sequenced on both strands over its entire length. The entire portion of the Vps26 ORF contained on the clone, however, was fully sequenced. The *Reclinomonas americana* Vps35 sequence was assembled from various clones. However, the sequence of each allele was assembled based on reads on both strands and using 12-to 18-fold coverage.

BLAST analysis

In all cases the BLAST score shown was obtained by using the conceptually translated amino acid sequence of the gene of interest as a query for a BLASTp search (Altschul, Madden et al. 1997). This was performed at the BLAST site at NCBI (http://www.ncbi.nlm.nih.gov/BLAST/) using default settings. As well, a comparative genomics survey of retromer components was performed as described in the Methods section of Chapter 2, with the named human representatives for Vps26, 29 and 35 used as the query sequences.

Alignment

For the various alignments in this chapter, all genes, with the exception of those obtained herein, were collected from Genbank using BLASTp at the nr database and retaining sequences that were retrieved with E values better than 0.05. In each case the collected sequences were aligned with the amino acid translations of obtained gene fragments using Clustal X at default parameters, and manually adjusted.

Vps26

All Vps26 homologues retrieved by BLAST were aligned in an initial dataset with 25 taxa and 282 sites. A second dataset, with clear lineage-specific duplicates removed, was constructed consisting of 19 taxa and 120 sites.

Vps29

Initial alignments were constructed by retrieving all sequences that had an E value of better than 0.05 in a BLASTp search using the *Trichomonas vaginalis* Vps29 deduced amino acid sequence as a query, as well as diverse examples of prokaryotic sequences. A 29-taxon data set with 126 sites was assembled, to test the monophyly of eukaryotic sequences compared to diverse prokaryotic ones. A final dataset of 14 taxa and 126 sites was assembled to resolve relations within the eukaryotic homologues.

Vps35

Three datasets were constructed for the analysis of Vps35 phylogeny. The first contained sequences from as diverse a range of taxa as possible, with 22 taxa and 387 positions. The second dataset contained only full-length sequences consisting of 15 taxa and 272 sites. The final dataset with long-branch taxa removed held 10 taxa and 273 sites.

Adaptin sigma

Protein sequences of the four adaptin-sigma paralogue families were aligned in a dataset containing 22 taxa and 119 sites.

B'-COP

Protein sequences of Beta-prime coatomer subunits, plus all F-Box proteins retrieved, were aligned into a final alignment of 20 taxa and 199 sites.

Phylogenetic analysis

Tree-Puzzle 4.0 (Strimmer and von Haeseler 1997) was used to calculate quartet puzzling support values as well as to estimate number of invariant sites and among-site-rate variation categories under a gamma model. These were then used in ML distance analyses using Puzzleboot (www.tree-puzzle.de) and full ML analyses using ProML (Felsenstein 1995). However full ML analysis was only performed on the final dataset for each protein. ProtML 2.2 (Adachi and Hasegawa 1996) with a q1000 search for each dataset was also performed in some cases to determine ML results. Resampling Estimated Log Likelihood values (RELLs) were calculated using ProtML 2.2 in conjunction with Mol2con (A. Stoltzfus, personal communication). Analyses incorporated a JTT amino acid substitution matrix and support values are based on 100 bootstrap replicates with global rearrangements and 3X jumbling when applicable.

Intron detection

The detection and in-silico splicing of putative introns was performed as described in Chapter 3.

Results

Retromer genes from taxa beyond animals and fungi

Functional characterization of retromer genes has been done in animal or fungal models. Before using retromer genes as evidence for the presence of a Golgi organelle, it is necessary to demonstrate that the retromer complex is, in fact, a widespread eukaryotic feature and not strictly an opisthokont novelty.

Degenerate PCR was used to amplify two fragments of the *Hartmanella vermiformis* Vps26 gene. These are 187 and 145 nts respectively, and overlap the same priming site. Although, when the priming sequence is excluded, there is a three amino acid gap in the sequence, these amino acids are conserved as MD/EG in all known Vps26 sequences. No introns were observable in the fragments, nor was there any evidence of variation that would suggest the presence of alleles. Similar regions were amplified from the *Trypanosoma brucei* Vps26 gene counting 178 and 149 nucleotides respectively. The conceptually translated protein sequences of both of these genes (136 amino acids from *H. vermiformis* and 125 amino acids for *T. brucei*) are clear Vps26 homologues (Table 5.2 and Figure 5.2A).

Figure 5.2: **Conserved regions of various Golgi-associated genes.** The taxa from which I obtained sequence are bolded. All numbering is according to the *Homo sapiens* protein sequence. Shaded regions indicate identity with the top line. (A) Aligned positions 179-207 of Vps26. (B) Aligned positions 131-153 of Vps35. The functionally critical, completely conserved, Asp123 position is denoted with a star. (C) Aligned positions 108-122 of B-COP. (D) Positions 207-222 and 243-253 of the B′-COP, with the ~~~ denoting the break in the alignment.

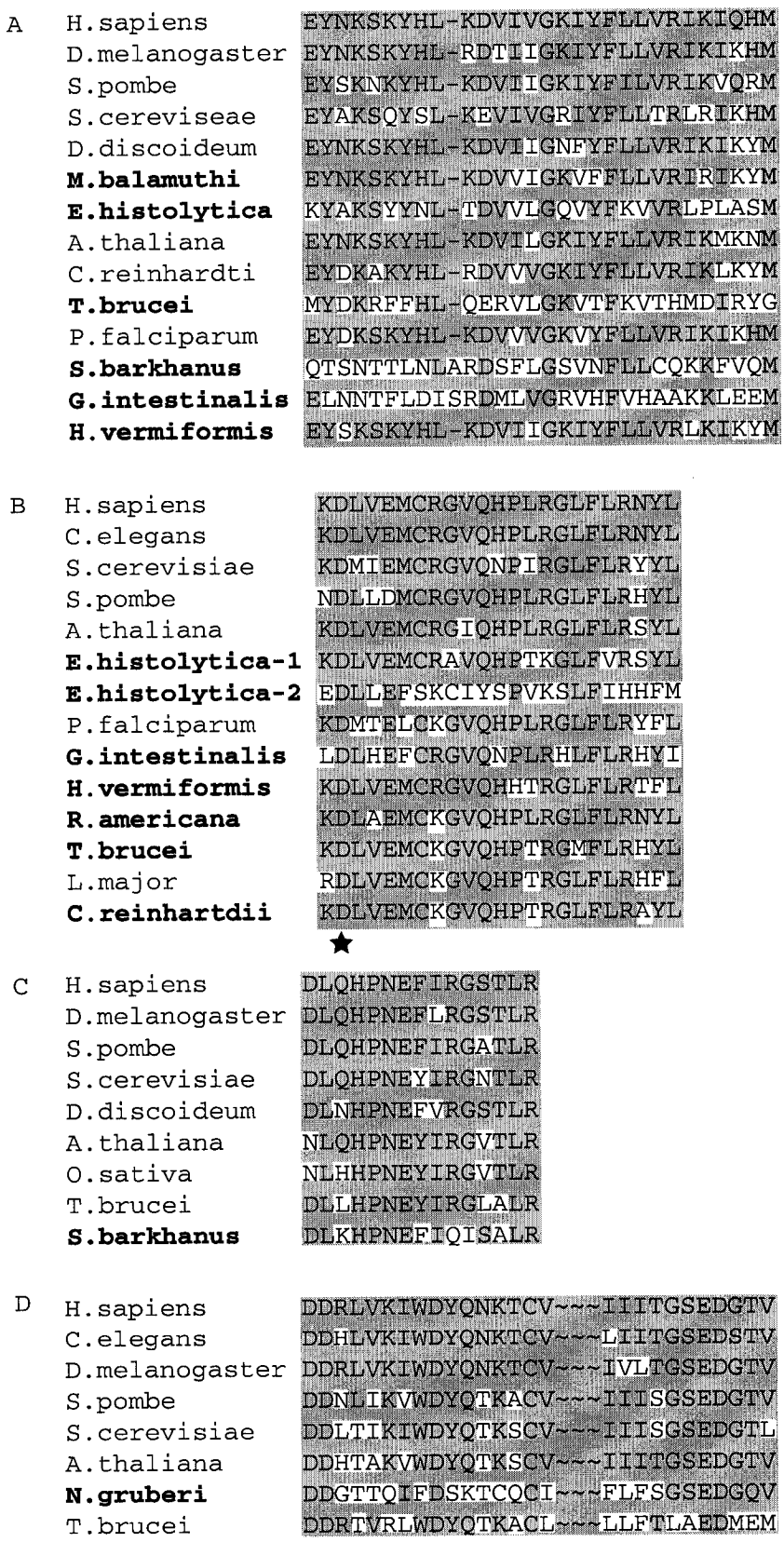


Figure 5.2: **Conserved regions of various Golgi-associated genes**

Accession #	Higher taxon	Organism	Assignment	Size	Top BLAST hit	Evalue	Top alternative hit	E value
#########	Amoeba	*H. vermiformis*	Vps 26	136AA	PepA-*D. discoideum*	3E-50	ilvI-*N.meningitidis*	3.5
#########	Kinetoplastid	*T. brucei*	Vps 26	125AA	VPS26-*D. melanogaster*	2E-20	AdeC-*M.kandleri*	4.3
AY193839	Entamoebids	*E. histolytica*	Vps 26	328AA	PepA-*D. discoideum*	1E-69	MepA-*P.multocida*	1.4
AY193840	Diplomonad	*G. intestinalis*	Vps 26	460AA	Vps-like/H-B-58-like-*A.thaliana*	1E-23	Unk. prot.-*Nostoc* sp.	1.2
AY193846	Diplomonad	*S. barkhanus*	Vps 26	214AA	VPS26-*D. melanogaster*	1E-09	Put prot.-*H.pylori*	2.3
AY193843	Pelobionts	*M. balamuthi*	Vps 26	57AA	PepA-*D. discoideum*	5E-18	None	N/A
#########	parabasalids	*T.vaginalis*	Vps 29	185AA	Vps29-*M.musculus*	5E-40	**Con.Hyp.-*A.fulgidus***	**1.0E-09**
#########	Amoeba	*H.vermiformis*	Vps 35	552AA	vps 35-*H. sapiens*	5E-93	argE-*Moritella* sp.	5.7
#########	Jakobids	*R. americana*	Vps 35	580AA	(unnamed),vps 35-*H. sapiens*	1E-93	rhoptry prot.-*P. yoelii*	6.4
#########	Kinetoplastid	*T. brucei*	Vps 35	788AA	possible vac. sort. prot.-*L. major*	1E-112	PKS1-*C. heterostrophus*	0.84
#########	Chlorophyte	*C. reinhartdii*	Vps 35	386AA	put. vac sort. prot. 35 -*A.thaliana*	7E-91	Hsp70-*T. rubripes*	1.9
#########	Slime mold	*D. discoideum*	Vps 35	453AA	(unnamed), vps35-*H. sapiens*	7E-56	laminin A-*M.musculus*	2
AY193838	Entamoebids	*E. histolytica*	Vps 35-1	343AA	Put.Vps35-*A. thaliana*	2E-47	C34D4.14.p-*C.elegans*	4.3
#########	Entamoebids	*E. histolytica*	Vps 35-2	655AA	Vps35-*H.sapiens*	2E-06	LOC233962-*M.musculus*	1.7
AY193841	Diplomonad	*G. intestinalis*	Vps 35	765AA	(CG5625), vps35-*M. musculus*	2E-16	None	N/A
AY193842	Pelobionts	*M. balamuthi*	AP 3-sigma	161AA	AP3-*M. musculus*	3E-53	**AP2-*S.pombe***	**1.0E-23**
AY193844	Heterolobosea	*N. gruberi*	B'-COP	267AA	B'-COP-*S.cerevisiae*	8E-13	**FBW7-*H.sapiens***	**2.0E-04**
AY193845	Diplomonad	*S. barkhanus*	B-COP	326AA	B-COP-*S.cerevisiae*	4E-13	AP-Beta-***B.taurus***	0.1

Table 5.2: **Golgi-associated genes obtained.** This table lists the genes obtained by Genbank accession number (if they have already been submitted), organismal source, proposed gene assignment, top BLAST hit, BLAST E value, top alternate hit and its BLAST E-value. N/A denotes a situation where the BLAST search retrieved no sequences other than homologues belonging to the proposed gene assignment. Bolded E values show alternate hits that had significant scores. In each of these cases the BLAST query was then confirmed in its assignment by phylogenetic analysis.

Fragments were amplified from the *Hartmanella vermiformis* Vps35 gene that corresponded to three slightly different versions of the Vps35 gene. The three clones sequenced were 2104, 2104 and 2101 nucleotides in length. At least 8 traditional GT-AG introns were detected in the amplified fragments. These introns ranged in size from 45 to 77 nucleotides. Since the amplified region does not span the entire ORF, other introns may also be present in the gene. When conceptually spliced, all three sequences encode proteins 552 amino acids long. Of the three sequences, *Hartmanella vermiformis* Vps26-5.1 and 5.2 (11 changes) are clearly closer to each other than either is to *Hartmanella vermiformis* Vps26-5.6 (32 and 28 changes, respectively, between 5.6 and 5.1 or 5.2). While the three sequences always retained the same intron boundaries and the majority of changes between the genes are either silent or within introns, there are a number of interesting differences. Intron 4 is actually two nucleotides shorter in 5.6 than in the other two. As well, there are one and three conservative amino acid changes, and two non-conservative changes observed between the sequences. Whether these sequences represent alleles in a population or copies within a genome is unclear and not critical to the larger question at hand. Given the extreme similarity of the alleles, the sequences were represented by a single allele in most subsequent analyses. BLAST analysis and conserved amino acid motifs demonstrate that the genes code for a Vps35 homologue (Table 5.2 and Figure 5.2B).

Fragments encoding the Vps35 gene were amplified from *Reclinomonas americana*. These yielded clones that resolved broadly into two classes, and so two consensus sequences were assembled, *Reclinomonas americana* Vps35-1 (2391

nts) and *Reclinomonas americana* Vps35-2 (2442 nts). The two sequences are 94% identical over their entire lengths. In each sequence, six traditional GT-AG introns could be detected, ranging in size from 58 to 151 nucleotides. Of these introns, five of the six are shared between the alleles. The conceptual translation of the sequences yielded a 580 amino acid protein which shows significant E values to other Vps35 orthologues (Table 5.2 and Figure 5.2B).

Nearly the full *Trypanosoma brucei* Vps35 gene sequence (2395 nts) was obtained, yielding a conceptually translated protein 788 amino acids long. No introns were detectable, nor was there any evidence of alleles. BLAST analysis and conserved sequence motifs show this gene clearly to encode a Vps35 homologue (Table 5.2 and Figure 5.2B).

Partial sequence of Vps35 genes from *Chlamydomonas reinhartdii* and *Dictyostelium discoideum* were obtained by sequencing cDNA clones. From the 5′ end of the *C. reinhartdii* gene, 1559 nucleotides of sequence yielded a protein 386 amino acids long that strongly retrieved Vps35 homologues (Table 5.2). Another 1016 nucleotides from the other end of the clone was sequenced but failed to retrieve any significant hits, indicating that the ORF had not yet been reached. The *D. discoideum* clone was 1368 nucleotides and contained the 3′ half of the gene. Nonetheless, this retrieved Vps35 homologues with significant E values (Table 5.2). The full sequence of a Vps29 gene (667 nucleotides, corresponding to an ORF encoding 185 amino acids) from *Trichomonas vaginalis* was also obtained from a cDNA. This also showed significant E values to Vps29 homologues in a BLAST search (Table 5.2).

Beyond the molecular biological investigation, a comparative genomic search was performed to determine the presence of retromer components in

Lineage	Organism	Vps26	Vps29	Vps35
Fungi	*S. cereviseae*	A	A	A
Land Plants	*A. thaliana*	A	C	B
Animals	*H. sapiens*	A	A	A
Kinetoplastid	*T. brucei*	F	E	F
Apicomplexa	*P. falciparum*	C	C	C
Slime molds	*D. discoideum*	A	NI	D
Red Algae	*P. yezoensis*	NI	NI	NI
Stramenopiles	*P. sojae*	NI	NI	NI
Green Algae	*C. reinhartdii*	E	E	D
Ciliates	*P. tetraurelia*	A	NI	NI

Table 5.3: **Comparative genomics survey of retromer components.** A = homologues published in separate analyses. B = homologues identified in the September 2001 search (Dacks and Doolittle 2001). C = genes not yet published but found in Genbank. D = genes listed on the respective genome initiative website. E shows when a homologue was found by reciprocal BLAST analysis. NI = a clear homologue was not reliably identified by any of the above criteria. F = sequence that was obtained and characterized in this chapter.

partial or completed eukaryotic genomes. As can be seen from Table 5.3, homologues of all three retromer component queries were found in the majority of genomes searched.

These homologues to retromer genes from diverse eukaryotes with stacked Golgi bodies demonstrate that the complex is present beyond animals and fungi and may be used as a marker for the presence of Golgi bodies.

Retromer genes in putatively 'Golgi-lacking' taxa

As evidence against the primitive absence of Golgi bodies in putatively 'Golgi-lacking' taxa, genes encoding components of the retromer complex were obtained.

A small but undeniably homologous fragment of the Vps26 gene was amplified from *Mastigamoeba balamuthi*. This sequence is 279 nucleotides long and contains a putative intron of 103 nucleotides with normal GT-AG boundaries. While the final spliced product corresponds to a conceptually translated protein only 57 amino acids long, this still retrieved Vps26 homologues with strong E values (Table 5.2) and contains conserved signatures for the protein (Figure 5.2A).

A fragment of a Vps26 gene (985 nts) was obtained from *Entamoeba histolytica*. As compared with the gene assembled from publicly available GSS reads, this fragment is missing only two amino acids from the N terminus and 46 amino acids from the C terminus of the protein, yielding a conceptual protein 328 amino acids long. In a BLAST search, this protein retrieved Vps26 homologues with high E values and, in fact, retrieved the *Dictyostelium discoideum* Vps26

homologue (called PepA) as its top hit, congruent with proposed organismal relationships (Arisue, Hashimoto et al. 2002; Bapteste, Brinkmann et al. 2002).

A GSS encoding part of the *Spironucleus barkhanus* Vps26 gene was obtained and sequenced in both directions to cover the ORF. Multiple stop codons were found in the relevant reading frame corresponding to the 3'end of the Vps26 gene in plants, yeast and other protist representatives. At the 5' end of the GSS, the ORF is interrupted by an in-frame stop codon. This could demonstrate that the gene is a pseudo-gene, or that the gene is truncated at its 5' end. The other possibility is that, despite the fact that no introns have yet been identified in *Spironucleus*, the reading frame is interrupted by an intron. Consistent with this is the fact that a 36-nucleotide region of the gene, with CT-AG boundaries and containing the stop codon, can be conceptually spliced out to restore an uninterrupted ORF. While the resulting upstream exon is only 28 amino acids long, it does share several highly conserved sites with Vps26 orthologues. In order to test whether this region does represent an intron, exact-match primers were designed to span the region and attempts were made to amplify the gene from cDNA. Unfortunately, while amplification was successful from the GSS, no amplification was observed from cDNA, despite testing upwards of 270 000 plaque-forming units. Regardless, the GSS does not contain the very 5' end of the gene. In a BLASTx, search the 5'-most end of the GSS sequence matches regions of Vps26 homologues at a point approximately 150 amino acids into their ORFs. A BLASTp search with the corresponding ORF from *S. barkhanus* (excluding the potential upstream exon) does shows significant E values to Vps26 proteins (Table 5.2).

A 1450-nucleotide fragment encoding the *Giardia intestinalis* Vps26 protein was amplified using exact-match primers based on single-pass reads from the *Giardia* genome project website. This corresponds to a conceptually translated protein 461 amino acids long. Unfortunately the single-pass reads did not contain the 5'-most portion of the gene, and this was not obtained as it corresponds to a phylogenetically uninformative region of the Vps 26 protein. By BLAST analysis, the gene encoded is clearly aVps26 homologue, as it retrieves a putative Vps26 gene from *Arabidopsis thaliana* as its top hit in a BLASTp search (Table 5.2).

Two separate alleles of the Vps35 gene were identified from *Entamoeba histolytica*. The two amplified regions share only 43% nucleotide identity and the GSS reads to which they correspond have different genomic contexts, demonstrating that they truly correspond to different coding regions within the genome. For *Entamoeba histolytica* Vps35-1, only the 5' region corresponding to an ORF of 1031 nucleotides was amplified. When conceptually translated, this yields a protein (343 amino acids) that retrieves Vps35 orthologues with high E values (Table 5.2). The *Entamoeba histolytica* Vps35-2 gene was amplified across most of its coding region (1968-nucleotide fragment), missing only approximately 55 amino acids from its N terminus and corresponds to an ORF of 655 amino acids. This is somewhat smaller than the average length of Vps35 proteins, which in *S. cereviseae, H. sapiens, A. thaliana, P. yoelii* are 937, 796, 789, and 938 amino acids, respectively. BLAST analysis did confirm *Entamoeba histolytica* Vps35-2 as a Vps35 homologue (Table 5.2).

The entire ORF encoding the *G. intestinalis* Vps35 was characterized. This yielded a conceptually translated protein 763 amino acids in length which strongly retrieved other Vps36 orthologues upon BLAST analysis (Table 5.2).

Retromer Phylogeny

The evolutionary history of the retromer complex has never been examined. Although somewhat tangential to the larger issue of primary *versus* secondary Golgi-lack in eukaryotes, phylogenetic analysis was performed on the three retromer-component protein families as an investigation into the evolution of an important piece of the endomembrane system machinery.

Vps26

All Vps26 homologues present in Genbank were retrieved by BLAST, aligned and analyzed by various phylogenetic methods. Analysis of the initial dataset of 25 taxa, 282 positions (Figure 5.3), robustly united the diplomonad sequences, as well as the animal ones and the two *Plasmodium* species. This dataset also confirmed that several *H. sapiens, D. melanogaster, C. elegans,* and *A. thaliana* sequences represented closely related lineage-specific duplicates (or possibly multiply annotated sequences in Genbank), and could therefore be represented by a single sequence in subsequent analyses.

A dataset with representative taxa and no lineage-specific duplicates was assembled (19 taxa, 120 sites) to reduce the number of taxa in the alignment, and subjected to rigorous phylogenetic analysis. In this case the clade consisting of *M. balamuthi, H. vermiformis* and *D. discoideum* was supported, as was a clade diplomonads (Figure 5.4).

	ML	QP	MLD
A	89	90	88
B	100	92	100
C	100	84	100
D	97	80	96
E	100	98	100
F	100	71	100
G	69	79	32
H	100	69	100

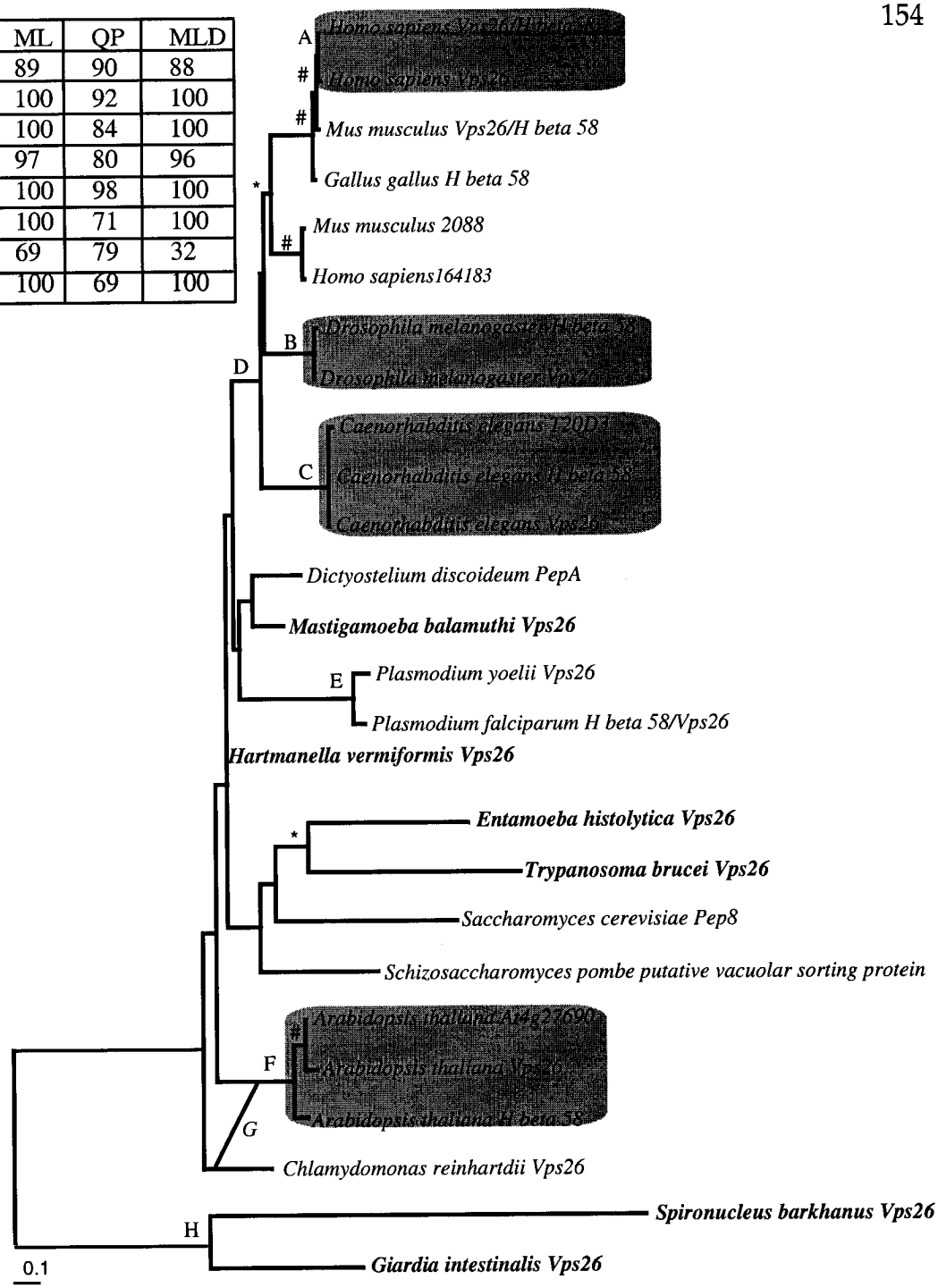


Figure 5.3: **Broad scale Vps26 phylogeny.** This figure shows the best ML distance tree with support values at nodes of interest. The diagonal line indicates an association between two taxa that is not reconstructed by the best topology, but is supported by other methods.

	PMB	ML	QP	MLD
A	95	100	97	100
B	23	62	33	25
C	29	86	81	84
D	99	100	93	100
E	58	97	60	70
F	97	100	46	100

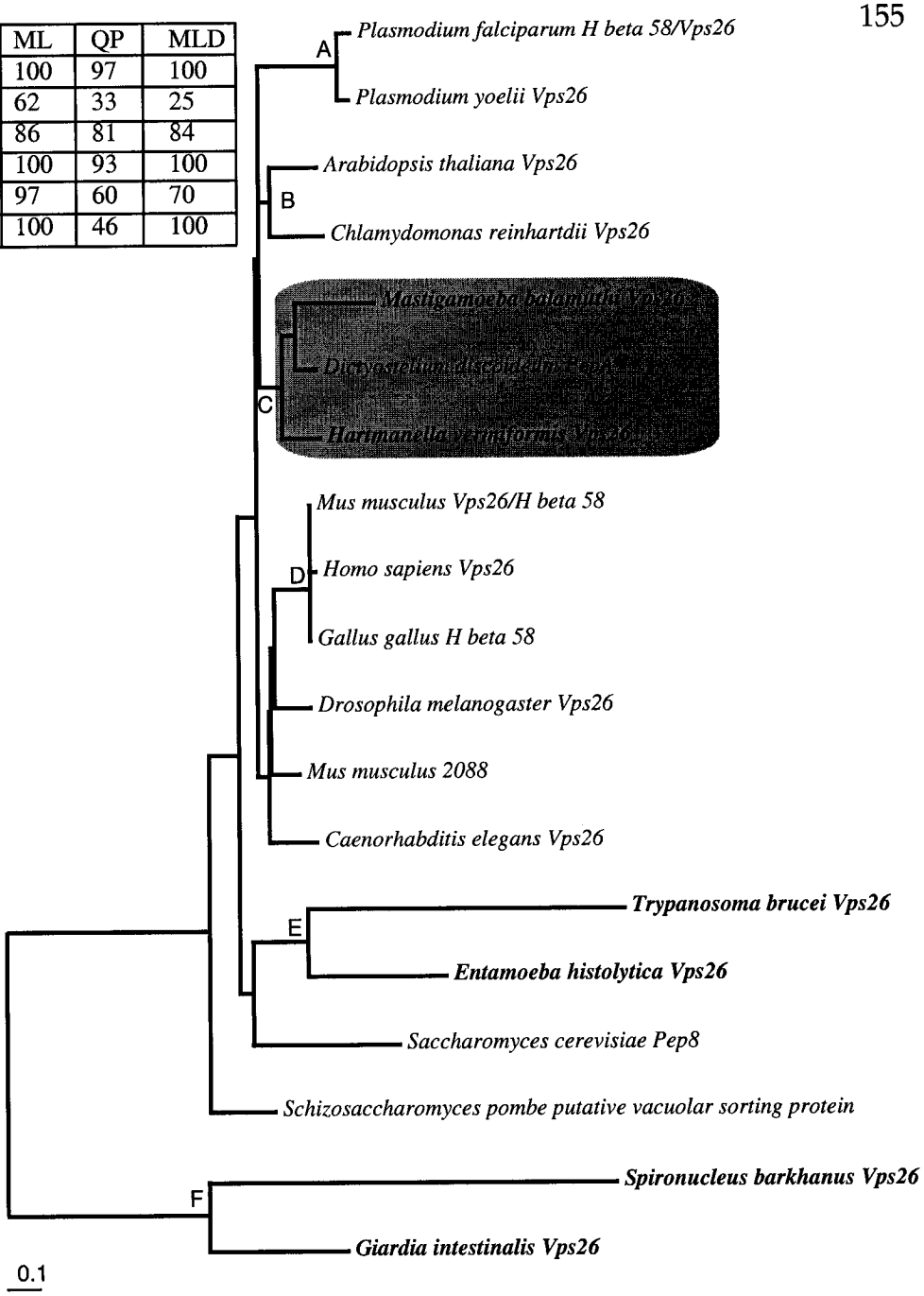
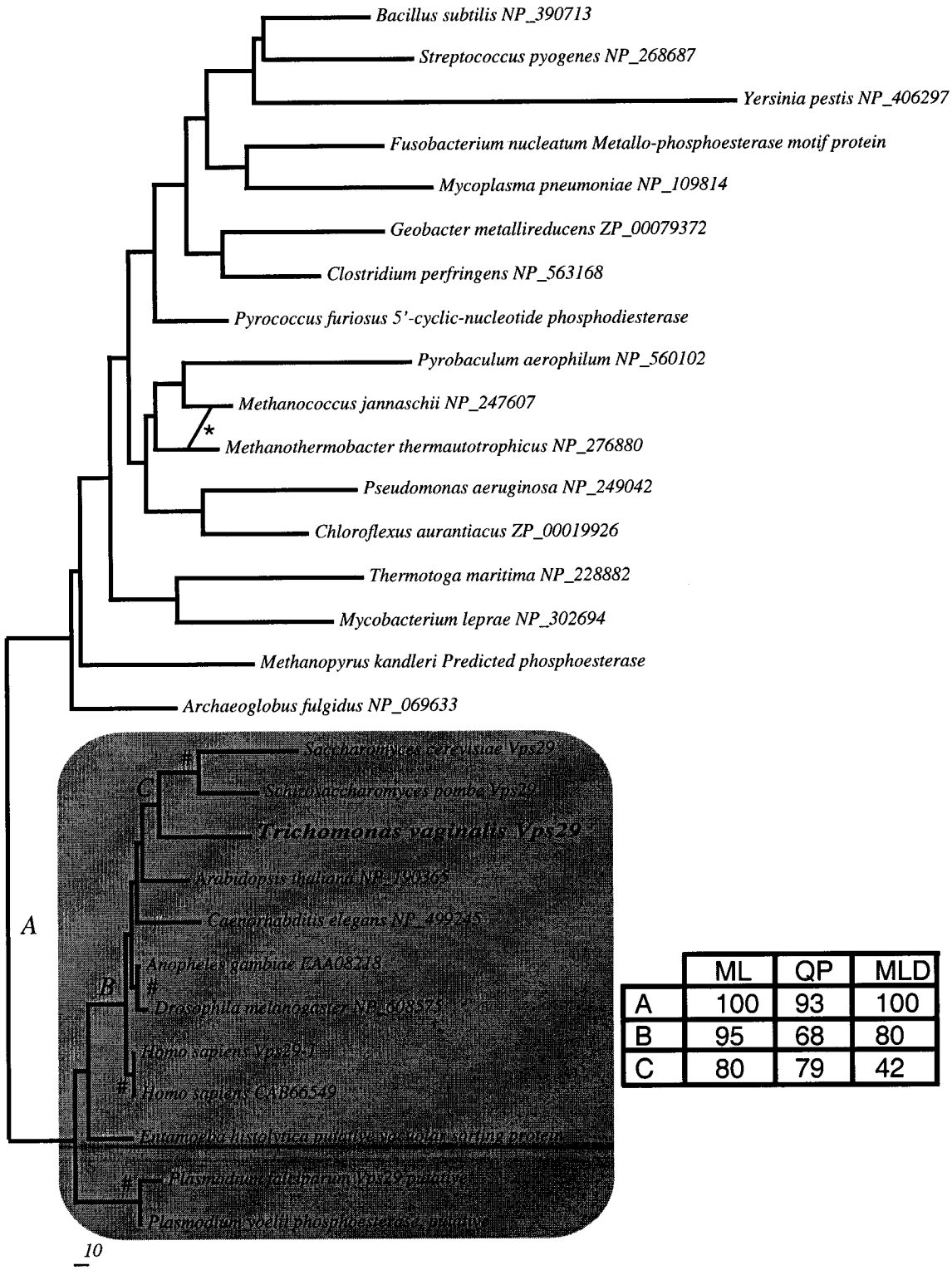


Figure 5.4: **Vps26 phylogeny with representative taxa.** This figure shows the best MLD tree with support values at nodes of relevance. Bootstrap support values from full ML analysis using ProML, are denoted here, and in subsequent figures, as PMB. Note the support for the clade with *Hartmanella, Dictyostelium* and *Mastigamoeba.*

Vps29

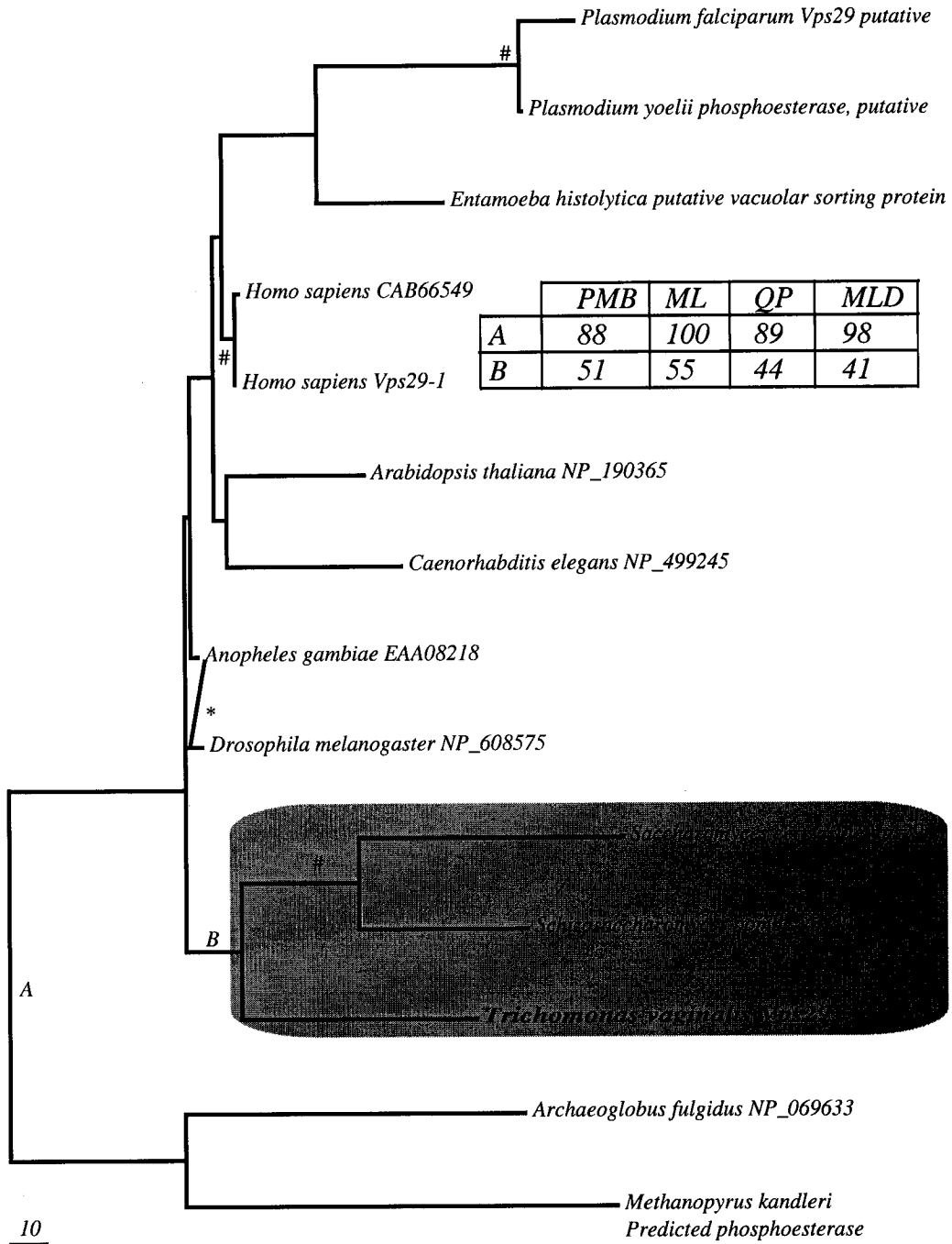
A BLASTp search using the *Trichomonas vaginalis* Vps29 protein was done to retrieve Vps29 sequences from diverse eukaryotes. Surprisingly, in addition to the eukaryotic sequences, proteins from several archaea were also retrieved with significant E values (1e-09 to 2e-04), indicating that they might be prokaryotic homologues. When the retrieved homologue from *Methanothermobacter thermautotrophicus* was used as a query, homologues were identified from diverse archael and bacterial lineages. A taxonomically broad range of prokaryotic homologues was retrieved and aligned with eukaryotic Vps29 sequences. Upon initial analysis with Tree-Puzzle, a number of the sequences failed the amino acid composition test and were discarded. The remaining sequences (29 taxa, 126 sites) were analysed by ML, quartet puzzling and ML distance (Figure 5.5), and showed a strong partition between the eukaryotic and prokaryotic sequences (node A). A reduced dataset consisting of 14 taxa and 126 sites was constructed to test the observed relationship of the *Trichomonas* sequence with the fungi, and also to allow a more rigorous full ML analysis to be performed. Upon analysis (Figure 5.6), the partition between the prokaryotic outgroups and the eukaryotic Vps29 homologues was confirmed. However, the relationship of *Trichomonas* with the fungi fell to negligible values, suggesting that this relationship is artifactual.

Figure 5.5: **Vps29 phylogeny with diverse prokaryotic outgroups.** This figure shows the best full ML topology, with support values at relevant nodes. Note the strong separation of the eukaryotic sequences from the prokaryotic homologues.



	ML	QP	MLD
A	100	93	100
B	95	68	80
C	80	79	42

Figure 5.5: **Vps29 phylogeny with diverse prokaryotic outgroups**



	PMB	ML	QP	MLD
A	88	100	89	98
B	51	55	44	41

Figure 5.6: **Phylogenetic analysis of Vps29, reduced taxon set.** This figure shows the best ML tree, with support values at relevant nodes. While the eukaryotic sequences are clearly separated from the prokaryotic ones, the support for the *T. vaginalis* sequence grouping with the fungi has been reduced to a negligible value.

Vps35

All homologues retrieved in a BLASTp search were aligned and any sequences representing duplicates or nearly identical lineage-specific paralogues were eliminated. A large dataset of 22 taxa, 387 sites was analyzed, but was so fraught with incomplete sequences that ML analysis using ProtML was impossible. In quartet puzzling and ML distance analyses, the multiple alleles found from *Hartmanella vermiformis* and *Reclinomonas americana* formed highly supported clades (Figure 5.7), indicating that these represent nearly identical sequences. On the other hand, the two *Entamoeba histolytica* Vps35 homologues did not form a clade, with the *Entamoeba histolytica* Vps35-2sequence presenting a very long branch. This, along with its aberrant size and divergent sequence at the key conserved amino acid regions (Figure 5.2), suggests that the *Entamoeba histolytica* Vps35-2 sequence is a highly diverged copy of the gene and not necessarily simply a dismissable lineage-specific sequence. Any incomplete Vps35 sequences were eliminated yielding a dataset with 15 taxa and 272 aligned positions. Phylogenetic analysis of this dataset produced little resolution beyond reconstructing the animal lineage, kinetoplastids and a weak affiliation of *R. americana* with kinetoplastids (data not shown). Of the full sequences, several failed the amino acid composition test in Tree-Puzzle, or were clearly long branches. A final dataset (10 taxa, 273 sites) with the long-branch taxa removed was rigorously analyzed. The animal (Figure 5.8, nodes A, B), fungal (node C) and opisthokont clades (weakly, node D) were reconstructed as was a clade uniting the *R. americana* and *T. brucei* sequences with moderate support (node E). Although the taxon sampling is severely limited, Vps35 is one of the few genes

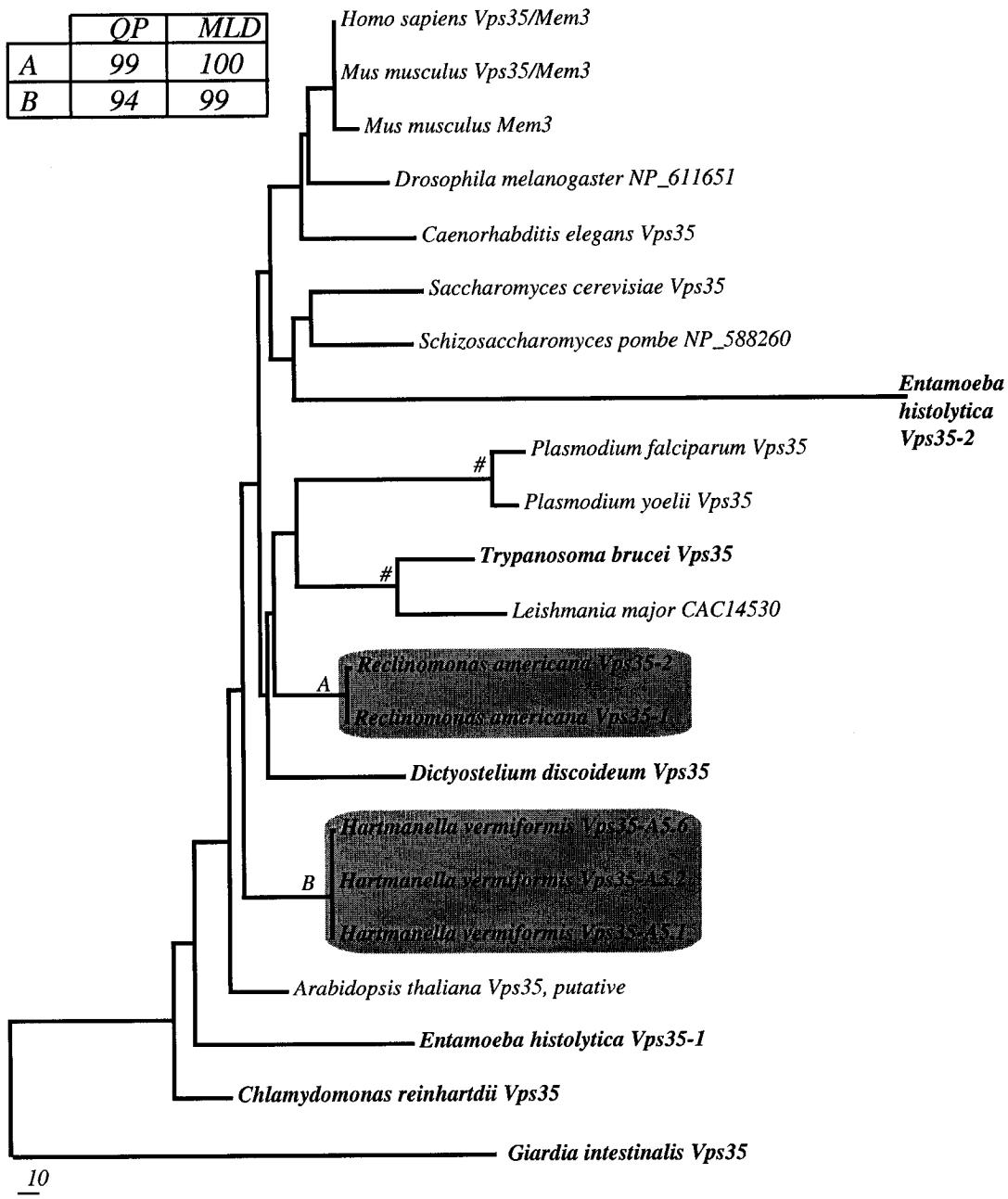


Figure 5.7: **Large Scale Vps35 phylogeny.** This figure shows the best full ML tree. Only quartet puzzling and LM distance support values were obtained for this dataset. These, however, did demonstrate that the alleles of the *R. americana* and *H. vermiformis* Vps35 genes formed strongly supported clades, justifying the use of a single representative sequence for each.

	PMB	ML	QP	MLD
A	100	100	100	100
B	99	100	98	97
C	97	100	94	97
D	73	66	29	32
E	77	79	86	76

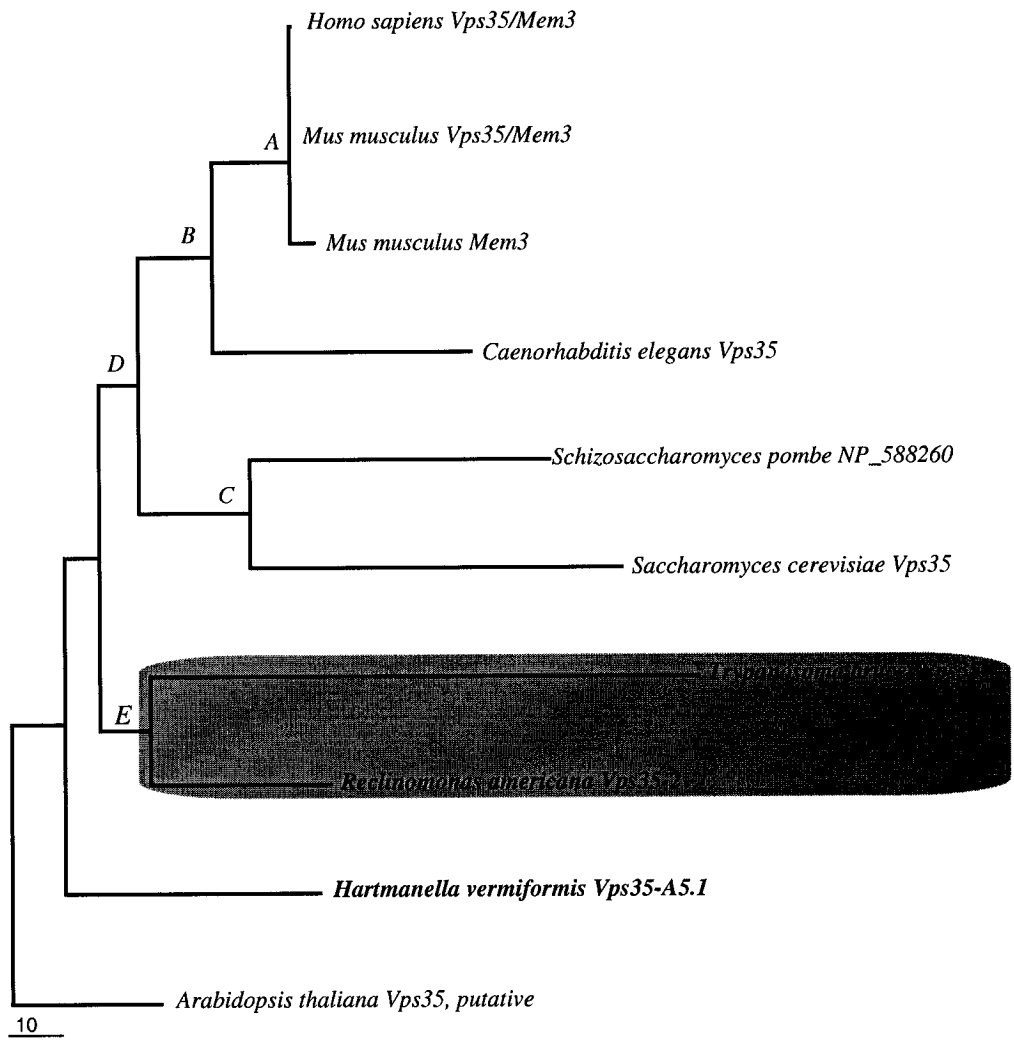


Figure 5.8: **Vps35 phylogeny with full sequences and long-branch taxa removed.** This figure shows the best full ML topolog,y with support values at various nodes. This analysis unites the *Reclinomonas americana* sequence with the *T. brucei* sequence, despite their difference in branch length.

that resolves the position of the jakobids. Re-inclusion of the *Giardia* sequences did not affect the support for this clade, despite the long-branch length of the *T. brucei* sequence (data not shown).

Golgi-specific vesicular-transport components

In addition to the genes encoding retromer components, a number of other putatively Golgi-associated genes were identified from organisms proposed to be 'Golgi-lacking'.

The adaptin complex is involved in coat formation on clathrin-coated vesicles. Adaptins 1 and 2 act at the Golgi and plasma membrane, respectively, while adaptin 3 is localized at both the endosome and Golgi and seems to be involved in *trans*-Golgi-network to endosomal transport (Robinson and Bonifacino 2001). A full-length cDNA from *M. balamuthi* was obtained which, by BLAST (Table 5.2) clearly codes for a small (sigma) subunit of an adaptin complex. A dataset of adaptin-sigma sequences from all four paralogue families was assembled (22 and 119). Full ML and ML distance analysis shows quite strongly that the *Mastigamoeba balamuthi* APs sequence belongs to the Adaptin 3 clade (Figure 5.9).

	PMB	QP	MLD
A	84	93	100
B	42	25	42
C	51	92	96
D	98	90	100
E	100	98	99

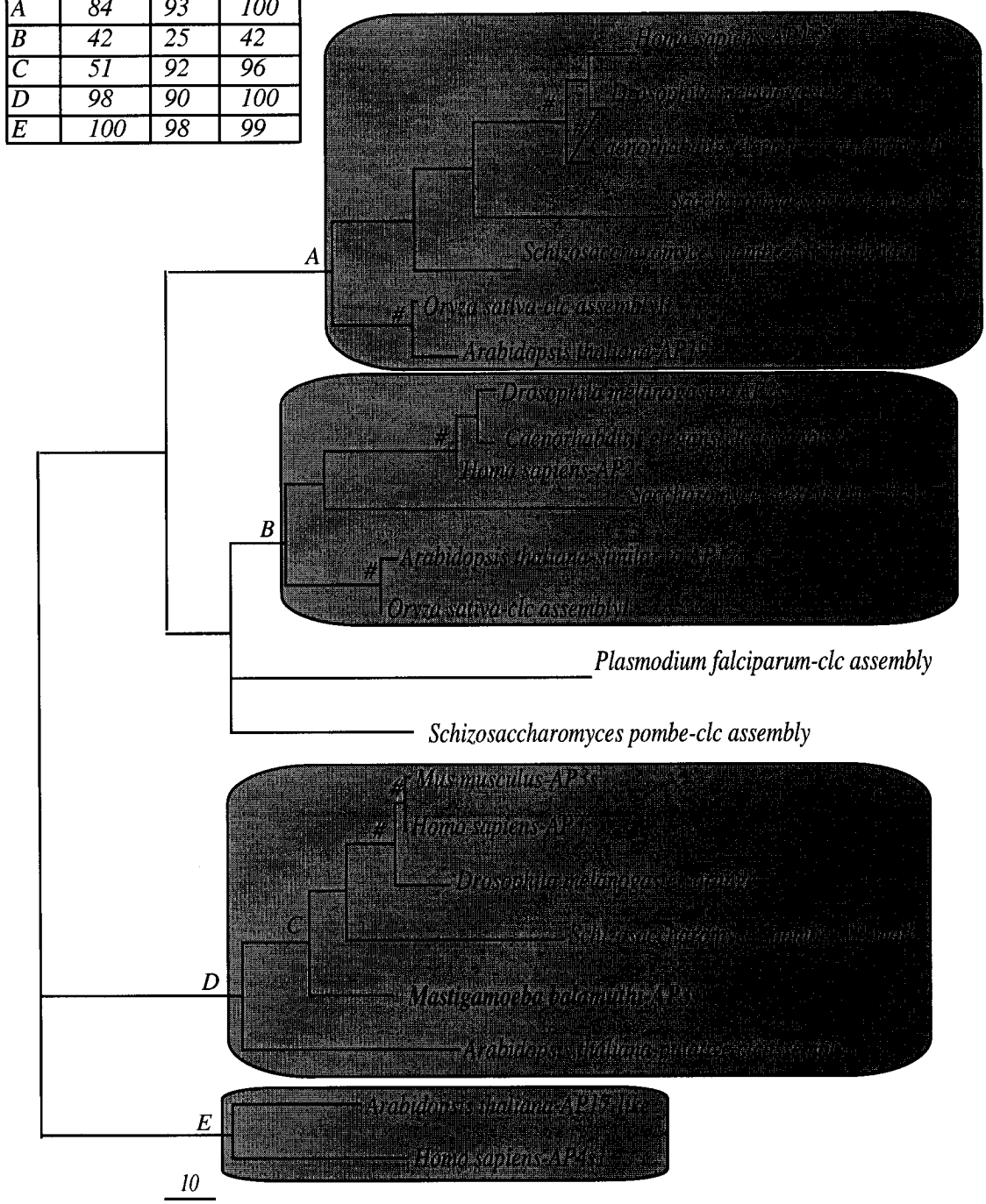


Figure 5.9: **Phylogeny of Adaptin sigma (small) subunit.** In this, the best full ML topology, the four major adaptin subfamiles are seen in boxes, with support values at the relevant nodes. This phylogeny shows the *Mastigamoeba balamuthi* APs (in bold) robustly nested in the AP3 clade.

The coatomer complex forms the coat for transport vesicles involved in both retrograde and possibly anterograde Golgi transport (Orci, Stamnes et al. 1997; Schekman and Mellman 1997). From GSS sequences, coatomer genes were obtained from both *Spironucleus barkhanus* (diplomonads) and *Naegleria gruberi* (heteroloboseids). When the *Spironucleus barkhanus* Beta-COP sequence was used as a query in BLASTp analysis (Table 5.2), only B-COP homologues were retrieved with significant E values. When the *Naegleria gruberi* Beta prime COP sequence was used as a query in a BLASTp search, B'-COP homologues were retrieved with significant E values, as are F-box protein homologues, albeit with much less significant E values (Table 5.2). In both ML and ML distance analyses, the *N. gruberi* sequence forms a clade with B'-COP genes to the exclusion of the F-Box proteins with 100% support, confirming the sequence as a B'-COP homologue (Figure 5.10, node B). Conserved signature motifs exist also for both the *S. barkhanus* (Figure 5.2c) and *N. gruberi* (Figure 5.2d) sequences that reinforce their assignment.

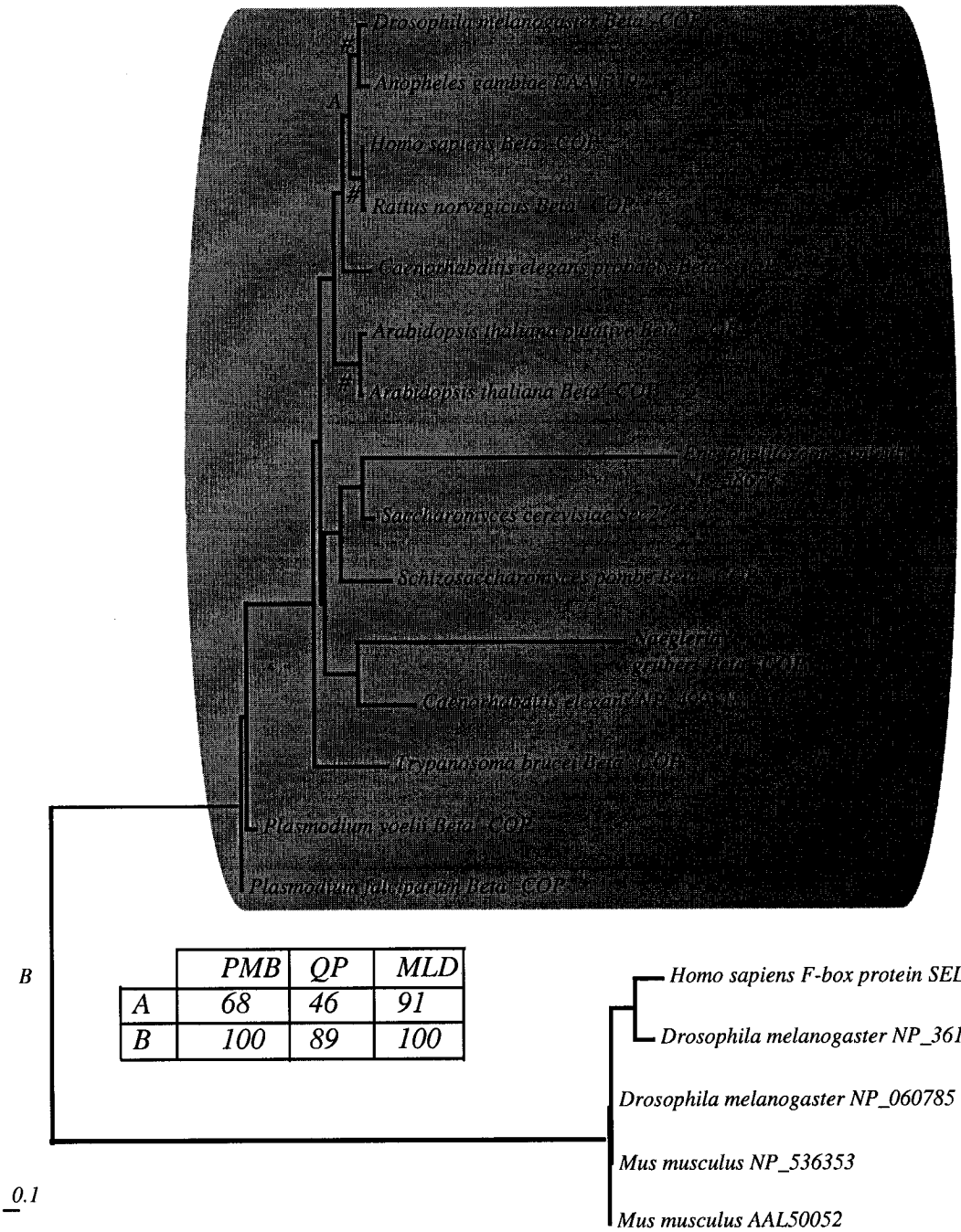


Figure 5.10: **Phylogeny of Beta-prime Coatomer.** This is the best full ML topology, with support values at several nodes. This figure shows the separation of Beta-prime COP sequences away from the F-Box homologues, demonstrating that the *Naegleria* sequence is a real Beta-prime Coatomer .

Discussion

The examination of retromer gene diversity and evolution uncovered several points worth noting.

Prokaryotic homologues

The retromer component Vps29 retrieves significant hits to prokaryotic sequences when used as a BLASTp query. These sequences are clearly homologous based on conserved regions along their length and as indicated from the high E values obtained in BLASTp searches. These are unlikely to be the result of a recent lateral gene transfer, since homologues are found in both archaeal domains and the major bacterial lineages. Additionally, a phylogenetic analysis demonstrates that the eukaryotic sequences form a clade robustly separated from the prokaryotic ones. Following the decision logic in Figure 2.1, the prokaryotic sequences likely represent precursors to Vps29. The majority of the prokaryotic sequences are hypothetical ORFs; however, a number of them are identified as putative phosphoesterases. The biological relevance of a phosphoesterase in membrane transport is unclear, but the homology of the proteins is undeniable. By contrast, when Vps26 and 35 sequences are used as BLAST queries, no significant prokaryotic hits are retrieved.

Eukaryotic relationships

The evolutionary relationship of the amoebae is based on a set of connected phylogenetic inferences: i.e. *Dictyostelium*, *Entamoeba* and *Mastigamoeba* are related (Arisue, Hashimoto et al. 2002; Bapteste, Brinkmann et al. 2002), *Dictyostelium* is related to *Acanthamoeba* (Baldauf, Roger et al. 2000; Dacks,

Marinets et al. 2002; Forget, Ustinova et al. 2002), and *Acanthamoeba* is related to *Hartmanella* (Gunderson, Goss et al. 1994). A recent ssu rDNA analysis unites the gymnamoebae and a larger amoebozoa clade, but without bootstrap support (Bolivar, Fahrni et al. 2001). The relationship of *Hartmanella* plus *Dictyostelium* and *Mastigamoeba*, seen in the Vps26 phylogeny, is therefore not terribly surprising. This phylogeny does, however, represent one of the few analyses that explicitly shows a relationship between these three taxa. Although the topology is unsupported in the full ML bootstrapping analyses, this is likely due to the affinity of the *Entamoeba histolytica* Vps26 sequence for the *Trypanosoma* sequence. Indeed, the relationship that received the highest amount of bootstrap support that was not included in the consensus tree was a relationship of *Entamoeba* and *Trypanosoma* with *Mastigamoeba*.

The evolutionary affinity of the jakobid flagellates is, on the other hand, a fairly open question. While strong morphological evidence supports an excavate clade (Simpson and Patterson 2001), published molecular data has thus far given little support for the jakobids with any of the excavate taxa. The Vps35 phylogeny, albeit quite limited in its taxonomic representation, supports a relationship of *Reclinomonas* with *Trypanosoma*, consistent with morphological and molecular data {Edgcomb, 2001 #831;Archibald, 2002 #898; A. G. B. Simpson, personal communication}.

Evolution of introns

Like mitochondria and Golgi, introns were once thought to be primitively absent from some eukaryotes (Logsdon 1998). It is now clear that this is not the case (Archibald, O'Kelly et al. 2002; Nixon, Wang et al. 2002; Simpson,

MacQuarrie et al. 2002), and yet there are some taxa for which examples of spliceosomal introns have not yet been identified. This is likely due to a scarcity of gene sequence derived from genomic DNA. From retromer component genes, putative introns were identified from *Hartmanella vermiformis* and *Spironucleus barkhanus*. The intron in the *Spironucleus barkhanus* Vps26 sequence is particularly interesting as it has putative CT-AG intron boundaries suggesting that the transition from GT-AG to the non-canonical (CT-AG) splice boundary may have occurred prior to the divergence of the two diplomonad lineages (Archibald, O'Kelly et al. 2002; Nixon, Wang et al. 2002; Simpson, MacQuarrie et al. 2002).

Although the implications and details will be expanded fully in the Conclusions chapter, it is clear, from the retromer-component genes and other genes of the vesicular-transport machinery identified here, that genes encoding proteins of putative Golgi function are present in at least 4 of the 7 major 'Golgi-lacking' lineages (Figure 5.11). This conclusion provides further evidence that these taxa are either secondarily lacking the organelle or else have shifted their organelle to an unrecognizable morphology.

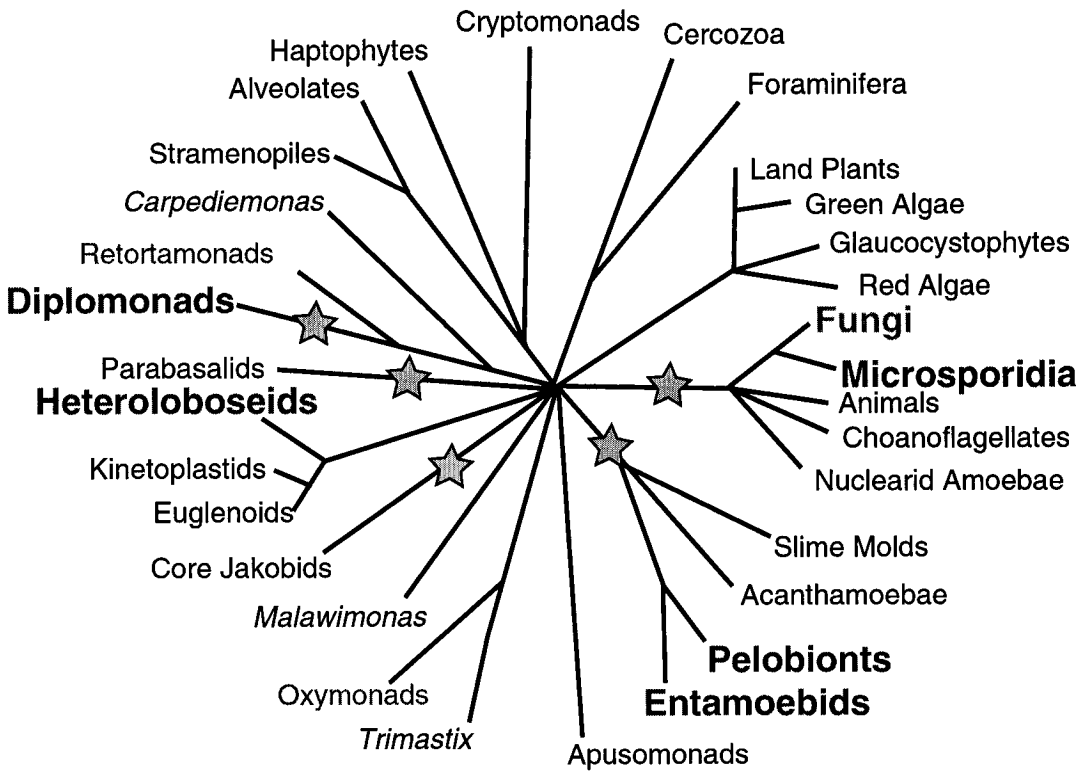


Figure 5.11: **Golgi-lacking taxa possessing direct genetic evidence for cryptic Golgi bodies.** This is the same figure as in Figure 5.1, but now taxa that possess genes encoding putatively Golgi-associated products are bolded.

Chapter 6: Conclusion

While some theses have discrete one question/one chapter formats, for better or for worse this one has gathered data that can be applied to its major questions through a number of different chapters. I therefore want to close with a synthesis of how my data, and that of others published in the field, finally come to bear on the evolution of the endomembrane system and the Golgi apparatus. As well, I suggest some additional studies or alternative angles that could be taken, if examinations of these questions were to be pursued.

Molecular evolution of the vesicular-transport machinery

The machinery involved in vesicular transport is an elaborate and elegant feature of the eukaryotic cell. Over the course of this thesis, I have examined several stages in the evolution of this machinery.

To begin with, a number of potential prokaryotic connections for proteins involved in vesicular transport were identified. The three types of GTPases (Rabs, Arfs and Sar) operating in the vesicular-transport machinery (Jahn and Sudhof 1999; Springer, Spang et al. 1999) most likely evolved from prokaryotic small GTPase proteins. In contrast to the prediction of Cavalier-Smith (Cavalier-Smith 2002), however, there was no evidence that a myxo-bacterial small GTPase was the specific ancestor of these proteins. The NSF and p97 proteins are derived from one or more Cdc48 prokaryotic homologues and the coat proteins of COPI and II vesicles probably were born from proteins possessing WD-40 domains. Finally, the retromer component Vps29 is clearly homologous to phosphoesterases from both prokaryotic domains of life.

There are other prokaryotic homologues of endomembrane system components. The co-translational translocation system is the point of entry for proteins into the vesicular-transport pathway. There are well-characterized homologues for various pieces of this machinery in prokaryotes, which serve similar, if not identical, roles in these cells (Rapoport, Jungnickel et al. 1996). A recent functional study has even shown that, when this system is blocked in *E. coli*, stacks of internal membranes with attached ribosomes accumulate in the cell (Herskovits, Shimoni et al. 2002), eerily reminiscient of the ER. In a similar vein, recent characterization of the archeon *Ignicoccus* revealed a clear double-membraned, intracellular, vesicle (Rachel, Wyschkony et al. 2002). While these structures may only be superficially similar, and not truly homologous, to the eukaryotic systems, they provide examples of the kind of structures found in the endomembrane system arising in a prokaryotic context. Having more than one example of this makes any suggested model of the process more plausible.

The proto-eukaryote then already possessed the building blocks that would eventually build the regulatory GTPases (Jahn and Sudhof 1999; Springer, Spang et al. 1999), complex-dissociating ATPases (Jahn and Sudhof 1999), vesicle coats (Springer, Spang et al. 1999) and several other pieces of the vesicular-transport machinery. The various components would come together to create an endomembrane system that appears already to have been well established by the time the Last Common Eukaryotic Ancestor arose. Based on the comparative genomic searches in Chapter 2 and 5, as well as the retromer components from *Trichomonas, Reclinomonas, Hartmanella* and *Mastigamoeba*, it is clear that many of the protein families that are involved in a generalized vesicular-transport process are present in a wide diversity of eukaryotes, and consequently were likely

present in an early eukaryotic ancestor. Some of the important duplications establishing major vesicular-transport component families had also occurred by this point, including the p97/NSF duplication as well as the Sar1/Arf duplication.

The comparative genomic survey showed that an ancestral syntaxin was present in the LCEA. From the study of syntaxins in Chapter 3, it appears that the complexification within the protein family also began early in eukaryotic evolution. I obtained 15 syntaxins from diverse eukaryotes (diplomonads, entamoebids, parabasalids, both red and green algae, kinetoplastids and stramenopiles). In addition, genomics projects provided sequence from animals, fungi, microsporidians, plants, apicomplexans, and slime molds. Phylogenetic analysis showed that the major families of syntaxins, by and large, group into 'mini' eukaryotic trees. For almost every taxon there was evidence that at least two of the major families had been created from a duplication that occurred before its divergence. Taken together, these data suggest that the syntaxin families had already established themselves at an early stage in the evolution of the endomembrane system. This complexification of the syntaxin system could potentially implicate them, in some way, in the origin or early evolution of the endomembrane system. One major obstacle in first evolving a permanent internal membrane system would be establishing a system that is both stable in the identity of its compartments and its maintenance within the cell, and yet dynamic enough to accommodate incoming and outgoing vesicles. *In vitro* reconstitution assays have shown that syntaxins not only play a role in vesicular transport but also in organellar reconstruction (Patel, Indig et al. 1998; Rabouille, Kondo et al. 1998; Roy, Bergeron et al. 2000; Wickner and Haas 2000). This

implies that they could be partially responsible for the "identity" of an organellar compartment and might have been able to fulfill the early role necessary for internal membrane stability and flexibility.

This complexification-by-duplication trend has continued from the establishment of protein families right up through to system-specific and lineage-specific components. This trend is particularly relevant when looking at the evolution of multicellular organisms. My phylogenies found that, in the PM-specific syntaxin family, both the plant and animal syntaxins had undergone lineage-specific duplications. This expansion seemingly has happened convergently in both systems and is symptomatic of a larger process. Simply by comparing the number of identified syntaxin homologues in genomes a clear trend is seen of multiplicity of syntaxins in multicellular (*H. sapiens* = 12, *C. elegans* = 9, *A. thaliana* = 24) versus single celled (*S. cerevisiae* = 7, *G. intestinalis* = 4) organisms. Although, this analysis is hamstrung somewhat by the limitations of BLAST (that might miss a highly diverged paralogue), the incompleteness of draft genomes and the different criteria used in the various studies as to what constitutes a distinct paralogue (Sanderfoot, Assaad et al. 2000; Bock, Matern et al. 2001), the trend is still likely to be robust.

From simple prokaryotic origins, it seems, the skeleton of the vesicular-transport machinery sprang forth quickly and well connected. In the case of the syntaxins, at least, then the system fleshed out and out with increasing complexity to the multicelled complements observed today.

All of these conclusions, however, are based on comparative genomics studies of limited taxa and analysis of only one gene family. A number of important areas of eukaryotic diversity (excavate, cercozoa) are poorly

represented, and there are a number of genome initiatives (pelobionts, trichomonads, choanoflagellates, chlorarachniophytes, euglenids and others) that are not yet publicly available. Repeating this study when these taxa can be included will allow a better approximation of the LCEA and, excitingly, perhaps show an intermediate stage in the incremental evolution of the endomembrane system that must have taken place. Using the same approach on a different system within the endomembrane machinery would also be a worthwhile endeavor. The clear prokaryotic connections of the co-translational translocation system make it an attractive candidate for this type of approach. The establishment of organellar identity may also have been a key event, and so machinery involved in post-mitotic reassembly of endomembrane compartments (other than syntaxins) would also be a fruitful place to look.

As pointed out in Chapter 2, though, the comparative genomics approach using current methods and incomplete genome sequence is, by necessity, broad but shallow in its scope. Many of the most interesting aspects will be about the more detailed evolution of protein machinery. These issues require reliable sequence, phylogenetic analysis and lots of it. The picture of syntaxin evolution, both functional evolution and diversification, will certainly be clearer for the new data that genome initiatives will provide as they become publicly available. The same approach taken in Chapter 3, however, could have been and should be taken with any one of the endomembrane system components. There have been a number of functional and evolutionary investigations into the Rab protein family (Bush, Franek et al. 1993; Field and Boothroyd 1995; Janoo, Musoke et al. 1999; Saito-Nakano, Nakazawa et al. 2001; Denny, Lewis et al. 2002; Jeffries, Morgan et al. 2002; Langford, Silberman et al. 2002). These have shown a similar tale of deep

duplications as well as lineage-specific expansions. Delving more deeply into the duplication stories surrounding NSF/p97 and Sar1/Arf, as well as the deeply nested expansions that gave rise to the some of the adaptin and Coatomer complex (Schledzewski, Brinkmann et al. 1999), should all provide useful information to further unravel the origin, evolution and diversification of the endomembrane system.

Previous attempts to explain the evolution of the endomembrane system have been highly speculative and vague. Nonetheless, as critical a transition as that from the prokaryotic to eukaryotic internal membrane organization and up to the present complexity deserves a carefully laid-out explanation. Such a model would begin with prokaryotic homologues present in the proto-eukaryote, and propose constructions and expansions of organelles and machinery based on information from gene phylogenies. The most important aspect of the model would be that it provides testable hypotheses upon which researchers could build. Even if the entire model were to be proven false, its creation would be an important step forward for the falsifying experiments that it would have engendered.

Evolution of the Golgi apparatus in eukaryotes

Involved in both secretion and endocytosis, the Golgi apparatus plays a deeply entrenched role in the life of the cell, and yet there are a few eukaryotic lineages that are thought to lack this organelle (Cavalier-Smith 1987; Patterson 1999). The evolution of the Golgi apparatus and whether putatively 'Golgi-

lacking' taxa do, in fact, possess a homologue of the organelle can be addressed with three types of data, of which two were obtained in this thesis.

All of the major eukaryotic lineages that 'lack Golgi' have phylogenetic affiliations with lineages that possess the stacked organelle. My contribution to this pool of data was to determine that the oxymonads are related to the Golgi-possessing lineage *Trimastix*. Using small subunit ribosomal rDNA analyses, I was responsible for the placement of an initial sequence of an oxymonad (*Pyrsonympha*), and for confirmation that the result was robust to the long-branch artifact that has plagued past ssu rDNA analyses. This initial result was then expanded to other oxymonad genera and to the full breadth of excavate taxa. The affiliation of oxymonads and *Trimastix* is highly supported and reliable.

In addition to the ssu rDNA phylogenies presented here, there is evidence for this relationship from several other lines. The affiliation of *Streblomastix strix* with the *Pyrsonympha* JD 2000 sequence, and these as sister taxa to *Trimastix*, has recently been confirmed based on ssu rDNA analyses (Keeling and Leander 2003). Unpublished protein phylogenies of EF1alpha (Yuji Inagaki, personal communication) and tubulin (Andrew Roger, personal communication) show the *Trimastix*/oxymonad relationship as well.

There are extensive morphological data supporting a common origin for excavate taxa (Simpson and Patterson 1999; Simpson and Patterson 2001). In an ultrastructural analysis of *Monocercomonoides*, an oxymonad, several homologies were proposed between its cytoskeleton and that of excavate taxa, particularly to *Trimastix* (Simpson, Radek et al. 2002). The pre-axostyle has been the long-standing unique feature of oxymonads. In that analysis it was shown that the oxymonad pre-axostyle is homologous to the I-fibre of excavate taxa. The

outstanding characteristic of excavates is the feeding groove formed by the left root, B-fibre, and singlet microtubule. Homologous structures for each of these features are present in oxymonads, providing a convincing argument for their excavate ancestry. The vanes on the C-fibre of oxymonads and *Trimastix*, as well as the structure and thickness of the I-fibre, also argue for an exclusive relationship of oxymonads and *Trimastix*, consistent with the molecular data.

The second type of data that bears on the issue of primary *versus* secondary Golgi evolution is the presence of genes whose products have functions specific to the Golgi apparatus. The various retromer and vesicular-transport components that I obtained in Chapter 5, as well as syntaxins from *Entamoeba* (PM and 5) and *Giardia* (PM, 16 and 18), serve as evidence against the idea that the taxa examined lack Golgi entirely and primitively. Rab proteins from *Giardia* (Langford, Silberman et al. 2002) and *Entamoeba* (Saito-Nakano, Nakazawa et al. 2001) have also been identified that are of putative Golgi affiliation. The genome of the microsporidian *Encephalitozoon cuniculi* contains homologues for Adaptin 1, six of the seven coatomer subunits, and multiple other Golgi-associated components (Katinka, Duprat et al. 2001). Biochemical data such as the sensitivity to Brefeldin A may also demonstrate the presence of a Golgi apparatus. *Entamoeba* and *Giardia* both show Brefeldin A sensitivity (Lujan, Marotta et al. 1995; Ghosh, Field et al. 1999).

These first two types of evidence are, to diminishing degrees, inferential. The final type of data is the most direct: identification of the organelle in its non-canonical form by immunolocalization. This has been done for *Giardia, Entamoeba* and microsporidians (Lujan, Marotta et al. 1995; Ghosh, Field et al. 1999; Sokolova, Snigirevskaya et al. 2001). The evidence for Golgi bodies in these taxa

is now overwhelming and it is clear that they no longer qualify as 'Golgi-lacking' eukaryotes. The other major lineages possess varying degrees of evidence and so the skeptic might well argue that their cases have not yet been closed. This thesis has been responsible for the first evidence of any kind against the primary lack of Golgi bodies in oxymonads, and the first data of the second type for the heteroloboseids and pelobionts.

That a gene is Golgi-associated in some taxa does not necessarily mean that it is in all of them. Localization and characterization of the gene products identified here as putatively Golgi-associated will be an important and exciting confirmatory step. Nonetheless, given that the homologues of the genes here have been functionally characterized (Bennett and Scheller 1993; Jahn and Sudhof 1999; Wickner and Haas 2000) in organisms both with stacked Golgi (mammals) and without (yeast), the most parsimonious hypothesis is to assume conservation of gene function. Even should this prove to be incorrect for one marker, there are many 'Golgi-associated' genes available, for diverse taxa, and the case would have to be made against each one independently. It is unlikely that they would all fail to be 'Golgi-associated'. The proposed 'Golgi-lacking' nature of the seven lineages was based on a presumption of deep-branching status and lack of observable stacked organelles, rather than on direct evidence against their presence (Cavalier-Smith 1983). Given the lack of direct evidence for Golgi absence, the multiple, likely Golgi-associated, genes described here make a strong case for the presence of Golgi organelles with alternative morphologies, as are observed in fungi and alveolates. Even excluding our genetic data, but assuming that the proposed relationships in Figure 4.8 are correct, the only way to refute the indirect evidence against primary Golgi lack is to place the root of

eukaryotes on a 'Golgi-lacking' lineage as in Figure 5.1. Since the root of eukaryotes can only be placed on a single lineage, however, this implies an unseen organelle in the others. This deduction is premised on a single origin of Golgi, which is highly probable, given the near universal distribution and conserved morphology of Golgi across the vast majority of eukaryotes (Becker and Melkonian 1996).

There are, nevertheless, several loose ends that need to be dealt with. A better resolved eukaryotic phylogeny would help to solidify some of the indirect inferences about secondary Golgi loss. It will be important to establish an outgroup for the oxymonad/*Trimastix* lineage. The current assignment of *Malawimonas* as the outgroup to this clade is certainly believable but based on exceedingly weak evidence (Simpson, Roger et al. 2002). Also, knowing the exact placement of *Stephanopogon* would be useful to confirm whether the loss of the stacked organelle is an independent shift (Cavalier-Smith 2002) or occurred as part of the shift in the heteroloboseans (Cavalier-Smith 2000). Finally, rooting the tree of eukaryotes remains a significant challenge. The DHFR/TS fusion marker (Stechmann and Cavalier-Smith 2002) represented a major step forward but this needs to be confirmed with other evidence and reconciled with other markers such as the enolase insertion (Keeling and Palmer 2000), and mitochondrial genome content (Lang, Burger et al. 1997).

More genes of proposed Golgi function from the groups that are already examined would also be useful. The diplomonads and entamoebids have fairly large bodies of evidence and more will certainly emerge from the GSS and genome projects. However, there are only one or two pieces of evidence each for the heteroloboseids and pelobionts. As the genome initiatives for these taxa

progress, hopefully more genes of Golgi function will be identified. Three lineages that lack cytological evidence for Golgi bodies currently have not had genes of proposed Golgi function obtained from them. Retortamonads are clearly related to diplomonads (Silberman, Simpson et al. 2002) and *Carpediemonas* (Simpson, Roger et al. 2002), and it is unlikely that the root of eukaryotes will fall inside this clade. Both the diplomonads and *Carpediemonas* have good evidence for Golgi bodies (Lujan, Marotta et al. 1995; Simpson and Patterson 1999). The oxymonads are clear, highly derived, relatives of *Trimastix*, which possess Golgi bodies and again it is unlikely that the root of eukaryotes will fall inside that clade. *Stephanopogon* is rarely proposed as either a major lineage or as deeply branching. Nonetheless, obtaining genes of proposed Golgi function from these three taxa will be important pieces of the puzzle of Golgi origins and useful for comparative cell-biological examinations of the Golgi apparatus.

Immunolocalization of proposed 'Golgi-associated' proteins has been done in yeast (Banfield, Lewis et al. 1995; Bevis, Hammond et al. 2002), *Giardia* (Lujan, Marotta et al. 1995; Marti, Li et al. 2003) and microsporidia (Sokolova, Snigirevskaya et al. 2001), but should be done for all 'Golgi-lacking' taxa. In addition to confirming the evidence against primary lack of Golgi bodies, these experiments may identify the physical organelle and provide powerful comparative data for Golgi morphology and the function of important gene products.

The sum of available data supports the idea that there are no primitively 'Golgi-lacking' taxa. A further question, then, is whether the Golgi apparatus itself was truly lost or simply shifted to a non-stacked and therefore non-obvious form in a given taxon. In determining the answer to this question, one can look to

several lines of argument. The role of the Golgi body is so central in both endo- and exocytosis that it is difficult to imagine how either process would proceed in the true absence of the organelle. On the other hand, examples of a shift from stacked to non-stacked Golgi can be found in both *S. cereviseae* and likely in *Giardia*. In these taxa, immuno-microscopy localizes Golgi-associated proteins to cytoplasmic vesicles (Banfield, Lewis et al. 1995; Lujan, Marotta et al. 1995; Bevis, Hammond et al. 2002; Marti, Li et al. 2003), thus identifying the vesicles as the physical organelle. The preservation of conserved functional residues in Golgi proteins, such as Vps35, also suggests the presence of a functional protein component, rather than a decaying product from a discarded organelle. Finally, the presence of multiple 'Golgi-associated genes' implies a functioning rather than decaying system.

It is likely that the Golgi apparatus evolved once in eukaryotes and, like mitochondria (Roger and Silberman 2002) and introns (Simpson, MacQuarrie et al. 2002), Golgi bodies were present in the last common ancestor of all extant eukaryotes. The shift from a stacked to non-stacked form likely happened independently in diverse eukaryotic lineages, a minimum of 4 times (Figure 6.1). *Stephanopogon* is not shown in that figure but likely represents an independent shift of morphology. This would count for one more shift, but rooting the tree of eukaryotes on diplomonads or within the Conosa (either on the pelobionts or the entamoebids) would reduce the number of shifts by one. The focus now must be on the diversity of form that this organelle can present, and diversity of gene sequence involved in this form. Together these will lead to a better understanding of how this diversity affects underlying function common to all eukaryotes and of the evolution of the organelle itself.

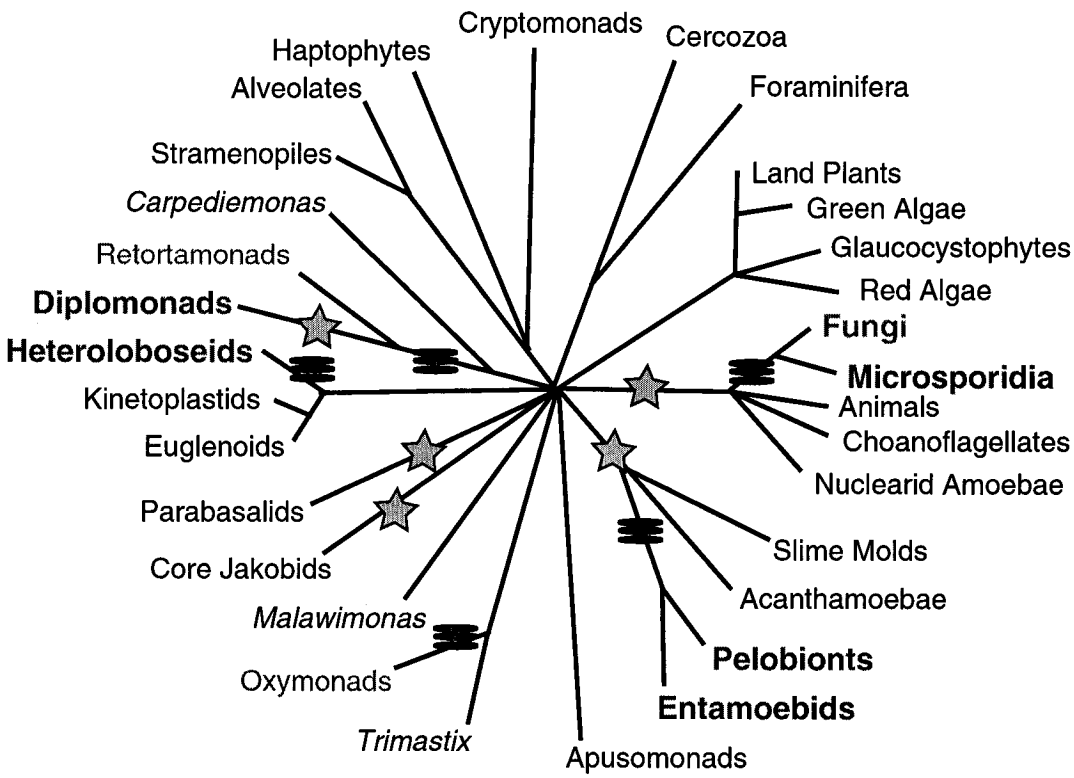


Figure 6.1: **Loss of Golgi bodies among eukaryotic lineages.** This figure shows, for the final time, the schematic of eukaryotic relationships, with potential rootings, and Golgi-lacking *versus* Golgi-possessing taxa. The cartoon Golgi stacks with a red X denote a shift from the stacked organelle to an unstacked or non-canonical morphology.

Eukaryogenesis and final conclusions

So what can be made of the prokaryote-to-eukaryote transition? The "missing link" eukaryote is an elusive beast indeed, as almost all eukaryotes examined seem to have almost all eukaryotic traits (Roger 1999). This is not to say that systems haven't been simplified or discarded. Using the same logic as for Golgi bodies, based on the variable presence of introns and splicing machinery in various eukaryotes and potential rootings of the eukaryotic tree, we have to deduce that the splicing machinery, and introns themselves, must have been streamlined in some lineages. This is consistent with the proposed neutral evolutionary origins of intron splicing (Stoltzfus 1999). In contrast, the endomembrane system seems surprisingly complete. The "minimal" core machinery deduced in Chapter 2 is quite complex, and at every opportunity that minimal compliment gets expanded. Gene families get added to the "minimal" list (syntaxins, the retromer complex, or Rabs) and "missing" organelles get found. This bespeaks the indispensability of the endomembrane system in the eukaryotic cell and hints at a possible principal role in eukaryogenesis (Dacks and Doolittle 2001; Dacks and Doolittle 2002).

Nonetheless, we are left with a few uncomfortable questions. Why did extant eukaryotes seemingly diversify in a rapid radiation? Why do all extant eukaryotes seem to have the nearly complete compliment of eukaryotic features or, in other words, why were there no transitional eukaryotes left behind? Finally, and just as importantly, why did a series of complex and unlikely features all arise in the one eukaryotic lineage?

The first question might be deflected with arguments of phylogenetic artifact. The eukaryotes appear to have radiated because the phylogeny is unresolvable, a product of mutational saturation over a long period of time (Philippe and Adoutte 1998). The second question could also be set aside due to insufficient eukaryotic sampling (Lopez-Garcia, Rodriguez-Valera et al. 2001; Dawson and Pace 2002) and consequent failure thus far to have uncovered our simple ancestor. Both of these responses are reasoned and entirely possible. The last question is somewhat more difficult to ignore. Is it possible that all of these questions though, have more to do with population biology and ecology than they do with sampling and phylogeny?

Suppose the following. It is well established by now that lateral gene transfer is a significant source of evolutionary novelty in prokaryotes (Doolittle 1999; Gogarten, Doolittle et al. 2002). Imagine that one lineage of prokaryotes receives a fortuitous combination of genes encoding an adaptation. This adaptation causes both a selective advantage and an isolation of that cell line. The isolation could be physical, possibly a size increase due to an endomembrane system. It could also have been an ecological or physiological isolation, such as mitochondria (Martin and Muller 1998) or chromatin (Kasinsky, Lewis et al. 2001). Irrespective of its nature, this isolation causes the rate of lateral gene transfer into that lineage to severely decrease. While clearly unicellular eukaryotes are not immune to transfer (Andersson, Sjögren et al. 2003), nor should they be (Doolittle 1998), the evidence thus far suggests that the rate of later gene transfer is much lower in eukaryotes than in prokaryotes (Katz 2002). On the other hand, of the prokaryotic genomes sampled, the majority tend to be smaller than eukaryotic genomes, and make less use extensive of

paralogous gene families (Friedman and Hughes 2001; Alberts 2002). As well, most of the prokaryotes examined seem to make use of an r-selected strategy rather than a k-selected one (Carlile 1982). It is possible that the isolation from competition, or possibly the selective advantage provided by the new adaptation, however, released the proto-eukaryotes from pressure selecting rapid replication through streamlined genome organization. The expansion of genome size, through parasitic DNA elements or gene duplication with paralogue diversification, would have then provided the source novel function acquisition previously supplied by LGT. Whether eukaryotic chromatin evolved as a product of this genome size increase or else facilitated the expansion is unclear. However, the origin of the histones from the two different prokaryotic domains hints at its proto-eukaryotic evolution (Kasinsky, Lewis et al. 2001), at least in this model. The evolution of mitosis would have facilitated the replication of the newly expanded genome. This genome increase and the subsequent complexification of replication further slows reproductive rate, but enhances fitness. Meiosis evolves possibly as a variation of the mitotic machinery or perhaps due to parasitic elements that would now be able to take hold in the non-streamlined eukaryotic genome. This provides the selectively advantageous ability to occasionally inject added novelty or diversity at times of need (Dacks and Roger 1999). Meiosis has an additional effect, however, of causing a reticulating population and stymieing divergent speciation. This, along with the low reproductive rate strategy of the proto-eukaryote, causes a stacking of novel eukaryotic traits in the single lineage and the lack of independent transitional eukaryotes. Eventually the tinkering produces the LCEA, whose subsequent progeny diversify in an explosive radiation (Philippe, Germot et al. 2000). This

radiation could have been in response to colonization of a novel niche, evolution of one key adaptation (again possibly mitochondria, or phagocytosis) or merely a shift back to r-selection as population size got larger and competition increased.

Of course, this hypothesis is vague and highly speculative. It is, however, consistent with some observations of genomic organization and ecological behavior difference between prokaryotes and eukaryotes. It also provides new angles on a few sticky problems in eukaryogenesis, namely the lack of transitional forms and trait-stacking in the eukaryotic line.

Regardless, the question of eukaryogenesis and the prokaryote-to-eukaryote transition is a fascinating one, and one that will be debated for decades to come. The studies aimed at this question will benefit research in general, whether basic (cell biology and computer science), or applied (parasitology and medicine), in addition to addressing one of the outstanding events in our cellular evolution.

References

Abeliovich, H., E. Grote, P. Novick and S. Ferro-Novick (1998). "Tlg2p, a yeast syntaxin homolog that resides on the Golgi and endocytic structures." J Biol Chem **273**(19): 11719-27.

Adachi, J. and M. Hasegawa (1996). "MOLPHY Version 2.3, Programs for molecular phylogenetics based on maximum likelihood." Computer Science Monographs **28**.

Adam, R. D. (2001). "Biology of *Giardia lamblia*." Clin Microbiol Rev **14**(3): 447-75.

Alberts, B. (2002). Molecular biology of the cell. New York, Garland Science.

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-402.

Anantharaman, V., E. V. Koonin and L. Aravind (2002). "Comparative genomics and evolution of proteins involved in RNA metabolism." Nucleic Acids Res **30**(7): 1427-64.

Andersson, J. O., A. M. Sjögren, L. A. Davis, T. M. Embley and A. J. Roger (2003). "Phylogenetic analyses of diplomonad genes reveal frequent lateral gene transfers affecting eukaryotes." Curr Biol **13**(2): 94-104.

Archibald, J. M., D. Longet, J. Pawlowski and P. J. Keeling (2003). "A novel polyubiquitin structure in cercozoa and foraminifera: evidence for a new eukaryotic supergroup." Mol Biol Evol **20**(1): 62-6.

Archibald, J. M., C. J. O'Kelly and W. F. Doolittle (2002). "The chaperonin genes of jakobid and jakobid-like flagellates: implications for eukaryotic evolution." Mol Biol Evol **19**(4): 422-31.

Arisue, N., T. Hashimoto, et al. (2002). "The phylogenetic position of the pelobiont *Mastigamoeba balamuthi* based on sequences of rDNA and translation elongation factors EF-1alpha and EF-2." J Eukaryot Microbiol **49**(1): 1-10.

Arisue, N., L. B. Sanchez, L. M. Weiss, M. Muller and T. Hashimoto (2002). "Mitochondrial-type hsp70 genes of the amitochondriate protists, *Giardia intestinalis*, *Entamoeba histolytica* and two microsporidians." Parasitol Int **51**(1): 9-16.

Armstrong, J. (2000). "Membrane traffic between genomes." Genome Biol **1**(1).

Baldauf, S. L. and W. F. Doolittle (1997). "Origin and evolution of the slime molds (Mycetozoa)." Proc Natl Acad Sci U S A **94**(22): 12007-12.

Baldauf, S. L. and J. D. Palmer (1993). "Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins." Proc Natl Acad Sci U S A **90**(24): 11558-62.

Baldauf, S. L., A. J. Roger, I. Wenk-Siefert and W. F. Doolittle (2000). "A kingdom-level phylogeny of eukaryotes based on combined protein data." Science **290**(5493): 972-7.

Banfield, D. K., M. J. Lewis and H. R. Pelham (1995). "A SNARE-like protein required for traffic through the Golgi complex." Nature **375**(6534): 806-9.

Bapteste, E., H. Brinkmann, 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**(3): 1414-9.

Barnes, W. M. (1994). "PCR amplification of up to 35-kb DNA with high fidelity and high yield from lambda bacteriophage templates." Proc Natl Acad Sci U S A **91**(6): 2216-20.

Bassham, D. C. and N. V. Raikhel (1999). "The pre-vacuolar t-SNARE AtPEP12p forms a 20S complex that dissociates in the presence of ATP." Plant J **19**(5): 599-603.

Becherer, K. A., S. E. Rieder, S. D. Emr and E. W. Jones (1996). "Novel syntaxin homologue, Pep12p, required for the sorting of lumenal hydrolases to the lysosome-like vacuole in yeast." Mol Biol Cell **7**(4): 579-94.

Becker, B. and M. Melkonian (1996). "The secretory pathway of protists: spatial and functional organization and evolution." Microbiol Rev **60**(4): 697-721.

Bennett, M. K., N. Calakos and R. H. Scheller (1992). "Syntaxin: a synaptic protein implicated in docking of synaptic vesicles at presynaptic active zones." Science **257**(5067): 255-9.

Bennett, M. K., J. E. Garcia-Arraras, et al. (1993). "The syntaxin family of vesicular transport receptors." Cell **74**(5): 863-73.

Bennett, M. K. and R. H. Scheller (1993). "The molecular machinery for secretion is conserved from yeast to neurons." Proc Natl Acad Sci U S A **90**(7): 2559-63.

Bevis, B. J., A. T. Hammond, C. A. Reinke and B. S. Glick (2002). "De novo formation of transitional ER sites and Golgi structures in *Pichia pastoris*." Nat Cell Biol **4**(10): 750-6.

Bezprozvanny, I., P. Zhong, R. H. Scheller and R. W. Tsien (2000). "Molecular determinants of the functional interaction between syntaxin and N-type Ca^{2+} channel gating." Proc Natl Acad Sci U S A **97**(25): 13943-8.

Biderre, C., G. Metenier and C. P. Vivares (1998). "A small spliceosomal-type intron occurs in a ribosomal protein gene of the microsporidian *Encephalitozoon cuniculi*." Mol Biochem Parasitol **94**(2): 283-6.

Bock, J. B., H. T. Matern, A. A. Peden and R. H. Scheller (2001). "A genomic perspective on membrane compartment organization." Nature **409**(6822): 839-41.

Bogdanovic, A., F. Bruckert, T. Morio and M. Satre (2000). "A syntaxin 7 homologue is present in *Dictyostelium discoideum* endosomes and controls their homotypic fusion." J Biol Chem **275**(47): 36691-7.

Bolivar, I., J. F. Fahrni, A. Smirnov and J. Pawlowski (2001). "SSU rRNA-based phylogenetic position of the genera *Amoeba* and *Chaos* (Lobosea, Gymnamoebia): the origin of gymnamoebae revisited." Mol Biol Evol **18**(12): 2306-14.

Brugerolle, G. (1991). "Flagellar and cytoskeletal systems in amitochondriate flagellates: Archamoeba, Metamonada and Parabasala." Protoplasma **164**: 70-90.

Bryant, N. J. and T. H. Stevens (1998). "Vacuole biogenesis in *Saccharomyces cerevisiae*: protein transport pathways to the yeast vacuole." Microbiol Mol Biol Rev **62**(1): 230-47.

Buhse, H. E., S. J. Stamler and H. E. Smith (1975). "Protracted maintenance of symbiotic polymastigote flagellates outside their termite host." J. Protozool **22**: 11A-12A.

Bush, J., K. Franek, J. Daniel, G. B. Spiegelman, G. Weeks and J. Cardelli (1993). "Cloning and characterization of five novel *Dictyostelium discoideum* rab-related genes." Gene **136**(1-2): 55-60.

Carlile, M. J. (1982). "Prokaryotes and eukaryotes: strategies and successes." Trends in Biochemical Science **7**: 128-130.

Carter, R. F. (1970). "Description of a *Naegleria* sp. isolated from two cases of primary amoebic meningo-encephalitis, and of the experimental pathological changes induced by it." J Pathol **100**(4): 217-44.

Cavalier-Smith, T. (1981). "Eukaryote kingdoms: seven or nine?" Biosystems **14**(3-4): 461-81.

Cavalier-Smith, T. (1983). A 6-kingdom classification and a unified phylogeny. Endocytobiology II. H. E. A. Schenk and W. Schwemmler. Berlin, Walter de Gruyter: 1027-1034.

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

Cavalier-Smith, T. (1987). "The origin of eukaryote and archaebacterial cells." Ann. New York Acad. Sci. **503**: 17-54.

Cavalier-Smith, T. (1987). The origin of Fungi and pseudo-fungi. Evolutionary Biology of the Fungi. A. D. M. Rayner, C. M. Brasier and D. M. Moore. Cambridge, Cambridge University Press: 339-353.

Cavalier-Smith, T. (1993). "Kingdom Protozoa and its 18 phyla." Microbiol. Rev. **57**(4): 953-994.

Cavalier-Smith, T. (1997). "Amoeboflagellates and Mitochondrial Cristae in Eukaryote Evolution : Megasystematics of the New Protozoan Subkingdoms Eozoa and Neozoa." Arch. Protistenk. **147**: 237-258.

Cavalier-Smith, T. (1998). "A revised six-kingdom system of life." Biol Rev Camb Philos Soc **73**(3): 203-66.

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

Cavalier-Smith, T. (2000). Flagellate megaevolution. The basis for eukaryote diversification. The Flagellates. B. S. C. Leadbeater and J. C. Green. London, Taylor and Francis. **The Systematics Association Special Volume ser 59:** 361-390.

Cavalier-Smith, T. (2002). "The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa." Int J Syst Evol Microbiol **52**(Pt 2): 297-354.

Cavalier-Smith, T. and E. E. Chao (1996). "Sarcomonad ribosomal RNA sequences, rhizopod phylogeny, and the origin of euglyphid amoebae." Arch. Protistenkd. **147**: 227-236.

Chow, V. T., M. K. Sakharkar, D. P. Lim and W. M. Yeo (2001). "Phylogenetic relationships of the seven coat protein subunits of the coatomer complex, and comparative sequence analysis of murine xenin and proxenin." Biochem Genet **39**(5-6): 201-11.

Clark, C. G. and A. J. Roger (1995). "Direct evidence for secondary loss of mitochondria in *Entamoeba histolytica*." Proc Natl Acad Sci U S A **92**(14): 6518-21.

Cleveland, L. R. (1956). "Brief account of the sexual sycles of the flagellates of *Cryptocercus*." Journal of Protozoology **3**(4): 161-180.

Dacks, J. and A. J. Roger (1999). "The first sexual lineage and the relevance of facultative sex." J Mol Evol **48**(6): 779-783.

Dacks, J. B. (1998). Phylogenetics and the origin of sex. Department of Zoology. Vancouver, B.C., University of British Columbia**:** 122.

Dacks, J. B. and W. F. Doolittle (2001). "Reconstructing/Deconstructing the earliest eukaryotes. How comparative genomics can help." Cell **107**(4): 419-25.

Dacks, J. B. and W. F. Doolittle (2002). "Novel syntaxin gene sequences from *Giardia*, *Trypanosoma* and algae: implications for the ancient evolution of the eukaryotic endomembrane system." J Cell Sci **115**(Pt 8): 1635-42.

Dacks, J. B., A. Marinets, W. Ford Doolittle, T. Cavalier-Smith and J. M. Logsdon, Jr. (2002). "Analyses of RNA Polymerase II genes from free-living protists: phylogeny, long branch attraction, and the eukaryotic big bang." Mol Biol Evol **19**(6): 830-40.

Dacks, J. B., J. D. Silberman, et al. (2001). "Oxymonads are closely related to the excavate taxon *Trimastix*." Mol Biol Evol **18**(6): 1034-1044.

Dawson, S. C. and N. R. Pace (2002). "Novel kingdom-level eukaryotic diversity in anoxic environments." Proc Natl Acad Sci U S A **99**(12): 8324-9.

Delwiche, C. F. (1999). "Tracing the thread of plastid diversity through the tapestry of life." Am Nat **154**(S4): S164-S177.

Denny, P. W., S. Lewis, et al. (2002). "*Leishmania* RAB7: characterisation of terminal endocytic stages in an intracellular parasite." Mol Biochem Parasitol **123**(2): 105-13.

Dessen, P., M. Zagulski, et al. (2001). "*Paramecium* genome survey: a pilot project." Trends Genet **17**(6): 306-8.

Donaldson, J. G. and J. Lippincott-Schwartz (2000). "Sorting and signaling at the Golgi complex." Cell **101**(7): 693-6.

Doolittle, W. F. (1998). "You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes." Trends in Genetics **14**(8): 307-311.

Doolittle, W. F. (1999). "Phylogenetic classification and the universal tree." Science **284**(5423): 2124-9.

Duboscq, O. and P. P. Grassé (1925). "Note sur les protistes des termites de France. IV. Appareil de Golgi, Mitochondries et vesicules sous-flagellaires de *Pyrsonympha vertens*." C. R. Soc. Biol. **93**: 345-347.

Dulubova, I., T. Yamaguchi, et al. (2003). "Convergence and divergence in the mechanism of SNARE binding by Sec1/Munc18-like proteins." Proc Natl Acad Sci U S A **100**(1): 32-7.

Edgcomb, V. P., A. G. Simpson, et al. (2002). "Pelobionts are degenerate protists: insights from molecules and morphology." Mol Biol Evol **19**(6): 978-82.

Edlind, T. D., J. Li, G. S. Visvesvara, M. H. Vodkin, G. L. McLaughlin and S. K. Katiyar (1996). "Phylogenetic analysis of beta-tubulin sequences from amitochondrial protozoa." Mol Phylogenet Evol **5**(2): 359-67.

Edwardson, J. M. (1998). "Membrane fusion: all done with SNAREpins?" Curr Biol **8**(11): R390-3.

Embley, T. M. and R. P. Hirt (1998). "Early branching eukaryotes?" Curr Opin Genet Dev **8**(6): 624-9.

Enserink, M. (2002). "Microbial genomics. TIGR begins assault on the anthrax genome." Science **295**(5559): 1442-3.

Fasshauer, D., R. B. Sutton, A. T. Brunger and R. Jahn (1998). "Conserved structural features of the synaptic fusion complex: SNARE proteins reclassified as Q- and R-SNAREs." Proc Natl Acad Sci U S A **95**(26): 15781-6.

Fast, N. M., J. C. Kissinger, D. S. Roos and P. J. Keeling (2001). "Nuclear-encoded, plastid-targeted genes suggest a single common origin for apicomplexan and dinoflagellate plastids." Mol Biol Evol **18**(3): 418-26.

Fast, N. M., J. M. Logsdon, Jr. and W. F. Doolittle (1999). "Phylogenetic analysis of the TATA box binding protein (TBP) gene from *Nosema locustae*: evidence for a microsporidia-fungi relationship and spliceosomal intron loss." Mol Biol Evol **16**(10): 1415-9.

Fast, N. M., A. J. Roger, C. A. Richardson and W. F. Doolittle (1998). "U2 and U6 snRNA genes in the microsporidian *Nosema locustae*: evidence for a functional spliceosome." Nucleic Acids Res **26**(13): 3202-7.

Felsenstein, J. (1978). "Cases in which parsimony or compatibility methods will be positively misleading." Systematic Zoology **25**: 401-410.

Felsenstein, J. (1995). PHYLIP(Phylogeny Inference Package). Seattle, Department of Genetics, University of Washington.

Field, M. C. and J. C. Boothroyd (1995). "*Trypanosoma brucei*: molecular cloning of homologues of small GTP-binding proteins involved in vesicle trafficking." Exp Parasitol **81**(3): 313-20.

Foran, P., G. W. Lawrence, C. C. Shone, K. A. Foster and J. O. Dolly (1996). "Botulinum neurotoxin C1 cleaves both syntaxin and SNAP-25 in intact and permeabilized chromaffin cells: correlation with its blockade of catecholamine release." Biochemistry **35**(8): 2630-6.

Forget, L., J. Ustinova, Z. Wang, V. A. Huss and B. F. Lang (2002). "*Hyaloraphidium curvatum*: a linear mitochondrial genome, tRNA editing, and an evolutionary link to lower fungi." Mol Biol Evol **19**(3): 310-9.

Friedman, R. and A. L. Hughes (2001). "Gene duplication and the structure of eukaryotic genomes." Genome Res **11**(3): 373-81.

Gajadhar, A. A., W. C. Marquardt, R. Hall, J. Gunderson, E. V. Ariztia-Carmona and M. L. Sogin (1991). "Ribosomal RNA sequences of *Sarcocystis muris*, *Theileria annulata* and *Crypthecodinium cohnii* reveal evolutionary relationships among apicomplexans, dinoflagellates, and ciliates." Mol Biochem Parasitol **45**(1): 147-54.

Gardner, M. J., N. Hall, et al. (2002). "Genome sequence of the human malaria parasite *Plasmodium falciparum*." Nature **419**(6906): 498-511.

Germot, A., H. Philippe and H. Le Guyader (1997). "Evidence for the loss of mitochondria in Microsporidia from a mitochondrial-type HSP70 in *Nosema locustae*." Mol. Biochem. Parasitol. **87**(2): 159-168.

Ghislain, M., R. J. Dohmen, F. Levy and A. Varshavsky (1996). "Cdc48p interacts with Ufd3p, a WD repeat protein required for ubiquitin-mediated proteolysis in *Saccharomyces cerevisiae*." Embo J **15**(18): 4884-99.

Ghosh, S. K., J. Field, et al. (1999). "Chitinase secretion by encysting *Entamoeba invadens* and transfected *Entamoeba histolytica* trophozoites: localization of secretory vesicles, endoplasmic reticulum, and Golgi apparatus." Infect Immun **67**(6): 3073-81.

Gogarten, J. P., W. F. Doolittle and J. G. Lawrence (2002). "Prokaryotic evolution in light of gene transfer." Mol Biol Evol **19**(12): 2226-38.

Grassé, P. P. (1952). Ordre des Oxymonadines. Traité de Zoologie. Paris, Masson et Cie. **1:** 802-823.

Grassé, P. P. (1952). Ordre des Pyrsonymphines. Traité de Zoologie. Paris, Masson + Cie: 789-800.

Grosovsky, B. D. D. and L. Margulis (1982). Termite Microbial Communities. Experimental Microbial Ecology. R. G. Burns and J. H. Slater. Oxford, Blackwell Scientific: 519-532.

Gunderson, J. H., S. J. Goss and M. L. Sogin (1994). "The sequence of the *Hartmannella vermiformis* small subunit rRNA coding region." J Eukaryot Microbiol **41**(5): 481-2.

Gupta, R. S. and G. B. Golding (1996). "The origin of the eukaryotic cell." Trends Biochem Sci **21**(5): 166-71.

Haeckel, E. H. P. A. (1866). Generelle Morphologie der Organismen allgemeine Grundzüge der organischen Formen-Wissenschaft : mechanisch begründet durch die von Charles Darwin reformirte Descendenz-Theorie. Berlin, Reimer.

Haft, C. R., M. de la Luz Sierra, R. Bafford, M. A. Lesniak, V. A. Barr and S. I. Taylor (2000). "Human orthologs of yeast vacuolar protein sorting proteins Vps26, 29, and 35: assembly into multimeric complexes." Mol Biol Cell **11**(12): 4105-16.

Hashimoto, T., Y. Nakamura, et al. (1994). "Protein phylogeny gives a robust estimation for early divergences of eukaryotes: phylogenetic place of a mitochondria-lacking protozoan, *Giardia lamblia*." Mol Biol Evol **11**(1): 65-71.

Hashimoto, T., L. B. Sanchez, T. Shirakura, M. Muller and M. Hasegawa (1998). "Secondary absence of mitochondria in *Giardia lamblia* and *Trichomonas vaginalis* revealed by valyl-tRNA synthetase phylogeny." Proc Natl Acad Sci U S A **95**(12): 6860-5.

Hatsuzawa, K., H. Hirose, K. Tani, A. Yamamoto, R. H. Scheller and M. Tagaya (2000). "Syntaxin 18, a SNAP receptor that functions in the endoplasmic reticulum, intermediate compartment, and *cis*-Golgi vesicle trafficking." J Biol Chem **275**(18): 13713-20.

Hay, J. C. (2001). "SNARE complex structure and function." Exp Cell Res **271**(1): 10-21.

Herskovits, A. A., E. Shimoni, A. Minsky and E. Bibi (2002). "Accumulation of endoplasmic membranes and novel membrane-bound ribosome-signal recognition particle receptor complexes in Escherichia coli." J Cell Biol **159**(3): 403-10.

Hillis, D. M., C. Moritz and B. K. Mable (1996). Molecular systematics. Sunderland, Mass., Sinauer Associates.

Hinkle, G., D. D. Leipe, T. A. Nerad and M. L. Sogin (1994). "The unusually long small subunit ribosomal RNA of *Phreatamoeba balamuthi*." Nucleic Acids Res **22**(3): 465-9.

Hirt, R. P., J. M. Logsdon, Jr., B. Healy, M. W. Dorey, W. F. Doolittle and T. M. Embley (1999). "Microsporidia are related to Fungi: evidence from the largest subunit of RNA polymerase II and other proteins." Proc Natl Acad Sci U S A **96**(2): 580-5.

Hollande, A. and J. Carruette-Valentin (1970). "Appariement chromosomique et complexes synaptonematiques dans les noyaux en cours de depolyploidisation chez *Pyrsonympha flagellata*: le cycle evolutif des Pyrsonymphines symbiontes de *Reticulitermes lucifugus*." Contes Rendues de L'Academie Scientifique de Paris **270**: 2550-2555.

Holthuis, J. C., B. J. Nichols, S. Dhruvakumar and H. R. Pelham (1998). "Two syntaxin homologues in the TGN/endosomal system of yeast." Embo J **17**(1): 113-26.

Horner, D. S. and T. M. Embley (2001). "Chaperonin 60 phylogeny provides further evidence for secondary loss of mitochondria among putative early-branching eukaryotes." Mol Biol Evol **18**(10): 1970-5.

Jahn, R. and T. C. Sudhof (1999). "Membrane fusion and exocytosis." Annu Rev Biochem **68**: 863-911.

Janoo, R., A. Musoke, C. Wells and R. Bishop (1999). "A Rab1 homologue with a novel isoprenylation signal provides insight into the secretory pathway of *Theileria parva*." Mol Biochem Parasitol **102**(1): 131-43.

Jeffries, T. R., G. W. Morgan and M. C. Field (2002). "TbRAB18, a developmentally regulated Golgi GTPase from *Trypanosoma brucei*." Mol Biochem Parasitol **121**(1): 63-74.

Kaiser, C. and S. Ferro-Novick (1998). "Transport from the endoplasmic reticulum to the Golgi." Curr Opin Cell Biol **10**(4): 477-82.

Kasinsky, H. E., J. D. Lewis, J. B. Dacks and J. Ausio (2001). "Origin of H1 linker histones." Faseb J **15**(1): 34-42.

Katinka, M. D., S. Duprat, et al. (2001). "Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*." Nature **414**(6862): 450-3.

Katz, L. A. (2002). "Lateral gene transfers and the evolution of eukaryotes: theories and data." Int J Syst Evol Microbiol **52**(Pt 5): 1893-900.

Katz, L. A., E. A. Curtis, M. Pfunder and L. F. Landweber (2000). "Characterization of novel sequences from distantly related taxa by walking PCR." Mol Phylogenet Evol **14**(2): 318-21.

Keeling, P. J. (2001). "Foraminifera and cercozoa are related in actin phylogeny: two orphans find a home?" Mol Biol Evol **18**(8): 1551-7.

Keeling, P. J. and W. F. Doolittle (1996). "Alpha-tubulin from early-diverging eukaryotic lineages and the evolution of the tubulin family." Mol Biol Evol **13**(10): 1297-305.

Keeling, P. J., M. A. Luker and J. D. Palmer (2000). "Evidence from beta-tubulin phylogeny that microsporidia evolved from within the fungi." Mol Biol Evol **17**(1): 23-31.

Keeling, P. J. and G. I. McFadden (1998). "Origins of microsporidia." Trends Microbiol **6**(1): 19-23.

Keeling, P. J. and J. D. Palmer (2000). "Parabasalian flagellates are ancient eukaryotes." Nature **405**(6787): 635-7.

Kirby, H. J. (1934). Protozoa in termites. Termites and Termite Control. C. A. Kofoid. Berkeley, University of California Press: 84-93.

Kirchhausen, T. (2000). "Clathrin." Annu Rev Biochem **69**: 699-727.

Klumperman, J. (2000). "Transport between ER and Golgi." Curr Opin Cell Biol **12**(4): 445-9.

Koidzumi, M. (1921). "Studies on the intestinal protozoa found in the termites of Japan." Parasitology **13**: 235-309.

Kuzoff, R. K. and C. S. Gasser (2000). "Recent progress in reconstructing angiosperm phylogeny." Trends Plant Sci **5**(8): 330-6.

Lang, B. F., G. Burger, et al. (1997). "An ancestral mitochondrial DNA resembling a eubacterial genome in miniature." Nature **387**(6632): 493-7.

Lang, B. F., C. O'Kelly, T. Nerad, M. W. Gray and G. Burger (2002). "The closest unicellular relatives of animals." Curr Biol **12**(20): 1773-8.

Langford, T. D., J. D. Silberman, et al. (2002). "*Giardia lamblia*: identification and characterization of Rab and GDI proteins in a genome survey of the ER to Golgi endomembrane system." Exp Parasitol **101**(1): 13-24.

Latterich, M., K. U. Frohlich and R. Schekman (1995). "Membrane fusion and the cell cycle: Cdc48p participates in the fusion of ER membranes." Cell **82**(6): 885-93.

Lauber, M. H., I. Waizenegger, et al. (1997). "The *Arabidopsis* KNOLLE protein is a cytokinesis-specific syntaxin." J Cell Biol **139**(6): 1485-93.

Lee, J. J., G. F. Leedale and P. Bradbury, Eds. (2002). The Illustrated Guide to the Protozoa. Lawrence, Kansas, Society of Protozoologists.

Leidy, J. (1881). "The parasites of termites." J. Acad. Nat. Sci. Philadelphia (series 2) **8**: 425-447.

Lewis, M. J., J. C. Rayner and H. R. Pelham (1997). "A novel SNARE complex implicated in vesicle fusion with the endoplasmic reticulum." Embo J **16**(11): 3017-24.

Leyman, B., D. Geelen, F. J. Quintero and M. R. Blatt (1999). "A tobacco syntaxin with a role in hormonal control of guard cell ion channels." Science **283**(5401): 537-40.

Li, D. and R. Roberts (2001). "WD-repeat proteins: structure characteristics, biological function, and their involvement in human diseases." Cell Mol Life Sci **58**(14): 2085-97.

Linstedt, A. D. (1999). "Stacking the cisternae." Curr Biol **9**(23): R893-6.

Logsdon, J. M., Jr. (1998). "The recent origins of spliceosomal introns revisited." Curr Opin Genet Dev **8**(6): 637-48.

Lopez-Garcia, P. and D. Moreira (1999). "Metabolic symbiosis at the origin of eukaryotes." Trends Biochem Sci **24**(3): 88-93.

Lopez-Garcia, P., F. Rodriguez-Valera, C. Pedros-Alio and D. Moreira (2001). "Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton." Nature **409**(6820): 603-7.

Low, S. H., M. Miura, P. A. Roche, A. C. Valdez, K. E. Mostov and T. Weimbs (2000). "Intracellular redirection of plasma membrane trafficking after loss of epithelial cell polarity." Mol Biol Cell **11**(9): 3045-60.

Lujan, H. D., A. Marotta, M. R. Mowatt, N. Sciaky, J. Lippincott-Schwartz and T. E. Nash (1995). "Developmental induction of Golgi structure and function in the primitive eukaryote *Giardia lamblia*." J Biol Chem **270**(9): 4612-8.

Lyons-Weiler, J. and G. A. Hoelzer (1999). "Null model selection, compositional bias, character state bias, and the limits of phylogenetic information." Mol Biol Evol **16**: 1400-1406.

Lyons-Weiler, J., G. A. Hoelzer and R. J. Tausch (1996). "Relative apparent synapomorphy analysis (RASA). I: The statistical measurement of phylogenetic signal." Mol Biol Evol **13**(6): 749-57.

Ma, S., P. Fey and R. L. Chisholm (2001). "Molecular motors and membrane traffic in *Dictyostelium*." Biochim Biophys Acta **1525**(3): 234-44.

Mai, Z., S. Ghosh, M. Frisardi, B. Rosenthal, R. Rogers and J. Samuelson (1999). "Hsp60 is targeted to a cryptic mitochondrion-derived organelle ("crypton") in the microaerophilic protozoan parasite *Entamoeba histolytica*." Mol Cell Biol **19**(3): 2198-205.

Mair, G., H. Shi, et al. (2000). "A new twist in trypanosome RNA metabolism: *cis*-splicing of pre-mRNA." Rna **6**(2): 163-9.

Mallard, F., B. L. Tang, et al. (2002). "Early/recycling endosomes-to-TGN transport involves two SNARE complexes and a Rab6 isoform." J Cell Biol **156**(4): 653-64.

Maniatis, T., E. F. Fritsch and J. Sambrook (1982). Molecular cloning : a laboratory manual. Cold Spring Harbor, N.Y., Cold Spring Harbor Laboratory.

Margulis, L. (1970). Origin of eukaryotic cells; evidence and research implications for a theory of the origin and evolution of microbial, plant, and animal cells on the Precambrian earth. New Haven, Yale University Press.

Marti, M., Y. Li, E. M. Schraner, P. Wild, P. Kohler and A. B. Hehl (2003). "The secretory apparatus of an ancient eukaryote: Protein sorting to separate export pathways occurs prior to formation of transient Golgi-like compartments." Mol Biol Cell **14**: 1433-1447.

Martin, W. (1999). "A briefly argued case that mitochondria and plastids are descendants of endosymbionts, but that the nuclear compartment is not." Proc. R. Soc. Lond. B. **266**: 1387-1395.

Martin, W. and M. Muller (1998). "The hydrogen hypothesis for the first eukaryote." Nature **392**(6671): 37-41.

McArthur, A. G., H. G. Morrison, et al. (2000). "The *Giardia* genome project database." FEMS Microbiol Lett **189**(2): 271-3.

McNew, J. A., F. Parlati, et al. (2000). "Compartmental specificity of cellular membrane fusion encoded in SNARE proteins." Nature **407**(6801): 153-9.

McNew, J. A., T. Weber, D. M. Engelman, T. H. Sollner and J. E. Rothman (1999). "The length of the flexible SNAREpin juxtamembrane region is a critical determinant of SNARE-dependent fusion." Mol Cell **4**(3): 415-21.

Medlin, L., H. J. Elwood, S. Stickel and M. L. Sogin (1988). "The characterization of enzymatically amplified eukaryotic 16S-like rRNA-coding regions." Gene **71**(2): 491-9.

Misura, K. M., R. H. Scheller and W. I. Weis (2000). "Three-dimensional structure of the neuronal-Sec1-syntaxin 1a complex." Nature **404**(6776): 355-62.

Moir, D., S. E. Stewart, B. C. Osmond and D. Botstein (1982). "Cold-sensitive cell-division-cycle mutants of yeast: isolation, properties, and pseudoreversion studies." Genetics **100**(4): 547-63.

Moreira, D., H. Le Guyader and H. Phillippe (2000). "The origin of red algae and the evolution of chloroplasts." Nature **405**(6782): 69-72.

Moreira, D. and P. Lopez-Garcia (1998). "Symbiosis between methanogenic archaea and delta-proteobacteria as the origin of eukaryotes: the syntrophic hypothesis." J Mol Evol **47**(5): 517-30.

Moriya, S., J. B. Dacks, et al. (In Press). "Molecular phylogeny of three oxymonad genera: *Pyrsonympha*, *Dinenympha* and *Oxymonas*." J. Eukaryot. Microbiol.

Moriya, S., M. Ohkuma and T. Kudo (1998). "Phylogenetic position of symbiotic protist *Dinenympha exilis* in the hindgut of the termite *Reticulitermes speratus* inferred from the protein phylogeny of elongation factor 1 alpha." Gene **210**(2): 221-7.

Mullock, B. M., C. W. Smith, et al. (2000). "Syntaxin 7 is localized to late endosome compartments, associates with Vamp 8, and is required for late endosome-lysosome fusion." Mol Biol Cell **11**(9): 3137-53.

Muniz, M., C. Nuoffer, H. P. Hauri and H. Riezman (2000). "The Emp24 complex recruits a specific cargo molecule into endoplasmic reticulum-derived vesicles." J Cell Biol **148**(5): 925-30.

Nagai, H., J. C. Kagan, X. Zhu, R. A. Kahn and C. R. Roy (2002). "A bacterial guanine nucleotide exchange factor activates ARF on *Legionella* phagosomes." Science **295**(5555): 679-82.

Nakamura, N., A. Yamamoto, Y. Wada and M. Futai (2000). "Syntaxin 7 mediates endocytic trafficking to late endosomes." J Biol Chem **275**(9): 6523-9.

Nickel, W., T. Weber, J. A. McNew, F. Parlati, T. H. Sollner and J. E. Rothman (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**(22): 12571-6.

Nickrent, D. L., C. L. Parkinson, J. D. Palmer and R. J. Duff (2000). "Multigene phylogeny of land plants with special reference to bryophytes and the earliest land plants." Mol Biol Evol **17**(12): 1885-95.

Nixon, J. E., A. Wang, et al. (2002). "A spliceosomal intron in *Giardia lamblia*." Proc Natl Acad Sci U S A **99**(6): 3701-3705.

O'Kelly, C. J., M. A. Farmer and T. A. Nerad (1999). "Ultrastructure of *Trimastix pyriformis* (Klebs) Bernard et al.: similarities of *Trimastix* species with retortamonad and jakobid flagellates." Protist **150**(2): 149-62.

Orci, L., M. Stamnes, et al. (1997). "Bidirectional transport by distinct populations of COPI-coated vesicles." Cell **90**(2): 335-49.

Pamnani, V., T. Tamura, et al. (1997). "Cloning, sequencing and expression of VAT, a CDC48/p97 ATPase homologue from the archaeon *Thermoplasma acidophilum*." FEBS Lett **404**(2-3): 263-8.

Parlati, F., T. Weber, J. A. McNew, B. Westermann, T. H. Sollner and J. E. Rothman (1999). "Rapid and efficient fusion of phospholipid vesicles by the alpha-helical core of a SNARE complex in the absence of an N-terminal regulatory domain." Proc Natl Acad Sci U S A **96**(22): 12565-70.

Patel, S. K., F. E. Indig, N. Olivieri, N. D. Levine and M. Latterich (1998). "Organelle membrane fusion: a novel function for the syntaxin homolog Ufe1p in ER membrane fusion." Cell **92**(5): 611-20.

Patterson, D. J. (1999). "The diversity of eukaryotes." The American Naturalist **154**: S96-S124.

Patterson, D. J., A. G. Simpson and A. Rogerson (2002). Amoebae of uncertain affinities. The Illustrated Guide to the Protozoa. J. J. Lee, G. F. Leedale and P. Bradbury. Lawrence, Kansas, Society of Protozoologists. **2:** 804-827.

Pennisi, E. (2002). "Genomics. Sequence tells mouse, human genome secrets." Science **298**(5600): 1863-5.

Peters, C., M. J. Bayer, S. Buhler, J. S. Andersen, M. Mann and A. Mayer (2001). "Trans-complex formation by proteolipid channels in the terminal phase of membrane fusion." Nature **409**(6820): 581-8.

Philippe, H. and A. Adoutte (1998). The molecular phylogeny of Eukaryota: solid facts and uncertainties. Evolutionary relationships among Protozoa. G. Coombs, K. Vickerman, M. Sleigh and A. Warren. London, Chapman & Hall: 25-56.

Philippe, H., A. Germot and D. Moreira (2000). "The new phylogeny of eukaryotes." Curr Opin Genet Dev **10**(6): 596-601.

Philippe, H., P. Lopez, et al. (2000). "Early-branching or fast-evolving eukaryotes? An answer based on slowly evolving positions." Proc R Soc Lond B Biol Sci **267**(1449): 1213-21.

Prekeris, R., J. Klumperman, Y. A. Chen and R. H. Scheller (1998). "Syntaxin 13 mediates cycling of plasma membrane proteins via tubulovesicular recycling endosomes." J Cell Biol **143**(4): 957-71.

Rabouille, C., H. Kondo, R. Newman, N. Hui, P. Freemont and G. Warren (1998). "Syntaxin 5 is a common component of the NSF- and p97-mediated reassembly pathways of Golgi cisternae from mitotic Golgi fragments *in vitro*." Cell **92**(5): 603-10.

Rabouille, C., T. P. Levine, J. M. Peters and G. Warren (1995). "An NSF-like ATPase, p97, and NSF mediate cisternal regrowth from mitotic Golgi fragments." Cell **82**(6): 905-14.

Rachel, R., I. Wyschkony, S. Riehl and H. Huber (2002). "The ultrastructure of *Ignicoccus*: Evidence for a novel outer membrane and for intracellular vesicle budding in an archeon." Archaea **1**(1): 9-18.

Rapoport, T. A., B. Jungnickel and U. Kutay (1996). "Protein transport across the eukaryotic endoplasmic reticulum and bacterial inner membranes." Annu Rev Biochem **65**: 271-303.

Reddy, J. V. and M. N. Seaman (2001). "Vps26p, a component of retromer, directs the interactions of Vps35p in endosome-to-Golgi retrieval." Mol Biol Cell **12**(10): 3242-56.

Robinson, M. S. and J. S. Bonifacino (2001). "Adaptor-related proteins." Curr Opin Cell Biol **13**(4): 444-53.

Roger, A. J. (1999). "Reconstructing early events in eukaryotic evolution." The American Naturalist **154**: S146-S163.

Roger, A. J., O. Sandblom, W. F. Doolittle and H. Philippe (1999). "An evaluation of elongation factor 1 alpha as a phylogenetic marker for eukaryotes." Mol Biol Evol **16**(2): 218-33.

Roger, A. J. and J. D. Silberman (2002). "Cell evolution: mitochondria in hiding." Nature **418**(6900): 827-9.

Roger, A. J., M. W. Smith, R. F. Doolittle and W. F. Doolittle (1996). "Evidence for the Heterolobosea from phylogenetic analysis of genes encoding glyceraldehyde-3-phosphate dehydrogenase." Journal of Eukaryotic Microbiology **43**(6): 475-85.

Roger, A. J., S. G. Svard, et al. (1998). "A mitochondrial-like chaperonin 60 gene in *Giardia lamblia*: Evidence that diplomonads once harbored an endosymbiont related to the progenitor of mitochondria." Proc. Natl. Acad. Sci. USA **95**: 229-234.

Rothman, J. E. and F. T. Wieland (1996). "Protein sorting by transport vesicles." Science **272**(5259): 227-34.

Roy, L., J. J. Bergeron, et al. (2000). "Role of p97 and syntaxin 5 in the assembly of transitional endoplasmic reticulum." Mol Biol Cell **11**(8): 2529-42.

Saito-Nakano, Y., M. Nakazawa, Y. Shigeta, T. Takeuchi and T. Nozaki (2001). "Identification and characterization of genes encoding novel Rab proteins from *Entamoeba histolytica*." Mol Biochem Parasitol **116**(2): 219-22.

Sanderfoot, A. A., F. F. Assaad and N. V. Raikhel (2000). "The *Arabidopsis* genome. An abundance of soluble N-ethylmaleimide-sensitive factor adaptor protein receptors." Plant Physiol **124**(4): 1558-69.

Sato, M. H., N. Nakamura, et al. (1997). "The AtVAM3 encodes a syntaxin-related molecule implicated in the vacuolar assembly in *Arabidopsis thaliana*." J Biol Chem **272**(39): 24530-5.

Schekman, R. and I. Mellman (1997). "Does COPI go both ways?" Cell **90**(2): 197-200.

Schledzewski, K., H. Brinkmann and R. R. Mendel (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**(6): 770-8.

Schroder-Kohne, S., F. Letourneur and H. Riezman (1998). "Alpha-COP can discriminate between distinct, functional di-lysine signals in vitro and regulates access into retrograde transport." J Cell Sci **111 (Pt 23)**: 3459-70.

Schulze, K. L., J. T. Littleton, et al. (1994). "rop, a *Drosophila* homolog of yeast Sec1 and vertebrate n-Sec1/Munc-18 proteins, is a negative regulator of neurotransmitter release in vivo." Neuron **13**(5): 1099-108.

Seaman, M. N., J. M. McCaffery and S. D. Emr (1998). "A membrane coat complex essential for endosome-to-Golgi retrograde transport in yeast." J Cell Biol **142**(3): 665-81.

Shimodaira, H. (2002). "An approximately unbiased test of phylogenetic tree selection." Syst Biol **51**(3): 492-508.

Shimodaira, H. and M. Hasegawa (2001). "CONSEL: for assessing the confidence of phylogenetic tree selection." Bioinformatics **17**(12): 1246-7.

Silberman, J. D., A. G. Simpson, et al. (2002). "Retortamonad flagellates are closely related to diplomonads--implications for the history of mitochondrial function in eukaryote evolution." Mol Biol Evol **19**(5): 777-86.

Simpson, A. G., E. K. MacQuarrie and A. J. Roger (2002). "Eukaryotic evolution: early origin of canonical introns." Nature **419**(6904): 270.

Simpson, A. G. and D. J. Patterson (2001). "On core jakobids and excavate taxa: the ultrastructure of *Jakoba incarcerata*." J Eukaryot Microbiol **48**(4): 480-92.

Simpson, A. G., R. Radek, J. B. Dacks and C. J. O'Kelly (2002). "How oxymonads lost their groove: an ultrastructural comparison of *Monocercomonoides* and excavate taxa." J Eukaryot Microbiol **49**(3): 239-48.

Simpson, A. G., A. J. Roger, et al. (2002). "Evolutionary history of "early-diverging" eukaryotes: the excavate taxon *Carpediemonas* is a close relative of *Giardia*." Mol Biol Evol **19**(10): 1782-91.

Simpson, A. G. B., C. Bernard and D. J. Patterson (2000). "The ultrastructure of *Trimastix marina*, an excavate flagellate." European Journal of Protistology **36**: 229-252.

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

Smith, H. E., S. J. Stamler and H. E. Buhse JR. (1975). "A scanning electron microscope survey of the surface features of polymastigote flagellates from *Reticulitermes flavipes*." Trans. Amer. Micros. Soc. **94**: 401-410.

Sogin, M. (1997). "History assignment: when was the mitochondrion founded?" Curr Opin Genet Dev **7**(6): 792-9.

Sogin, M. L. (1991). "Early evolution and the origin of eukaryotes." Cur. Op. Gen. Dev. **1**: 457-463.

Sokolova, Y., E. Snigirevskaya, E. Morzhina, S. Skarlato, A. Mironov and Y. Komissarchik (2001). "Visualization of early Golgi compartments at proliferate and sporogenic stages of a microsporidian *Nosema grylli*." J Eukaryot Microbiol **Suppl**: 86S-87S.

Sollner, T., S. W. Whiteheart, et al. (1993). "SNAP receptors implicated in vesicle targeting and fusion." Nature **362**(6418): 318-24.

Springer, S., A. Spang and R. Schekman (1999). "A primer on vesicle budding." Cell **97**(2): 145-8.

Stanier, R. (1970). Some aspects of the biology of cells and their possible evolutionary significance. Organization and Control in Prokaryotic and Eukaryotic Cells. H. Charles and B. Knight. Cambridge, Great Britain, Syndics of the Cambridge University Press: 1-38.

Stanier, R. Y., M. Douderoff and E. Adelberg (1963). The microbial world. Englewood Cliffs, N.J., Prentice-Hall.

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

Sterud, E., T. A. Mo and T. T. Poppe (1997). "Ultrastructure of *Spironucleus barkhanus* N. Sp. (Diplomonadida: Hexamitidae) from Grayling *Thymallus thymallus* (L.) (Salmonidae) and Atlantic Salmon *Salmo salar* L. (Salmonidae)." Journal of Eukaryotic Microbiology **44**(5): 399-407.

Stiller, J. W., E. C. Duffield and B. D. Hall (1998). "Amitochondriate amoebae and the evolution of DNA-dependent RNA polymerase II." Proc Natl Acad Sci U S A **95**(20): 11769-74.

Stiller, J. W. and B. D. Hall (1997). "The origin of red algae: implications for plastid evolution." Proc Natl Acad Sci U S A **94**(9): 4520-5.

Stiller, J. W. and B. J. Hall (1999). "Long-Branch Attraction and the rDNA Model of Early Eukaryotic Evolution." M.B.E. **16**(9): 1270-1279.

Stoltzfus, A. (1999). "On the possibility of constructive neutral evolution." J Mol Evol **49**(2): 169-81.

Strimmer, K. and A. von Haeseler (1997). Puzzle. Zoologisches Institut. Munich, Universitat Muenchen.

Sulli, C., Z. Fang, U. Muchhal and S. D. Schwartzbach (1999). "Topology of Euglena chloroplast protein precursors within endoplasmic reticulum to Golgi to chloroplast transport vesicles." J Biol Chem **274**(1): 457-63.

Sulli, C. and S. D. Schwartzbach (1995). "The polyprotein precursor to the *Euglena* light-harvesting chlorophyll a/b-binding protein is transported to the Golgi apparatus prior to chloroplast import and polyprotein processing." J Biol Chem **270**(22): 13084-90.

Swofford, D. L. (1998). PAUP*. Phylogenetic Analysis Using Parsimony (* and Other Methods). Sunderland, Massachusetts, Sinauer Associates.

Swofford, D. L., G. J. Olsen, P. J. Waddell and D. M. Hillis (1996). Phylogenetic Inference. Molecular Systematics. D. M. Hillis, C. Moritz and B. K. Mable. Sunderland, Mass., Sinauer Associates Inc.: 407-514.

Tachezy, J., L. B. Sanchez and M. Muller (2001). "Mitochondrial type iron-sulfur cluster assembly in the amitochondriate eukaryotes *Trichomonas vaginalis* and *Giardia intestinalis*, as indicated by the phylogeny of IscS." Mol Biol Evol **18**(10): 1919-28.

Tanigawa, G., L. Orci, M. Amherdt, M. Ravazzola, J. B. Helms and J. E. Rothman (1993). "Hydrolysis of bound GTP by ARF protein triggers uncoating of Golgi-derived COP-coated vesicles." J Cell Biol **123**(6 Pt 1): 1365-71.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin and D. G. Higgins (1997). "The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools." Nucleic Acids Res **25**(24): 4876-82.

Tovar, J., A. Fischer and C. G. Clark (1999). "The mitosome, a novel organelle related to mitochondria in the amitochondrial parasite *Entamoeba histolytica*." Mol Microbiol **32**(5): 1013-21.

Ungermann, C., A. Price and W. Wickner (2000). "A new role for a SNARE protein as a regulator of the Ypt7/Rab-dependent stage of docking." Proc Natl Acad Sci U S A **97**(16): 8889-91.

Ungermann, C., K. Sato and W. Wickner (1998). "Defining the functions of trans-SNARE pairs." Nature **396**(6711): 543-8.

Van de Peer, Y., A. Ben Ali and A. Meyer (2000). "Microsporidia: accumulating molecular evidence that a group of amitochondriate and suspectedly primitive eukaryotes are just curious fungi." Gene **246**(1-2): 1-8.

van den Ent, F., L. A. Amos and J. Lowe (2001). "Prokaryotic origin of the actin cytoskeleton." Nature **413**(6851): 39-44.

Volker, A., Y. D. Stierhof and G. Jurgens (2001). "Cell cycle-independent expression of the *Arabidopsis* cytokinesis-specific syntaxin KNOLLE results in mistargeting to

the plasma membrane and is not sufficient for cytokinesis." J Cell Sci **114**(Pt 16): 3001-12.

Vossbrinck, C. R., J. V. Maddox, S. Friedman, B. A. Debrunner-Vossbrinck and C. R. Woese (1987). "Ribosomal RNA sequence suggests microsporidia are extremely ancient eukaryotes." Nature **326**(6111): 411-4.

Wada, Y., N. Nakamura, Y. Ohsumi and A. Hirata (1997). "Vam3p, a new member of syntaxin related protein, is required for vacuolar assembly in the yeast *Saccharomyces cerevisiae*." J Cell Sci **110 (Pt 11)**: 1299-306.

Wang, H., L. Frelin and J. Pevsner (1997). "Human syntaxin 7: a Pep12p/Vps6p homologue implicated in vesicle trafficking to lysosomes." Gene **199**(1-2): 39-48.

Waters, M. G., T. Serafini and J. E. Rothman (1991). "'Coatomer': a cytosolic protein complex containing subunits of non-clathrin-coated Golgi transport vesicles." Nature **349**(6306): 248-51.

Wickner, W. and A. Haas (2000). "Yeast homotypic vacuole fusion: a window on organelle trafficking mechanisms." Annu Rev Biochem **69**: 247-75.

Wilihoeft, U., E. Campos-Gongora, S. Touzni, I. Bruchhaus and E. Tannich (2001). "Introns of *Entamoeba histolytica* and *Entamoeba dispar*." Protist **152**(2): 149-56.

Williams, B. A., R. P. Hirt, J. M. Lucocq and T. M. Embley (2002). "A mitochondrial remnant in the microsporidian *Trachipleistophora hominis*." Nature **418**(6900): 865-9.

Wolf, Y. I., F. A. Kondrashov and E. V. Koonin (2001). "Footprints of primordial introns on the eukaryotic genome: still no clear traces." Trends Genet **17**(9): 499-501.

Wolfe, K. H. and D. C. Shields (1997). "Molecular evidence for an ancient duplication of the entire yeast genome." Nature **387**(6634): 708-13.

Yamin, M. A. (1979). "Flagellates of the orders Trichomonadida Kirby, Oxymonadida Grasse, and Hypermastigida Grassi and Foa reported from lower termites (Isoptera families Mastotermitidae, Kalotermitidae, Hodotermitidae, Termopsidae, Rhinotermitidae and Serritermitidae) and from the wood-feeding Cryptocercus (Dictyoptera: Cryptocercidae)." Sociobiology **3**: 3-117.

Ye, Y., H. H. Meyer and T. A. Rapoport (2001). "The AAA ATPase Cdc48/p97 and its partners transport proteins from the ER into the cytosol." Nature **414**(6864): 652-6.

Yoon, H. S., J. D. Hackett, G. Pinto and D. Bhattacharya (2002). "The single, ancient origin of chromist plastids." Proc Natl Acad Sci U S A **99**(24): 15507-12.

Zerial, M. and H. McBride (2001). "Rab proteins as membrane organizers." Nat Rev Mol Cell Biol **2**(2): 107-17.

Zettler, L. A. A., T. A. Nerad, C. J. O'Kelly and M. L. Sogin (2001). "The nucleariid amoebae: more protists at the animal-fungal boundary." J Eukaryot Microbiol **48**(3): 293-7.

Zhang, X., A. Shaw, et al. (2000). "Structure of the AAA ATPase p97." Mol Cell **6**(6): 1473-84.