Phylogeny of endocytic components yields insight into the process of nonendosymbiotic organelle evolution

Joel B. Dacks*†, Pak P. Poon‡, and Mark C. Field*

*Department of Pathology, University of Cambridge, Tennis Court Road, Cambridge CB2 1QP, United Kingdom; and ‡Departments of Microbiology and Immunology and Biochemistry and Molecular Biology, Dalhousie University, 5850 College Street, Halifax, NS, Canada B3H 1X5

Edited by Antonio Lazcano, Universidad Nacional Autónoma de México, Mexico D.F., Mexico, and accepted by the Editorial Board November 15, 2007 (received for review August 3, 2007)

The process by which some eukaryotic organelles, for example the endomembrane system, evolved without endosymbiotic input remains poorly understood. This problem largely arises because many major cellular systems predate the last common eukaryotic ancestor (LCEA) and thus do not provide examples of organelle genesis in progress. A model is emerging whereby gene duplication and divergence of multiple "specificity-" or "identity-" encoding proteins for the various endomembranous organelles produced the diversity of nonendosymbiotically derived cellular compartments present in modern eukaryotes. To address this possibility, we analyzed three molecular components of the endocytic membrane-trafficking machinery. Phylogenetic analyses of the endocytic syntaxins, Rab 5, and the β-adapts each reveal a pattern of ancestral, undifferentiated endocytic homologues in the LCEA. Subsequently, these undifferentiated progenitors independently duplicated in widely divergent lineages, convergently producing components with similar endocytic roles, e.g., β1 and β2-adaptin. In contrast, β3, β4, and all other adaptin complex subunits, as well as paralogues of the syntaxins and Rab specific for the other membrane-trafficking organelles, all evolved before the LCEA. Thus, the process giving rise to the differentiated organelles of the endocytic system appears to have been interrupted by the major speciation event that produced the extant eukaryotic lineages. These results suggest that although many endocytic components evolved before the LCEA, other major features evolved independently and convergently after diversification into the primary eukaryotic superfamilies. This finding provides an example of a basic cellular system that was simpler in the LCEA than in many extant eukaryotes and yields insight into nonendosymbiotic organelle evolution.

adaptin | Rab | SNARE | trafficking | autogenous

Eukaryotic cells are characterized by internal membrane-bound compartments, which include the nucleus, mitochondria, plastids, peroxisomes, and organelles of the membrane-trafficking system. By contrast, prokaryotes (eubacteria and archaeabacteria) generally lack such internal membrane-bound organelles. Based on molecular (1), and paleontological evidence (2), prokaryotes most likely preceded the eukaryotes, making autogenously encoded proteins for the various endomembranous organelles the likely components of the eukaryotic endomembrane system (ref. 15 and references therein). The system encompasses the endoplasmic reticulum (ER), the Golgi complex, the plasma membrane, and a variety of endocytic compartments. Movement between compartments is accomplished by packaging of cargo into vesicular or tubular membranous carriers and delivery by fusion of the carrier membrane with the target organelle. The protein factors required by each transport pathway include SNAREs, coat proteins, and Rabs (16). SNAREs are coiled-coil trans-membrane proteins that function as components of the fusion machinery and also as identifiers in providing membrane specificity (17, 18). Rabs, members of the Ras small GTPase superfamily, regulate membrane fusion and coordinate specificity among the multiple transport pathways and between the factors responsible for membrane fusion (19). Coat proteins participate in formation of carriers through membrane deformation, cargo selection, and other interactions (16).

Many factors involved in membrane trafficking are the products of gene duplications, with each member of the protein family performing similar functional roles but at distinct cellular locations (16). Comparative genomic and phylogenetic analysis of Rabs (20), SNAREs (21, 22), vesicle coats (23, 24), and several other trafficking components (25), sampling a broad diversity of eukaryotes, resolved these protein families into clades that correspond to organelle-specific paralogues. Importantly, each paralogue clade encompasses the full diversity of the taxa sampled, in many cases from five of the six eukaryotic supergroups (26), strongly implying that the major organelar paralogue families of Rabs (20), SNAREs (22), and vesicle coats (27) are at different stages in their transition from free-living organism to cellular compartment. However, some organelles appear to have evolved autogenously, from preexisting components in the protoeukaryote, without a significant contribution from endosymbiosis (7, 8). Understanding the process of autogenous organelle evolution is hampered by the lack of available intermediate forms because most basic eukaryotic cellular features appear established before the last common eukaryotic ancestor (LCEA) (9, 10). Incidents of proposed primitive simplicity in some eukaryotic lineages, with respect to mitochondria, peroxisomes, introns, and the Golgi complex, all now appear to be better explained via secondary loss (11–14), removing any potential transitional forms.

Organelles thought to have autogenous origins include those of the membrane-trafficking system (ref. 15 and references therein). The system encompasses the endoplasmic reticulum (ER), the Golgi complex, the plasma membrane, and a variety of endocytic compartments. Movement between compartments is accomplished by packaging of cargo into vesicular or tubular membranous carriers and delivery by fusion of the carrier membrane with the target organelle. The protein factors required by each transport pathway include SNAREs, coat proteins, and Rabs (16). SNAREs are coiled-coil trans-membrane proteins that function as components of the fusion machinery and also as identifiers in providing membrane specificity (17, 18). Rabs, members of the Ras small GTPase superfamily, regulate membrane fusion and coordinate specificity among the multiple transport pathways and between the factors responsible for membrane fusion (19). Coat proteins participate in formation of carriers through membrane deformation, cargo selection, and other interactions (16).

Many factors involved in membrane trafficking are the products of gene duplications, with each member of the protein family performing similar functional roles but at distinct cellular locations (16). Comparative genomic and phylogenetic analysis of Rabs (20), SNAREs (21, 22), vesicle coats (23, 24), and several other trafficking components (25), sampling a broad diversity of eukaryotes, resolved these protein families into clades that correspond to organelle-specific paralogues. Importantly, each paralogue clade encompasses the full diversity of the taxa sampled, in many cases from five of the six eukaryotic superfamilies (26), strongly implying that the major organelar paralogue families of Rabs (20), SNAREs (22), and vesicle coats (27)

Author contributions: J.B.D., P.P.P., and M.C.F. designed research; J.B.D. performed research; J.B.D. analyzed data; and J.B.D., P.P.P., and M.C.F. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. A.L. is a guest editor invited by the Editorial Board.

1To whom correspondence should be addressed. E-mail: jbd26@cam.ac.uk.

This article contains supporting information online at www.pnas.org/cgi/content/full/0707318105DC1.

© 2008 by The National Academy of Sciences of the USA
were established before the divergence of the existing eukaryotic groups, i.e., before the LCEA.

Incorporation of phylogenetic data has allowed for a molecular model of autogenous organellogenesis to emerge explaining how a single ancestral endomembrane, formed by invagination of the plasma membrane, differentiated into a variety of chemically and topologically distinct membranes (8, 15, 28). In this model, an undifferentiated endomembrane organelle, possessing identity-encoding proteins, would have evolved into two or more distinct organelles via duplications of the genes encoding the identity proteins [supporting information (SI) Fig. 6]. As the organelles replicate at cell division, and the protein duplicates diverge in sequence and function over many generations, novel organelle identity is acquired (SI Fig. 6). Various authors have proposed vesicle coats (8), SNAREs (22), Qa-SNAREs or syntaxins (21), small GTPases (29), and Sec1/Munc18 (SM) proteins (30) as the evolutionarily critical identity-encoding protein. However, because organelar identity is likely to depend on multiple factors, and each of these protein families share the same basic evolutionary pattern of paralogous expansion before the LCEA, it is probable that they all contributed to the origins and evolution of novel endomembrane compartments.

If organellogenesis requires more than one protein family, and if the duplications giving rise to the organellar specific paralogues were not concurrent, then there would be periods when some protein families of an emerging organelle have duplicated although others have not. If, at that point, the protoeukaryotic lineage underwent a radiative speciation event, i.e., the eukaryotic “big bang” (31), then those proteins that had already duplicated would resolve into organelle-specific clades encompassing all taxa, whereas those proteins associated with the same organelles or pathway-specific complexes that had not yet duplicated, would resolve into lineage-specific clades (SI Fig. 6 A and C) or else would retain undifferentiated versions active at multiple organelles.

Given that the basic complement of eukaryotic membrane-trafficking organelles and major protein families were already established in the LCEA (10, 13, 20, 22, 27, 32), we focus here on a more comprehensive and critical phylogenetic analysis of three protein families of the endocytic system. Essentially, the endocytic system is composed of two interconnected pathways: (i) recycling machinery returning receptors and plasma membrane components to the cell surface and (ii) degradative machinery transporting molecules to the digestive organelle, i.e., the lysosome or vacuole. We performed phylogenetic analysis of endocytic syntaxins, Rab 5, and the large subunits of adaptin proposed vesicle coats (8), SNAREs (22), Qa-SNAREs or syntaxins (21), small GTPases (29), and Sec1/Munc18 (SM) proteins (30) as the evolutionarily critical identity-encoding protein. Unlike the SynE proteins, the Rab5 corresponding to the recycling (Rab 5) and degradative pathways (Rab 7) form organelle- or pathway-specific clades that evolved before the LCEA (20), with clear phylogenetic separation of Rab 7 and Rab 5 (37). Rab 5 mediates fusion of vesicles to early endosomes (38, 39). Trypanosome Rab 5A and Rab 5B localize to separate endosomal compartments and associate with distinct cargos (39). Rab 5 paralogues in yeast also appear to possess distinct functions, although some redundancy likely exists (40), and similarly nonredundant functionality has been described for mammalian Rab 5 isoforms (41). Previous phylogenetic analyses of Rab 5 paralogues from these three taxon suggested that the last common ancestor of this gene in trypanosomes, yeast, and humans are derived from a single Rab5 gene that was then duplicated in these lineages (42). However, the analysis did not encompass diverse eukaryotic taxa or use methods accounting for rate variation, an important factor that can cause artifacts in phylogenetic reconstruction (31).

Sampling genomes of diverse taxon allowed us to build a more comprehensive Rab 5 dataset, which we subjected to phylogenetic analysis to identify and remove long-branch taxa (SI Fig. 9). Analysis of the resulting dataset shows that Rab 5 paralogues in vertebrates and in kinetoplastids (Fig. 3 and SI Fig. 10) are likely the result of independent lineage-specific duplications. Unlike the SynE proteins, the Rabs corresponding to the recycling (Rab 5) and degradative (Rab 7) pathways must have already been distinct genes very early in eukaryotic evolution and before the LCEA (20, 37). However, later lineage-specific functional expansion in the endocytic system appears to have continued for Rab 5 (SI Fig. 8B).
endocytic system (43). There are four separate complexes, each composed of four subunits (small, medium, and the large $\beta$- and $\gamma$-subunits). These complexes derive from duplications of genes ancestrally related to the F-COP subcomplex of coatamer (44).

The two large subunits are homologous, the product of an even more-ancient gene duplication (23). $\beta$-AP1 and $\beta$-AP2 may not be as simple as for the other AP complexes (23, 47). Whereas the large $\gamma$, the medium ($\delta$), and the small ($\epsilon$) subunits of AP1, AP2, AP3, and AP4 complexes resolve into discrete clades, the $\beta$-subunits form clear $\beta$AP3 and $\beta$AP4 clades but did not resolve separate $\beta$1 and $\beta$2 clades (23, 47). However, previous analyses did not include the full diversity of eukaryotes, and used methods susceptible to artifact (31). We therefore analyzed the phylogeny of the $\beta$-subunits of all four AP complexes with taxon sampling expanded to represent eukaryotic diversity. We also undertook phylogeny of the large $\gamma$-subunit to confirm that the other subunits had indeed diverged into four distinct complexes before the LCEA.

A broadly sampled dataset of the large $\gamma$-subunit was analyzed, allowing us to identify and remove long-branch sequences (SI Fig. 11). In the resulting dataset, the four AP complexes each formed separate clades encompassing the diversity of eukaryotes sampled, with support of 1.00/100%/100%/100%, 1.00/100%/100%/100%, 1.00/100%/100%/100%, and 0.99/99%/99%/99% for $\gamma$, $\alpha$, $\delta$, and $\epsilon$ clades, respectively (SI Fig. 12; support values given in same order as in figures). This finding robustly confirms that the duplications giving rise to AP complexes preceded the LCEA. Similar results were recently shown for the medium AP subunit (24, 46).

Upon phylogenetic analysis to remove long-branch taxa (SI Fig. 13), analysis of the $\beta$-subunits (SI Fig. 14) showed strongly resolved support for a joint $\beta$1/2 clade (1.00/100%/100%/87%) and separate clades of $\beta$3 (1.00/100%/100%/100%) and $\beta$4 (1.00/
Lineages have a single paralogue; specifically the deuterostome divergence, because all other metazoa and sister Leishmania likely occurred. Even more clearly, in animals, the rise to all subunits of all AP complexes having occurred before the divergence of Monosiga brevicolis of the choanoflagellate Caenorhabditis elegans and Drosophila melanogaster. These data are most consistent with the duplications giving rise to all subunits of all AP complexes having occurred before the LCEA, with one exception. The β1- and β2-adaptin cluster by organismal lineage. It was found, but Oryza sativa and Oryza sativa RHA1 and its sister lineages have a single paralogue; specifically the β1/2-subunits of the choanoflagellate Monosiga brevicolis and the proteosome invertebrates Caenorhabditis elegans and Drosophila melanogaster.

These data are most consistent with the duplications giving rise to all subunits of all AP complexes having occurred before the LCEA, with one exception. The β-subunit of AP1 and AP2 duplicated independently in multiple eukaryotic lineages, after their divergence from the common ancestor (SI Fig. 8C). Thus, it seems that in the LCEA, AP1 and AP2 shared a common β-subunit despite being fully differentiated with respect to all of the remaining subunits in each complex. Many taxa do not, in fact, possess separate β1 and β2 homologues (47) (SI Table 1). Phylogenetic analysis of the β1/2 sequences alone did not resolve clades corresponding to AP1 and AP2, instead showing lineage-specific duplications (Fig. 4). Within flowering plants, a duplication specific for Arabidopsis thaliana was found, but Oryza sativa and Oryza sativa RHA1 each possess a single β1/2-subunit.

In the kinetoplastids, a duplication into AP1 and AP2 before the divergence of Trypanosoma and Leishmania likely occurred. Even more clearly, in animals, the β1 and β2 paralogues duplicated after the proterostome/deuterostome divergence, because all other metazoa and sister lineages have a single paralogue; specifically the β1/2-subunits of the choanoflagellate Monosiga brevicolis and the proteosome invertebrates Caenorhabditis elegans and Drosophila melanogaster.

Discussion

The ancestral eukaryote appears to have been a remarkably complex cell, with mitochondria, peroxisomes, introns, and nearly all eukaryotic hallmarks, including an endomembrane system (10–12, 14). In particular, all major vesicle coats and the organelles associated with those coats (10, 13, 32), as well as aspects of a sophisticated endocytic system, were already established before the LCEA (27). Our new analysis of three endocytic components, SynE, Rab 5, and β-adaptin, by contrast, reveals features that differentiated much later.

Instead of respective paralogues clustering according to intracellular localization, as observed for most paralogues functioning in membrane transport pathways originating before the LCEA, SynE, Rab 5, and β-adaptin cluster by organismal lineage. It is not possible to reconstruct the exact number of endocytic homologues present in the LCEA, due in part to the lack of resolution in the deepest nodes of the phylogenies in Figs. 1 and 3 and in part to the formal possibility that other endocytic genes may have been present in this ancestor that have been subsequently lost in multiple lineages. However, the phylogenetic pattern observed is most consistent with there being fewer paralogues of these endocytic components in the ancestral eukaryote than there are in many eukaryotes today.
Duplications giving rise to distinct AP1 and AP2 subunits had not undergone functional divergence, yielding increased specificity of function. Post-LCEA SynE and Rab 5 continued the trajectory of gene duplication and functional divergence, yielding increased specificity of function. The endocytic system of the LCEA consisted, most parsimoniously, of an ancestral SynE homologue and at least one endocytic Rab (Fig. 5). Although the order in which the Rab subfamilies evolved from a single prototypic Rab protein is unknown, it is clear that Rab 5 and Rab 7, which today demarcate the early endosomes and lysosomes/vacuoles (19), were already present and distinct from one another (37). Thus, the LCEA differentiated recycling and degradative endocytic functions (Fig. 5). However, because the LCEA may have possessed only a single SynE homologue, this implies that the Rab 5 and Rab 7 compartments/subdomains were serviced by an undifferentiated ancestral SynE protein. The duplications giving rise to the endosomal and lysosomal/vacuolar SynE paralogues therefore occurred independently, with convergence of function in at least vertebrates, streptophytes, and ascomycetes. The ability of the A. thaliana AtVAM3 and AtPEP12 to complement vam3 and pep12 yeast mutants, respectively, are examples of experimentally demonstrated functional convergence (51, 52). Thus, in this respect, early endosomes and lysosomes/vacuoles were likely less distinct than they appear in some modern systems. The duplications producing multiple Rab 5 paralogues also occurred independently in different lineages, implying continued evolution and specialization of the early endosomal system in many lineages subsequent to the LCEA (Fig. 5). That the β-adaptin duplications giving rise to distinct AP1 and AP2 subunits had not yet occurred in the LCEA suggests that the process of organellar differentiation begun by the other AP subunits was incomplete before the divergence of the major eukaryotic superfamilies (SI Fig. 8). Consistent with this idea of evolutionary plasticity in the eukaryotic endocytic system are recent studies of dynamin that suggest that its role in endocytosis may have evolved convergently in animals and ciliates (24). The loss of key endocytic components in eukaryotic lineages and the novel endocytic adaptors in opisthokonts, such as GGAs, and epsins also suggest evolutionary lability in the endocytic machinery (45).

Understanding the process by which eukaryotic organelles arose is a fundamental aim of evolutionary biology. If a cellular system evolving by paralogous organellar expansion (8, 28) was interrupted by a speciation event, a phylogenetic pattern could result whereby some of the organelle-specific machinery would resolve into clades by cellular function and some by lineage (SI Fig. 6). This pattern is precisely what is observed for the endocytic factors analyzed here. One important step to understanding the origins of eukaryotic cell organelles is to reconstruct the LCEA by comparative analysis. To achieve this, one must differentiate between truly ancestral eukaryotic characters and convergently evolved features restricted to well-studied model organisms. Here, we show that certain gene duplications during the evolution of the endocytic system, previously considered ancestral for all eukaryotes, actually occurred much later. It will be important to assess in more detail when the various pieces of the endocytic machinery evolved relative to the LCEA, emphasizing a need for experimental work in multiple and evolutionarily disparate organisms to determine which details of endocytosis are common to all eukaryotic cells.

**Methods**

**Sequence Sampling Strategy.** To ensure representation from the major eukaryotic superfamilies, to allow taxonomic comparability between analyses of the various proteins, to avoid issues of paralogy, and to maintain taxon cardinality amenable to rigorous computational analysis, an *a priori* sampling strategy was used. Amino acid sequences of Rab 5, SynE, and the large AP subunits were retrieved from representatives of the five eukaryotic supergroups with completed genome sequences. Homology was assessed by BLAST (53) and reciprocal BLAST, whereby candidates retrieved the initial query or named homologues thereof with at least five orders of magnitude better than the next named homologue of a different subfamily. Otherwise, the putative candidate was excluded from analyses to avoid paralogy issues. Sequences were retrieved from the National Center for Biotechnology Information (NCBI) (www.ncbi.nlm.nih.gov), the Joint Genome Institute (JGI) (www.jgi.doe.gov/), or from the URLs given in brackets. Accession numbers for proteins used for the analyses can be found in SI Table 2.

The taxa included in preliminary phylogenetic analyses were *A. thaliana* (NCBI), *C. elegans* (NCBI), *Chlamydomonas reinhardtii* (JGI), *Cryptococcus neoformans* (NCBI), *Cyndioschyzon merolae* (http://merolae.biols.u-tokyo.ac.jp), *Danio rerio* (NCBI), *Dictyostelium discoideum* (NCBI), *D. melanogaster* (NCBI), *Entamoeba histolytica* (NCBI), *Giardia intestinalis* (NCBI), *Homo sapiens* (NCBI), *Leishmania major* (www.genedb.org), *M. brevicolis* (JGI), *Naegleria gruberi* (JGI), *O. tauri* (JGI), *Trypanosoma brucei* (JGI), *Trypanosoma cruzi* (JGI), *Trypanosoma cruzi* (NCBI). To maintain comparability between analyses, *C. merolae*, *N. gruberi*, T. thermophila, and C. reinhardtii, although sampled, were excluded, because in each case at least two of the four possible genes examined could not be confidently identified by the above criteria. *Trichomonas vaginalis* was also excluded because of massive expansion of its endocytic machinery (54).

**Alignment and Phylogeny.** Amino acid sequences were aligned by using T.Coffee v.1.37 (55) and were manually adjusted. Only unambiguously aligned regions of homology were retained. Details of the datasets, including size and models of sequence evolution, are in SI Table 3. Alignments are available on request. For each protein, an initial phylogenetic analysis was performed to identify long-branch sequences, which were likely to contribute artifacts to subsequent analyses. These sequences were then removed, and a new mask was created, allowing for more alignment positions to be analyzed and for

---

**Fig. 5.** Evolution of endocytic organelles with rabs and syntaxins. The phylogenetic data here are most consistent with the endocytic system having evolved by the process of autogenous organelle evolution illustrated in SI Fig. 8 but being interrupted by the eukaryotic big bang. A single undifferentiated endocytic compartment would have initially been serviced by both undifferentiated endocytic rabs and syntaxins. The rabs duplicated and diverged before the eukaryotic big bang, however, the endocytic syntaxins did not. Post-LCEA SynE and Rab 5 continued the trajectory of gene duplication and functional divergence, yielding increased specificity of function.
more-intensive phylogenetic analysis involving Bayesian, two methods of protein maximum-likelihood (ML), and ML-corrected distance (MLD) methodologies. All datasets contained representatives from the five eukaryotic supergroups with sequenced genomes. Further subalignments were created and analyzed to test specific clade robustness (SI Fig. 10) and to test the resolution of character functionally homologues into either organelle- or lineage-specific groups (Fig. 2).

The model of sequence evolution for each dataset was determined by using ProtTest version 1.3 (56) and incorporated corrections for rate variation and invariable sites where relevant. Trees were built by using Fitch from the PHYLIP v.3.6a3 package (60) based on distance matrices calculated by Tree-Puzzle (61) and phylodeplicate datasets. Apart from the datasets for SI Figs. 10 and 11, which were generated by using Fitch from the PHYLIP v.3.6a3 package (60) based on distance matrices calculated by Tree-Puzzle (61) and phylodeplicate datasets. All datasets contained representatives from the five eukaryotic supergroups with sequenced genomes. Further subalignments were created and analyzed to test specific clade robustness (SI Fig. 10) and to test the resolution of character functionally homologues into either organelle- or lineage-specific groups (Fig. 2).

The model of sequence evolution for each dataset was determined by using ProtTest version 1.3 (56) and incorporated corrections for rate variation and invariable sites where relevant. Trees were built by using Fitch from the PHYLIP v.3.6a3 package (60) based on distance matrices calculated by Tree-Puzzle (61) and phylodeplicate datasets. Apart from the datasets for SI Figs. 10 and 11, which were generated by using Fitch from the PHYLIP v.3.6a3 package (60) based on distance matrices calculated by Tree-Puzzle (61) and phylodeplicate datasets. All datasets contained representatives from the five eukaryotic supergroups with sequenced genomes. Further subalignments were created and analyzed to test specific clade robustness (SI Fig. 10) and to test the resolution of character functionally homologues into either organelle- or lineage-specific groups (Fig. 2).

The model of sequence evolution for each dataset was determined by using ProtTest version 1.3 (56) and incorporated corrections for rate variation and invariable sites where relevant. Trees were built by using Fitch from the PHYLIP v.3.6a3 package (60) based on distance matrices calculated by Tree-Puzzle (61) and phylodeplicate datasets. Apart from the datasets for SI Figs. 10 and 11, which were generated by using Fitch from the PHYLIP v.3.6a3 package (60) based on distance matrices calculated by Tree-Puzzle (61) and phylodeplicate datasets. All datasets contained representatives from the five eukaryotic supergroups with sequenced genomes. Further subalignments were created and analyzed to test specific clade robustness (SI Fig. 10) and to test the resolution of character functionally homologues into either organelle- or lineage-specific groups (Fig. 2).

The model of sequence evolution for each dataset was determined by using ProtTest version 1.3 (56) and incorporated corrections for rate variation and invariable sites where relevant. Trees were built by using Fitch from the PHYLIP v.3.6a3 package (60) based on distance matrices calculated by Tree-Puzzle (61) and phylodeplicate datasets. Apart from the datasets for SI Figs. 10 and 11, which were generated by using Fitch from the PHYLIP v.3.6a3 package (60) based on distance matrices calculated by Tree-Puzzle (61) and phylodeplicate datasets. All datasets contained representatives from the five eukaryotic supergroups with sequenced genomes. Further subalignments were created and analyzed to test specific clade robustness (SI Fig. 10) and to test the resolution of character functionally homologues into either organelle- or lineage-specific groups (Fig. 2).

The model of sequence evolution for each dataset was determined by using ProtTest version 1.3 (56) and incorporated corrections for rate variation and invariable sites where relevant. Trees were built by using Fitch from the PHYLIP v.3.6a3 package (60) based on distance matrices calculated by Tree-Puzzle (61) and phylodeplicate datasets. Apart from the datasets for SI Figs. 10 and 11, which were generated by using Fitch from the PHYLIP v.3.6a3 package (60) based on distance matrices calculated by Tree-Puzzle (61) and phylodeplicate datasets. All datasets contained representatives from the five eukaryotic supergroups with sequenced genomes. Further subalignments were created and analyzed to test specific clade robustness (SI Fig. 10) and to test the resolution of character functionally homologues into either organelle- or lineage-specific groups (Fig. 2).