Chapter 10
The Emergence of Cellular Complexity at the Dawn of the Eukaryotes: Reconstructing the Endomembrane System with In Silico and Functional Analyses

Lila V. Koumandou and Mark C. Field

Abstract Eukaryotic cells depend on a complex network of intracellular organelles to perform endocytosis and exocytosis. These trafficking routes underlie many vital cellular processes, including nutrition, responses to environmental cues, defense from pathogens, and differentiation. Multiple disease mechanisms arise from defects in these pathways. How this complex system arose, especially when compared to the simpler trafficking systems of prokaryotes, remains largely unanswered. However, the availability of fully sequenced genomes from many diverse eukaryotic taxa and representing distinct lineages, increasingly facilitates the reconstruction of very early events in eukaryotic evolution. Studies based on comparative genomics and phylogenetics point to great complexity being already present in the last common ancestor of all eukaryotes and enriched with lineage-specific variability/flexibility. Here we describe the methodology and limitations behind such studies, how conclusions can be enhanced by functional analysis, as well as recent results relating to evolution of Rab small GTPases and the retromer complex.

10.1 Introduction

The endomembrane system of eukaryotic cells mediates uptake from the environment and transport of proteins, nutrients, and a variety of other molecules within the cell as well as release from the cell by secretion. The system comprises various

L.V. Koumandou
Department of Pathology, University of Cambridge, Tennis Court Road, Cambridge CB2 1QP, UK
Biomedical Research Foundation, Academy of Athens, Soranou Efesiou 4, 115 27 Athens, Greece
e-mail: koumandou@cantab.net

M.C. Field
Department of Pathology, University of Cambridge, Tennis Court Road, Cambridge CB2 1QP, UK

P. Pontarotti (ed.), *Evolutionary Biology – Concepts, Biodiversity, Macroevolution and Genome Evolution*, DOI 10.1007/978-3-642-20763-1_10,

organelles and membrane-bound transport intermediates, as well as recognition factors and effectors that direct traffic between the different compartments. Exocytosis can be thought of as starting at the endoplasmic reticulum where new proteins are synthesized and translocated across the ER membrane. From the ER, they are transported to the Golgi complex for post-translational modifications and packaged into vesicles that travel through the cell, and eventually fuse with the plasma membrane to release their contents to the outside of the cell. Conversely, endocytosis starts at the plasma membrane, where selected cargo is packaged into vesicles and trafficked to the endosomes, from where the cargo can either be rapidly recycled back to the plasma membrane, or proceed to the late endosome, the multivesicular body, and the lysosome for breakdown. Retrograde routes from the endosome to the Golgi complex, and from the Golgi complex to the ER also exist.

Common to all these routes are the processes of (a) cargo selection and vesicle formation, (b) vesicle transport along the cytoskeleton, and (c) recognition of the target compartment and vesicle fusion with the target membrane (Bonifacino and Glick 2004). Small GTPases of the Rab family mediate all these steps, and are crucial to orchestrating additional factors such as adaptins for cargo selection before vesicle budding, recruitment of the vesicle coat polymer, uncoating factors, effectors for interactions with the cytoskeleton, tethering factors for recognition of the target compartment, as well as SNARE proteins, which mediate membrane fusion for release of the vesicle's contents into the target compartment. Most of these proteins are members of large protein families, with paralogues restricted to specific cellular locations. For example, different members of the Rab GTPase family are restricted to specific cellular locations and responsible for specific trafficking routes (Stenmark and Olkkonen 2001). Members of the SNARE protein family and the adaptins are also restricted to specific compartments (Chen and Scheller 2001). Remarkably, the overall functions of orthologues appear to remain well conserved across the eukaryotes, so that, e.g., Rab11 is involved in recycling endocytic pathways in plants, mammals, chromalveolates, and excavates (Brighouse et al. 2010).

Distinct vesicle coats exist for endocytic vesicles (clathrin), for vesicles mediating ER to Golgi transport (COPII), and for vesicles traveling along the retrograde routes between the endosome and the Golgi (retromer) and the Golgi and ER (COPI). In a similar fashion to the Rabs, the vesicle coats (clathrin, COPI, and COPII) are related; the evolutionary history of these "protocoatomer" systems has yet to be fully elucidated and at present, some of the relationships are based on secondary structural conservation only (DeGrasse et al. 2009; Devos et al. 2004). Significantly, each coat system has a restricted cellular location (Devos et al. 2004). It is also possible that the protocoatomer has a direct prokaryotic origin as proteins with similar architectures have been reported in some bacterial lineages (Santarella-Mellwig et al. 2010). Introducing some variability on this theme, the tethering factors are protein complexes, each with a distinct cellular localization. Not all are members of the same protein family,

although several may be structurally related to protocoatomer (Koumandou et al. 2007; Nickerson et al. 2009).

This pattern of evolution of large protein families with specific localization to distinct cellular compartments, poses the question of whether the families expanded as new compartments arose with increasing eukaryotic complexity. However, the general pattern that has emerged to date, from a variety of studies, points to the early emergence of a complex eukaryotic cell, the components of which are shared among all extant eukaryotic lineages. This argument has two correlates, one pointing to an ancient and possibly rapid diversification of eukaryotic lineages, and the other to the universal conservation of most proteins involved in endocellular trafficking among all eukaryotic lineages.

As most functional studies of the endomembrane trafficking system are carried out in yeast and mammals, these provide only a limited sampling of eukaryotic diversity. To examine the early evolution of the system's complexity across all eukaryotes, a much wider sampling of organisms is necessary. The evolution of eukaryotic lineages has been examined by phylogenetic, phylogenomic, and morphological methods, with increasingly available genomic data allowing fine-tuning of the overall picture. Molecular phylogenies group eukaryotes into five major lineages: (1) the Opisthokonta, including the animals and fungi, (2) the Amoebozoa, (3) the Archaeplastida, (4) the Excavata, and (5) the SAR clade, which includes the Rhizaria, the Alveolates (ciliates, dinoflagellates, Apicomplexa), and the Stramenopiles (brown algae and diatoms amongst others). It remains to be established whether the Haptophyta and Cryptophyceae possibly also belong to the SAR group (Adl et al. 2005; Burki et al. 2007; Hackett et al. 2007; Keeling et al. 2005; Simpson and Roger 2004). Regardless, all of these groups are uniformly deep-branching, meaning that the steps toward the formation of the last eukaryotic common ancestor (LECA) are largely inaccessible to us by phylogenetic methods. However, reconstructing the LECA is possible to a large extent (Field and Dacks 2009).

Previous studies on the evolution of various protein families, such as Rabs (Dacks and Field 2004; Pereira-Leal 2008), vesicle coats (Devos et al. 2004), tethers (Koumandou et al. 2007), ESCRTs (Leung et al. 2008), and SNAREs (Dacks and Doolittle 2004), reveal a highly complex LECA with evidence for all major organelles and trafficking routes having become established before the diversification of the eukaryotic lineages. Such analyses can also identify earlier and later diverging members within each protein family, as well as secondary losses in lineages where certain factors were nonessential. Although protein family expansion is common within the Metazoa, a surprising level of diversity is also shared across all eukaryotic lineages, so that innovation is a general phenomenon. This underscores the need for comparative cell biology to fully understand the functionality present in these diverse lineages. Here we describe our general strategy used in these studies, and extend the analysis to some new results for the Rab protein system and the retromer complex.

10.2 Bioinformatic Workflow

While most molecular cell biology studies tend to focus on a restricted group of select organisms, especially yeast, mammals, and invertebrate metazoan models, assessment of the origins of eukaryotic cellular functional diversity and capacity depends on sampling the full diversity of organisms. Plenty of fully sequenced genomes are now available for opisthokonts, representing both classical model systems (e.g., *Saccharomyces cerevisiae, Schizosaccahromyces pombe, Drosophila melanogaster, Ceanorhabditis elegans*), and species which diverged earlier in the lineage (e.g., *Rhizopus orhyzae* in the fungi; *Nematostella vectensis* and *Monosiga brevicollis* for the Metazoa). Plant and algal diversity can be encompassed by including both multicellualr and unicellular representatives, as well as both red and green algae. For the excavates, the amoebozoa, and the SAR group (stramenopiles, alveolates, rhizaria), fully sequenced species represent vastly differing lifestyles, and only recently does the set of organisms with fully sequenced genomes begin to cover the true group diversity. Inevitably, the focus has been on organisms of economic or public health priority, with an obvious bias toward highly derived species that are frequently pathogenic; this is, however, being overcome as the cost of sequencing even large eukaryotic genomes has fallen, and hence both better and denser sampling is now apparent. Genomes for the cryptophyte *Guillardia theta* and the haptophyte *Emiliana huxleyi* have become available only very recently, and the phylogenetic position of these organisms is still under debate. Where possible, two or more taxa must be included from any supergroup to facilitate detection of secondary losses versus absence from an entire group and to minimize detection failure because of species-specific divergence or incompleteness in the database.

For comparative genomic analysis, genomic databases can be retrieved from the NCBI BLAST interface (http://www.ncbi.nlm.nih.gov/BLAST/), the Joint Genome Institute (JGI) (http://genome.jgi-psf.org/euk_cur1.html), the Broad Institute (http://www.broadinstitute.org/), the Sanger Institute (http://www.genedb.org/), EuPathdb (http://eupathdb.org/eupathdb/), as well as organism-specific BLAST servers, e.g., for *C. merolae* (http://merolae.biol.s.utokyo.ac.jp/blast/blast.html), *T. gondii* (http://www.toxodb.org/), *C. parvum* (http://www.cryptodb.org/cryptodb/), *Giardia intestinalis* (http://www.giardiadb.org/giardiadb/). In addition, predicted proteomes for most species can be downloaded by ftp from the respective database for local analysis.

The strategy for finding orthologous genes has a number of steps to ensure robustness of the results, and at present is heuristic (Fig. 10.1). Searches are done using protein sequences, as they are overall more highly conserved than nucleotide sequences for distantly related species, and avoid any effects from codon bias. In our approach, BLASTp searches are largely performed manually, and checked by hand, as expect value (E-value) cutoff thresholds can vary considerably between different organisms, and for different proteins, especially for large vs. small proteins or those that retain more restricted structural features. The BLOSUM62

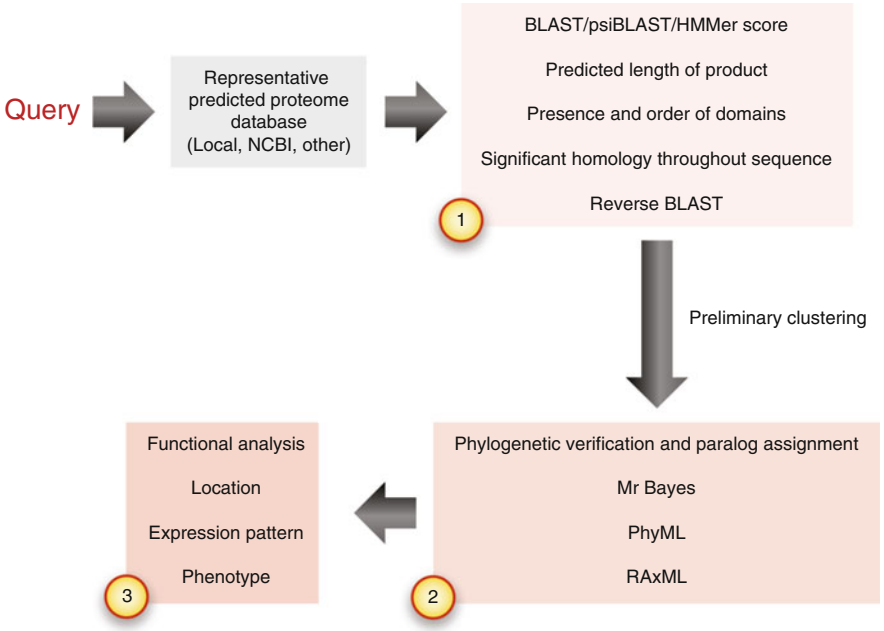

Fig. 10.1 An informatics workflow for comparative genomics. A query sequence is subjected to several tests to ensure that orthology is confidently predicted. These include a number of criteria designed to reduce miscalls due to regions of local similarity, and also frequently rely on the use of high-quality phylogenetic algorithms. See text for fuller discussion. Depending on the precise question being asked, progression from box 1 to box 2 or box 3 may not be required for fulfillment of an accurate call. The deeper intensity of the background color signifies the increased burden of moving from one box to the next

or BLOSUM45 substitution matrix is normally used, and, in general, E-values below e-3 are considered significant. Initial query sequences may be from yeast or human, and are used to search individually against each different organism; we find that this organism-by-organism approach leads to less overinterpretation when compared to a broader search. If the search identifies one clear hit, i.e., the top BLAST hit has an E-value much lower than all subsequent hits, then only the top hit is examined further. If the search identifies multiple top hits with similar E-values, then all top hits are examined further to determine if they represent paralogues. For example, small G proteins frequently generate multiple high-quality hits due to the conservation of the GTPase-binding site, necessitating further inspection for assignment (Fig. 10.1).

The candidate BLAST hits are subsequently tested by reverse BLAST, i.e., used to search the yeast or human proteome, or the nr database, and should return the original query or annotated orthologues from other species within the top five hits. In addition, the length of the putative orthologues should be similar to the original query, and any domains identified in the original query should also be conserved in the orthologue, which can easily be done by parsing through the NCBI conserved

domain database (CDD) or similar. We find this to be simple to perform, but frequently is an excellent discriminator; many candidates have only moderate support from E-value alone, but taken with conserved size and domain architecture, a good case can often be made for an evolutionary relationship.

Finally, further support is provided if the alignment between orthologues spans the full length of the sequence, and is not only concentrated in the conserved domain regions. This is particularly important for common domains, as the presence of the domain alone does not guarantee that the hit corresponds to a true orthologue; many domains are very highly conserved, providing respectable BLAST scores, but only taking account of comparatively restricted regions of the protein; if in doubt, a region of high homology can be removed from the sequence and the BLAST analysis rerun to ascertain if the remaining portions of the candidate have any relationship to the initial query. For analyses of many proteins (e.g., for the retromer cargo proteins presented here), the BLAST and reverse BLAST searches can be automated, e.g., with a BioPerl script that retrieves the BLAST results, and only records homologues if the reverse BLAST to the original query has an E-value better than a set threshold (e.g., e-3). However, examination of at least some of the results by hand is strongly advised, to check for length and domain agreement, as described above, and any negative results (not found) should be treated with caution and may need to be reexamined (e.g., with HMMer, see below).

In cases where no hits are retrieved by the original query, or no correspondence is found by reverse BLAST, or where the length and domain information are dubious, more detailed searches can be performed. One strategy is loosely termed "genome walking," i.e., using as an original query an orthologue from a closely related organism (e.g., using an *Arabidopsis* protein as query to search in *Chlamydomonas*). If this also fails, HMMer or PSI-BLAST can be used; these use the entire set of sequences for each protein family to generate a profile or consensus sequence and search based on that against any proteome in which BLAST did not recover a homologue. These are more sensitive than BLAST, and usually guaranteed to identify even highly divergent hits. The important downside here is that the sequences retrieved may have very weak sequence similarity and in fact lack a true evolutionary relationship – this is a particular issue when sequences contain coiled-coil regions, which are rather frequent in trafficking factors, and underscores the importance of manual curation of datasets.

In cases where all these attempts fail, or the results are rejected based on phylogeny (see below), the conclusion is that an orthologous protein is not found in this organism. This may be due to a true loss, due to the limits of detection of similarity search algorithms for highly divergent sequences, or due to sampling/misannotation errors in the available genome sequence. Examination of the genomic context of the gene may provide further clues here. For example, if the gene is within a syntenic region, where the order of genes is conserved between different species, one can examine whether the neighboring genes are conserved, although this is only of use when examining closely related taxa. If that is indeed the case, and the protein-coding sequence is absent, this is a powerful argument for secondary loss, which usually means that the gene's function was redundant or

nonessential. Presence or absence in closely related species can also be examined before secondary loss is invoked for a whole lineage. Conversely, if a protein is retained only in certain species or lineages where it might not be expected by parsimonious evolution (i.e., suggesting multiple independent acquisitions or losses), related factors can be examined to add robustness to the results; in some cases, examination for genes that contribute toward a given pathway can be helpful. For example, in the analysis of retromer cargo, the cation-independent mannose 6-phosphate receptor (CIMPR) is classically responsible for lysosomal delivery of proteins bearing the mannose-6-phosphate modification. As only some species outside the Metazoa were found to posses CIMPR, it was necessary to look for the presence of genes encoding the two enzymes required for mannose-6-phosphate modification, N-acetylglucosamine-1-phosphotransferase subunits α/β precursor (GNPTAB) and N-acetylglucosamine-1-phosphodiester α (NAGPA).

Finally, all candidate orthologues are examined by phylogeny to confirm orthology. Bayesian as well as maximum likelihood methods are used, including repeated sampling rounds to obtain posterior probability, and bootstrap support values, respectively. In cases where multiple orthologues are found in certain species, phylogeny can distinguish whether these are from ancient or recent gene duplications. Bayesian phylogeny can be run locally using Mr Bayes (http://mrbayes.csit.fsu.edu/), but for large datasets, a processor cluster is essential. Finally, maximum likelihood methods can be performed using remote web servers (http://www.hiv.lanl.gov/content/sequence/PHYML/interface.html and http://phylobench.vital-it.ch/raxml-bb/index.php provide good options), or can be run locally.

10.3 Results

One of the most important families of proteins involved in membrane trafficking are the Rab GTPases, and these proteins have received considerable attention (Stenmark 2009). They function to coordinate the actions of vesicle budding, targeting and fusion, and interact with a large number of proteins (Lee et al. 2009). As Ras-like GTPases, their intrinsic enzymatic activity is poor and hence hydrolysis of GTP requires the intervention of a GTPase activating protein (GAP). Rabs constitute a large family, with over 70 in *H. sapiens* and over 300 in *T. vaginalis*. As this topic has been reviewed extensively recently (Brighouse et al. 2010; Elias 2010), we will focus on a few specific issues here and the reader is referred elsewhere for a broader perspective.

A major goal has been the derivation of a Rab phylogeny (Pereira-Leal and Seabra 2001). As Rab orthologues are almost always associated with the same organelle, even across large evolutionary distances, these proteins conceptually provide an atlas of the compartments present, and critically this makes such information accessible for organisms that are hard to analyze experimentally for technical reasons. Hence, being able to determine the Rab complement in any lineage would provide an extremely valuable insight into the structure of the

endomembrane system. Further, understanding how Rab proteins have evolved would also solve the problem of when specific compartments arose, how they were expanded, and how they were lost. Additionally, such an analysis avoids the asymmetry problem, whereby it is easy to find conservation or secondary losses in divergent taxa, but due to significantly less direct experimental data, identification of true novel features in divergent lineages is rather more challenging.

Deriving a robust Rab phylogeny is more complex than might appear, as the database is misannotated, Rabs are small proteins (~250 amino acids) with a combination of highly variable C-termini and extremely well-conserved GTP-binding regions, plus the dataset is huge. This combination of factors has confounded many attempts to provide high-resolution phylogenies as there are too few informative character states for accurate resolution. However, a new method, "ScrollSaw," that essentially examines subsets of sequence data, and then combines these has been derived, which facilitates great improvement over traditional methods (Elias et al. manuscript in preparation). What is surprising is that the reconstruction predicts a considerable Rab complement in LECA, but which is consistent with the emerging view of great complexity in this organism (Fig. 10.2). This also indicates that secondary loss has played a significant role in evolution of the Rab protein family.

As a means to validate this conclusion, we have also performed detailed analysis of the TBC (Tre-2, Bub2, Cdc16) domain Rab GAP family (Gaberet-Castello et al. manuscript in preparation). This family accounts for most known Rab GAPs (Pan et al. 2006), and was selected as it is paralagous and the TBC domain facilitates reliable identification. The number of TBC proteins encoded in the genomes of most organisms is similar to the number of Rabs. The specificity of most TBC GAPs is poorly defined (Barr and Lambright 2010; Will and Gallwitz 2001), raising the issue of how activity is regulated. Applying the ScrollSaw approach to TBC evolution, it is clear that innovation of TBC GAPs is complex. Again, a large cohort is predicted to be present in the LECA, but it is also clear that lineage-specific innovations postdate the LECA; it is remarkable that the TBC GAPs and Rab phylogenies achieve such an overall degree of congruence, which gives confidence that the conclusion is probably correct.

As an example of a smaller and more restricted system, we have also analyzed the retromer complex, which is involved in retrograde traffic from the endosome to the Golgi. In *S. cerevisiae* and mammalian cells, retromer comprises five subunits: a sorting nexin dimer (Vps5 and Vps17 in yeast, SNX1 and SNX2 in mammals) which mediates membrane binding via PX domains and senses membrane curvature via BAR domains, and a trimeric subcomplex formed of Vps26, Vps29, and Vps35, which is responsible for cargo selection.

Retromer mediates recycling of vacuolar hydrolase receptors in yeast and mammals (Seaman et al. 1998; Arighi et al. 2004; Mari et al. 2008), as well as trafficking of the polymeric immunoglobulin receptor (Verges et al. 2004), plasma membrane iron transporters (Strochlic et al. 2007), Wntless (Eaton 2008), and processing of the amyloid precursor protein (He et al. 2005). Recently, retromer was also implicated in clearance of apoptotic bodies (Chen et al. 2010) and trafficking from the mitochondria to the peroxisome (Braschi et al. 2010). Using the comparative

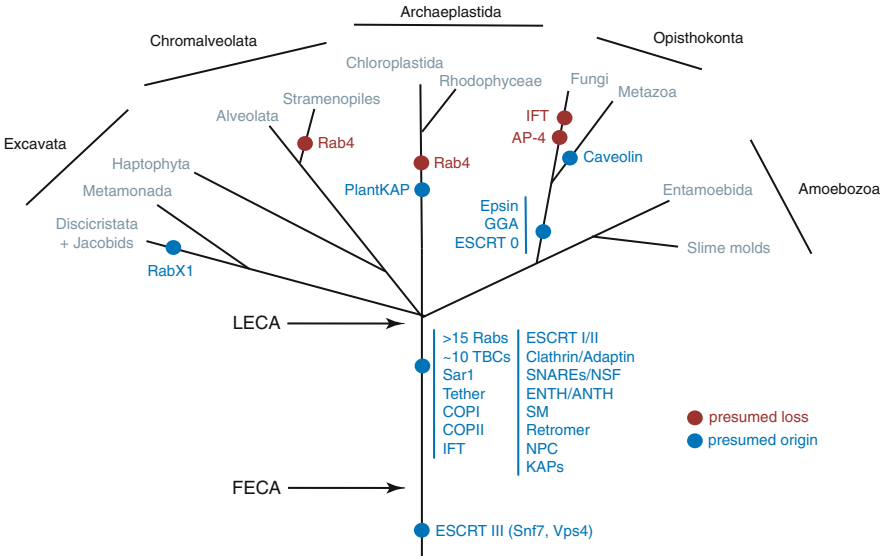

Fig. 10.2 Schematic eukaryotic phylogeny (sensu Adl) highlighting origins of trafficking components. Much of the basic bauplan for the eukaryotic cell predates the radiation of the eukaryotes, so that major coats, the factors required for vesicle specificity and the various control elements were all in place by the last eukaryotic common ancestor (LECA). The tree omits much detail, especially the origins and losses of a great many members of paralagous families from specific taxa, and also ignores any potential lateral gene transfer. *ANTH* AP180 N-Terminal Homology (ANTH) domain, *COP* coatomer, *ENTH* Epsin N-terminal homology (ENTH) domain, *ESCRT* endosomal sorting complex required for transport (a late endosomal membrane bending system also involved in cytokinesis), *FECA* first eukaryotic common ancestor (broadly equivalent to the eukaryogenesis event itself), *IFT* intraflagellar transport, *KAP* karyopherin (nucleocytoplasmic transport receptors), *NSF* NEM-sensitive factor (an ATPase that mediates SNARE protein complex disassembly), *NPC* nuclear pore complex or nucleoporins, *SM* Sec1/Munc18-like proteins (involved in SNARE-mediated vesicle fusion), *SNARE* SNAP (Soluble NSF attachment protein) receptors (coiled coil proteins required for vesicle fusion), *TBC* Tre-2, Bub2, Cdc16 domain (Rab GTPase activating proteins)

genomics workflow described above, we find that the Vps26/Vps29/Vps35 subcomplex is extremely well conserved. Interestingly, all cargo recognition subunits show expansions, with Vps26 the most widely expanded. Phylogenetics indicates that most expansions are species-specific (Koumandou et al. 2011).

The Vps5/Vps17 membrane-attachment subcomplex is also well conserved but less well than the cargo recognition complex. Vps17 is specific to the fungi, but SNX5/6 are probable functional analogues in Metazoa (Wassmer et al. 2007). Vps5 has been duplicated into SNX1/2 in Metazoa. Humans possess a total of 33 sorting nexins, fungi about eight, and non-Opisthokonta about four, indicating a huge and specific expansion in Metazoa. While speculative, this may reflect the huge complexity of endosomal systems in metazoan organisms where these trafficking systems perform important roles in cell–cell adhesion, signaling pathways, immune

defense, and development, which may require specific adaptors to traffic distinct cohorts of molecules through the late endosomal pathway.

Given the extremely good conservation of retromer throughout the eukaryotes, we also examined conservation of the retromer cargo. Vps10, the best characterized of this group, is a transmembrane lysosomal hydrolase receptor orthologous to mammalian sortilins (Mari et al. 2008). We find that Vps10 is broadly conserved with expansions in *Homo sapiens, Danio rerio, Nematostella vectensis, Monosiga brevicollis, Saccharomyces cerevisiae, Rhizopus oryzae,* and *Tetrahymena thermophila*. The *S. cerevisiae, R. oryzae*, and *T. thermophila* expansions are species-specific while metazoan-specific expansions have generated several distinct metazoan sortilin/Vps10 families. However, Vps10 is absent from several lineages, suggesting multiple secondary losses and, importantly, a role for retromer in sorting distinct sets of cargo in different organisms.

We therefore performed comparative genomics for the 14 previously reported retromer cargo proteins, as well as the retromer-interacting protein EHD1 (Gokool et al. 2007). Most studies of retromer and its cargo are from opisthokonts and indeed many of the reported cargo proteins are specific to the Metazoa, e.g., PIGR, EGFR, and Wntless, as they are involved in metazoan-specific signaling or immune defense. The vacuolar sorting receptor VSR1, originally identified in *A. thaliana* (Yamazaki et al. 2008), is restricted to the Archaeplastida. However, we also found widely distributed cargo, namely, EHD1, STE13, KEX2, and the FET3/FTR1 iron transporter. Importantly, all species lacking Vps10 contain at least one putative alternative cargo (Koumandou et al. 2011).

Sequence conservation alone does not immediately signify conservation of function. As part of our workplan, we use *Trypanosoma brucei* as a both accessible and highly divergent organism to facilitate comparisons with mammalian or other model systems. *In S. cerevisiae*, retromer mutants exhibit variable deficiencies with highly fragmented vacuoles in Vps5 and Vps17 mutants, moderate fragmentation in Vps26 mutants and no observable morphological defects for Vps29 and Vps35 mutants (Raymond et al. 1992). Mammalian SNX1/2 colocalize with endosomal markers EEA1 and Rab5, and with the mammalian Vps26-Vps29-Vps35 trimer (Haft et al. 2000; Kurten et al. 2001; Teasdale et al. 2001). Mouse SNX1 and SNX2 double knockouts arrest embryonic development, as do mutations in mammalian Vps26, and transcriptome analysis implicates Vps35 in Alzheimer's disease, underlining the importance of retromer to mammalian systems (Schwarz et al. 2002; Radice et al. 1991; Small et al. 2005).

In trypanosomes Vps5/Vps26/Vps29 and Vps35 mRNA expression is strongly upregulated in the mammalian form (Koumandou et al. 2008, 2011) where endocytosis is also more active (Natesan et al. 2007) and consistent with a role for retromer in endocytic activity. Retromer is also clearly essential as knockdowns of several subunits arrest cell proliferation (Koumandou et al. 2011). Trypanosome Vps26 and Vps5 exhibit both diffuse cytoplasmic localization plus distinct puncta located between the nucleus and kinetoplast, likely corresponding to endosomes (Field and Carrington 2009). TbVps26 partially colocalized with the clathrin heavy chain, and markers of the early and recycling endosomes; it was also proximal to

Vps28, which marks the multivesicular body (MVB), and to the lysosomal marker p67. Overall, these data indicate an endosomal location (Koumandou et al. 2011). Knockdowns also confirm a role in endosomal trafficking and result in a modest increase in p67 expression, representing a possible increase to lysosomal traffic via a block of retrograde retromer traffic from the endosome to the Golgi. Further, there is a decrease in intracellular levels of ISG75, a transmembrane protein that undergoes ubiquitin-dependent degradation (Chung et al. 2008; Leung et al. 2008), which is likely due to accelerated turnover. Finally, silencing Vps26 in mammalian cells results in fragmentation of the Golgi complex (Seaman 2004); a similar effect was found in Vps26-silenced trypanosomes (Koumandou et al. 2011). Together, the locations and effects of retromer subunit knockdowns suggest functional conservation between trypanosome, mammalian and yeast retromer.

10.4 Conclusions and Challenges

Understanding how the modern eukaryotic cell architecture arose, and was subsequently modified by differential selective pressures, has been a major goal in evolutionary cell biology. This is not solely of academic interest as many eukaryotic pathogens invest considerably in their cell surface as a host–pathogen interface and a site for immune evasion. An understanding of what such lineages have on board in terms of molecular components and function is a potentially potent weapon for combating infection and agricultural pathogens.

What is now very clear is the great complexity of LECA, and that this complexity encompasses many different families of proteins, adding confidence that this view is correct. We are also beginning to suspect that many extant organisms are in fact simpler than LECA with respect to their trafficking systems. This suggests that secondary losses, as well as paralogous expansions, are a major evolutionary driver responsible for the diversification of different lineages. A few factors appear to have been carried over from prokaryotic origins, but the majority are probably *de novo* innovations restricted to eukaryotes (Fig. 10.3).

Combined obstacles have made the elucidation of the evolutionary history of cellular functions a challenging task, and it remains incompletely addressed. In part, our view of some of the earliest events in eukaryogenesis is still very uncertain, but with the increase in genome sequencing, it is now possible to utilize molecular sequence data to address such problems. While still capable of generating equivocal or poorly supported models, with the inevitable controversies, such approaches have significant advantages. For the unwary, however, there remain some major pitfalls, which become even more pressing with increased size and complexity of datasets, and which necessitate the use of automated search algorithms.

We have developed a workflow that attempts to address at least some of these issues, and which is predicated on a reliance for heuristic analysis and the application of some biological principles. Such approaches are labor intensive and slower than fully automated approaches, but we consider them to be ultimately more accurate. The full workflow can incorporate functional studies, as we describe

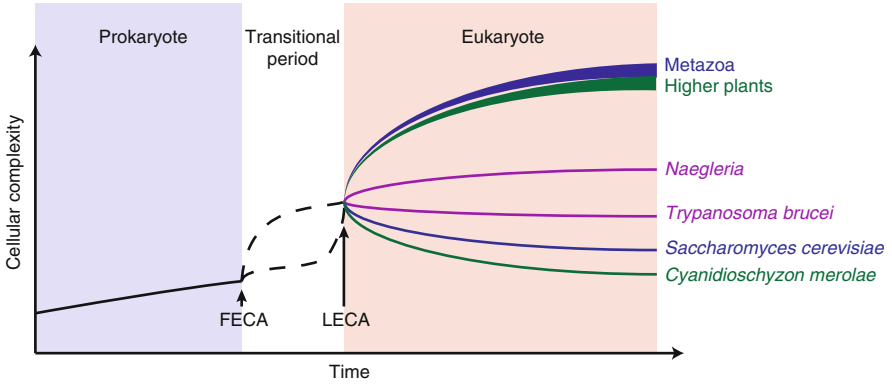

Fig. 10.3 Schematic for evolution of complexity in eukaryotic cells. Prokaryotic evolution (*blue*) proceeds in this model to increase complexity over time, but at the point of eukaryogenesis, equivalent to the first eukaryotic common ancestor (FECA, *arrow*), the acquisition of a nucleus propels the ability to increase complexity. Two extreme potential evolutionary trajectories are shown (*dotted lines*), where the initial event was followed by a period of rapid innovation (*top*), or a period of little innovation followed by rapidly increased complexity (*lower*). All trajectories between these extremes are possible, with the eventual arrival at the last eukaryotic common ancestor (LECA, *arrow*), which likely also represents an extreme bottleneck as all eukaryotes appear to radiate from a single lineage. Following eukaryotic radiation, many taxa evolved to increased complexity, of which the most potent examples are the Metazoa and higher plants. Some lineages appear to resemble the LECA in complexity, e.g., *Naegleria gruberi*, while many other taxa, including *Trypanosoma*, yeasts, and the extremophile red algae *C. merolae* have become less complex due to secondary losses of many components. Opisthokonts are in *blue*, Archaeplastida in *green*, and Excavata in *purple*. The diagram is heavily schematic and seeks to make general points only

here for retromer and several Rab GTPases. Such studies are complex, expensive, and prone to interpretive error as a great deal of interpretive expertise needs to be used, and one is frequently operating in organisms where the level of knowledge is sparse. However, such analysis can provide substantial support for *in silico* calls. For example, Tsg101/Vps23, an ESCRT complex subunit, in trypanosomes has very low similarity to the mammalian orthologue; knockdown and localization studies, however, confirm that the gene product plays a role in late endosomal transport, providing a strong argument that the *in silico* assignment is correct (Leung et al. 2008). Additionally, localization of SNARE proteins in trypanosomes has helped increase confidence in *in silico* assignments, which for these proteins can be difficult (Besteiro et al. 2006).

Several issues remain as challenges, of which at least three are paramount. First is the ongoing issue of asymmetry, whereby most *ab initio* identification of proteins involved in trafficking pathways is performed in a very small number of opisthokont taxa (Dacks and Field 2007). While using these organisms as a basis for comparative genomics will identify conserved elements and potential secondary losses or opisthokont lineage-specific innovations, by definition it fails to capture innovations in other supergroups. Analyses that focus on broader paralogous

families can go some way to solving this problem, and as more genome data becomes available, this may fade as a major issue. Second, search algorithms themselves are problematic. BLAST itself is rather insensitive, but using pattern recognition as implemented by HMMER or PSI-BLAST greatly increases the potential for false positives. Regardless, even these algorithms can fail to capture candidates, as demonstrated recently with a HMMER-based reconstruction of nuclear pore complex evolution – this is significantly better than BLAST alone, but still failed to identify many gene products (DeGrasse et al. 2009; Neumann et al. 2010). Third, sequence relationships are not the same as functional equivalence, which necessitates the expense and expertise required to gain direct functional insight. However, without such evidence, much valuable insight can be simply overlooked.

In summary, it is clear that the LECA was a complex organism. While some processes were likely inherited directly from prokaryotic predecessors, a spectacular level of innovation seems to have accompanied progression from the eukaryogenesis event itself to LECA.

Acknowledgments Many of these studies have been supported in part by the Wellcome Trust. We are also most grateful to collaborators and colleagues for discussions and access to unpublished data, and especially Carme Gabernet-Castello, Joel B. Dacks, Michael P. Rout, and Marek Elias.

References

Adl SM, Simpson AG, Farmer MA, Andersen RA, Anderson OR, Barta JR, Bowser SS, Brugerolle G, Fensome RA, Fredericq S, James TY, Karpov S, Kugrens P, Krug J, Lane CE, Lewis LA, Lodge J, Lynn DH, Mann DG, McCourt RM, Mendoza L, Moestrup O, Mozley-Standridge SE, Nerad TA, Shearer CA, Smirnov AV, Spiegel FW, Taylor MF (2005) The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. J Eukaryot Microbiol 52:399–451

Arighi CN, Hartnell LM, Aguilar RC, Haft CR, Bonifacino JS (2004) Role of the mammalian retromer in sorting of the cation-independent mannose 6-phosphate receptor. J Cell Biol 165:123–133

Barr F, Lambright DG (2010) Rab GEFs and GAPs. Curr Opin Cell Biol 22:461–470

Besteiro S, Coombs GH, Mottram JC (2006) The SNARE protein family of *Leishmania major*. BMC Genomics 7:250

Bonifacino JS, Glick BS (2004) The mechanisms of vesicle budding and fusion. Cell 116:153–166

Braschi E, Goyon V, Zunino R, Mohanty A, Xu L, McBride HM (2010) Vps35 mediates vesicle transport between the mitochondria and peroxisomes. Curr Biol 20:1310–1315

Brighouse A, Dacks JB, Field MC (2010) Rab protein evolution and the history of the eukaryotic endomembrane system. Cell Mol Life Sci 67:3449–3465

Burki F, Shalchian-Tabrizi K, Minge M, Skjaeveland A, Nikolaev SI, Jakobsen KS, Pawlowski J (2007) Phylogenomics reshuffles the eukaryotic supergroups. PLoS ONE 2:e790

Chen YA, Scheller RH (2001) SNARE-mediated membrane fusion. Nat Rev Mol Cell Biol 2:98–106

Chen D, Xiao H, Zhang K, Wang B, Gao Z, Jian Y, Qi X, Sun J, Miao L, Yang C (2010) Retromer is required for apoptotic cell clearance by phagocytic receptor recycling. Science 327:1261–1264

Chung WL, Leung KF, Carrington M, Field MC (2008) Ubiquitylation is required for degradation of transmembrane surface proteins in trypanosomes. Traffic 9:1681–1697

Dacks JB, Doolittle WF (2004) Molecular and phylogenetic characterization of syntaxin genes from parasitic protozoa. Mol Biochem Parasitol 136:123–136

Dacks JB, Field MC (2004) Eukaryotic cell evolution from a comparative genomic perspective: the endomembrane system. In: Hirt R, Horner D (eds) Organelles, genomes and eukaryote phylogeny: an evolutionary synthesis in the age of genomics. GRC Press, Boca Raton, pp 309–334

Dacks JB, Field MC (2007) Evolution of the eukaryotic membrane-trafficking system: origin, tempo and mode. J Cell Sci 120:2977–2985

DeGrasse JA, DuBois KN, Devos D, Siegel TN, Sali A, Field MC, Rout MP, Chait BT (2009) Evidence for a shared nuclear pore complex architecture that is conserved from the last common eukaryotic ancestor. Mol Cell Proteomics 8:2119–2130

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

Eaton S (2008) Retromer retrieves wntless. Dev Cell 14:4–6

Elias M (2010) Patterns and processes in the evolution of the eukaryotic endomembrane system. Mol Membr Biol 27:469–489

Field MC, Carrington M (2009) The trypanosome flagellar pocket. Nat Rev Microbiol 7:775–786

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

Gokool S, Tattersall D, Seaman MN (2007) EHD1 interacts with retromer to stabilize SNX1 tubules and facilitate endosome-to-Golgi retrieval. Traffic 8:1873–1886

Hackett JD, Yoon HS, Li S, Reyes-Prieto A, Rummele SE, Bhattacharya D (2007) Phylogenomic analysis supports the monophyly of cryptophytes and haptophytes and the association of rhizaria with chromalveolates. Mol Biol Evol 24:1702–1713

Haft CR, de la Luz SM, Bafford R, Lesniak MA, Barr VA, Taylor SI (2000) Human orthologs of yeast vacuolar protein sorting proteins Vps26, 29, and 35: assembly into multimeric complexes. Mol Biol Cell 11:4105–4116

He X, Li F, Chang WP, Tang J (2005) GGA proteins mediate the recycling pathway of memapsin 2 (BACE). J Biol Chem 280:11696–11703

Keeling PJ, Burger G, Durnford DG, Lang BF, Lee RW, Pearlman RE, Roger AJ, Gray MW (2005) The tree of eukaryotes. Trends Ecol Evol 20:670–676

Koumandou VL, Dacks JB, Coulson RM, Field MC (2007) Control systems for membrane fusion in the ancestral eukaryote; evolution of tethering complexes and SM proteins. BMC Evol Biol 7:29

Koumandou VL, Natesan SK, Sergeenko T, Field MC (2008) The trypanosome transcriptome is remodelled during differentiation but displays limited responsiveness within life stages. BMC Genomics 9:298

Koumandou VL, Klute MJ, Herman EK, Nunez-Miguel R, Dacks JB, Field MC (2011) Evolutionary reconstruction of the retromer complex and its function in *Trypanosoma brucei*. J Cell Sci 124:1496–1509

Kurten RC, Eddington AD, Chowdhury P, Smith RD, Davidson AD, Shank BB (2001) Self-assembly and binding of a sorting nexin to sorting endosomes. J Cell Sci 114:1743–1756

Lee MT, Mishra A, Lambright DG (2009) Structural mechanisms for regulation of membrane traffic by rab GTPases. Traffic 10:1377–1389

Leung KF, Dacks JB, Field MC (2008) Evolution of the multivesicular body ESCRT machinery; retention across the eukaryotic lineage. Traffic 9:1698–1716

Mari M, Bujny MV, Zeuschner D, Geerts WJ, Griffith J, Petersen CM, Cullen PJ, Klumperman J, Geuze HJ (2008) SNX1 defines an early endosomal recycling exit for sortilin and mannose 6-phosphate receptors. Traffic 9:380–393

Natesan SK, Peacock L, Matthews K, Gibson W, Field MC (2007) Activation of endocytosis as an adaptation to the mammalian host by trypanosomes. Eukaryot Cell 6:2029–2037
Neumann N, Lundin D, Poole AM (2010) Comparative genomic evidence for a complete nuclear pore complex in the last eukaryotic common ancestor. PLoS ONE 5:e13241
Nickerson DP, Brett CL, Merz AJ (2009) Vps-C complexes: gatekeepers of endolysosomal traffic. Curr Opin Cell Biol 21:543–551
Pan X, Eathiraj S, Munson M, Lambright DG (2006) TBC-domain GAPs for Rab GTPases accelerate GTP hydrolysis by a dual-finger mechanism. Nature 442:303–306
Pereira-Leal JB (2008) The Ypt/Rab family and the evolution of trafficking in fungi. Traffic 9:27–38
Pereira-Leal JB, Seabra MC (2001) Evolution of the Rab family of small GTP-binding proteins. J Mol Biol 313:889–901
Radice G, Lee JJ, Costantini F (1991) H beta 58, an insertional mutation affecting early postimplantation development of the mouse embryo. Development 111:801–811
Raymond CK, Howald-Stevenson I, Vater CA, Stevens TH (1992) Morphological classification of the yeast vacuolar protein sorting mutants: evidence for a prevacuolar compartment in class E vps mutants. Mol Biol Cell 3:1389–1402
Santarella-Mellwig R, Franke J, Jaedicke A, Gorjanacz M, Bauer U, Budd A, Mattaj IW, Devos DP (2010) The compartmentalized bacteria of the planctomycetes-verrucomicrobia-chlamydiae superphylum have membrane coat-like proteins. PLoS Biol 8:e1000281
Schwarz DG, Griffin CT, Schneider EA, Yee D, Magnuson T (2002) Genetic analysis of sorting nexins 1 and 2 reveals a redundant and essential function in mice. Mol Biol Cell 13:3588–3600
Seaman MN (2004) Cargo-selective endosomal sorting for retrieval to the Golgi requires retromer. J Cell Biol 165:111–122
Seaman MN, McCaffery JM, Emr SD (1998) A membrane coat complex essential for endosome-to-Golgi retrograde transport in yeast. J Cell Biol 142:665–681
Simpson AG, Roger AJ (2004) The real 'kingdoms' of eukaryotes. Curr Biol 14:R693–R696
Small SA, Kent K, Pierce A, Leung C, Kang MS, Okada H, Honig L, Vonsattel JP, Kim TW (2005) Model-guided microarray implicates the retromer complex in Alzheimer's disease. Ann Neurol 58:909–919
Stenmark H (2009) Rab GTPases as coordinators of vesicle traffic. Nat Rev Mol Cell Biol 10:513–525
Stenmark H, Olkkonen VM (2001) The Rab GTPase family. Genome Biol 2:REVIEWS3007
Strochlic TI, Setty TG, Sitaram A, Burd CG (2007) Grd19/Snx3p functions as a cargo-specific adapter for retromer-dependent endocytic recycling. J Cell Biol 177:115–125
Teasdale RD, Loci D, Houghton F, Karlsson L, Gleeson PA (2001) A large family of endosome-localized proteins related to sorting nexin 1. Biochem J 358:7–16
Verges M, Luton F, Gruber C, Tiemann F, Reinders LG, Huang L, Burlingame AL, Haft CR, Mostov KE (2004) The mammalian retromer regulates transcytosis of the polymeric immunoglobulin receptor. Nat Cell Biol 6:763–769
Wassmer T, Attar N, Bujny MV, Oakley J, Traer CJ, Cullen PJ (2007) A loss-of-function screen reveals SNX5 and SNX6 as potential components of the mammalian retromer. J Cell Sci 120:45–54
Will E, Gallwitz D (2001) Biochemical characterization of Gyp6p, a Ypt/Rab-specific GTPase-activating protein from yeast. J Biol Chem 276:12135–12139
Yamazaki M, Shimada T, Takahashi H, Tamura K, Kondo M, Nishimura M, Hara-Nishimura I (2008) *Arabidopsis* VPS35, a retromer component, is required for vacuolar protein sorting and involved in plant growth and leaf senescence. Plant Cell Physiol 49:142–156