

Figure S1

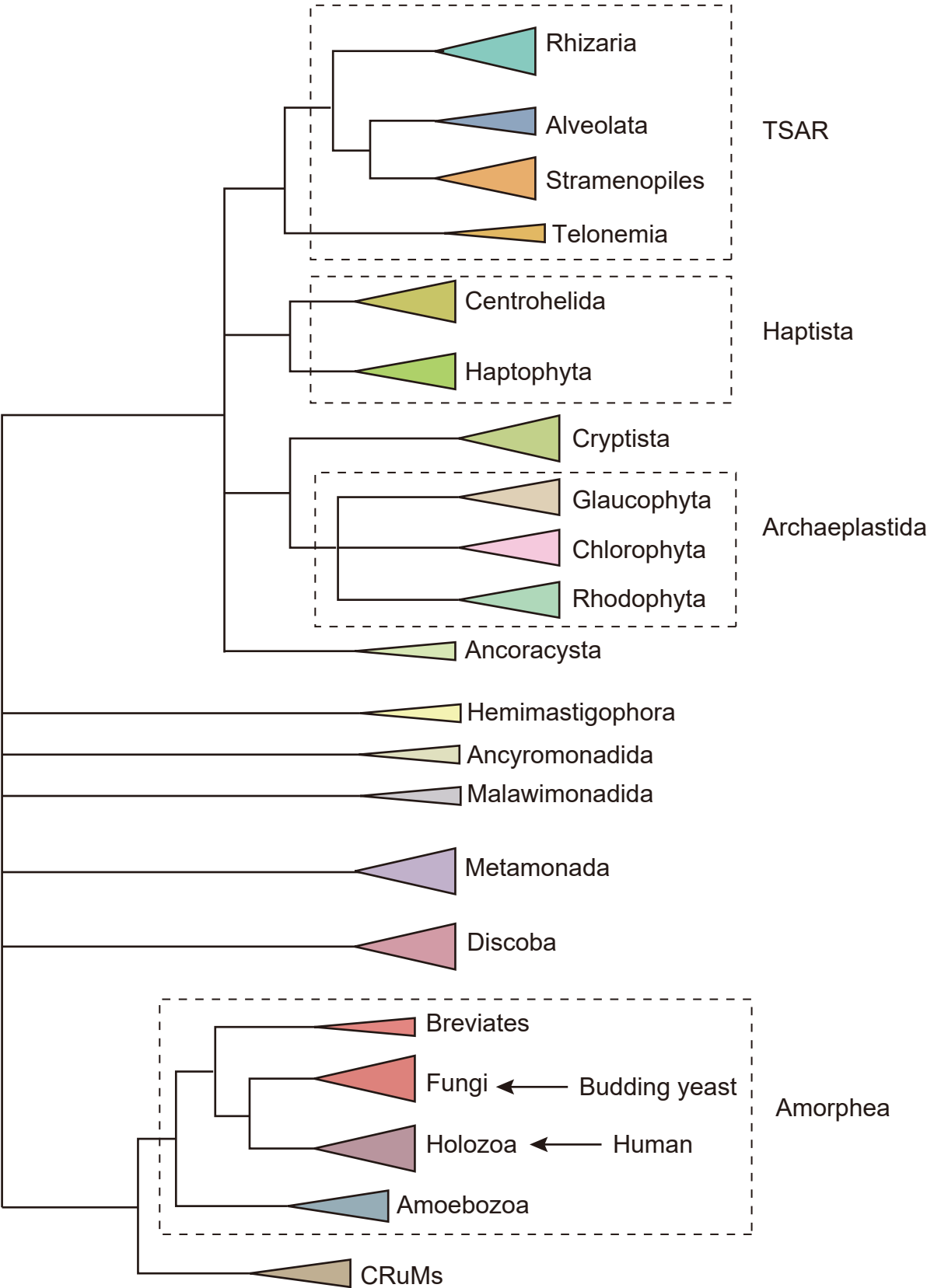


Figure S1. The phylogenetic tree of the 21 clades containing the 94 eukaryotic species included in this study. A list of species names by the clades can be found in Table S1. TSAR, Telonemia, Stramenopiles, Alveolata and Rhizaria; CRuMs, collodictyonids, rigifilids, and Mantamonas.

Figure S2

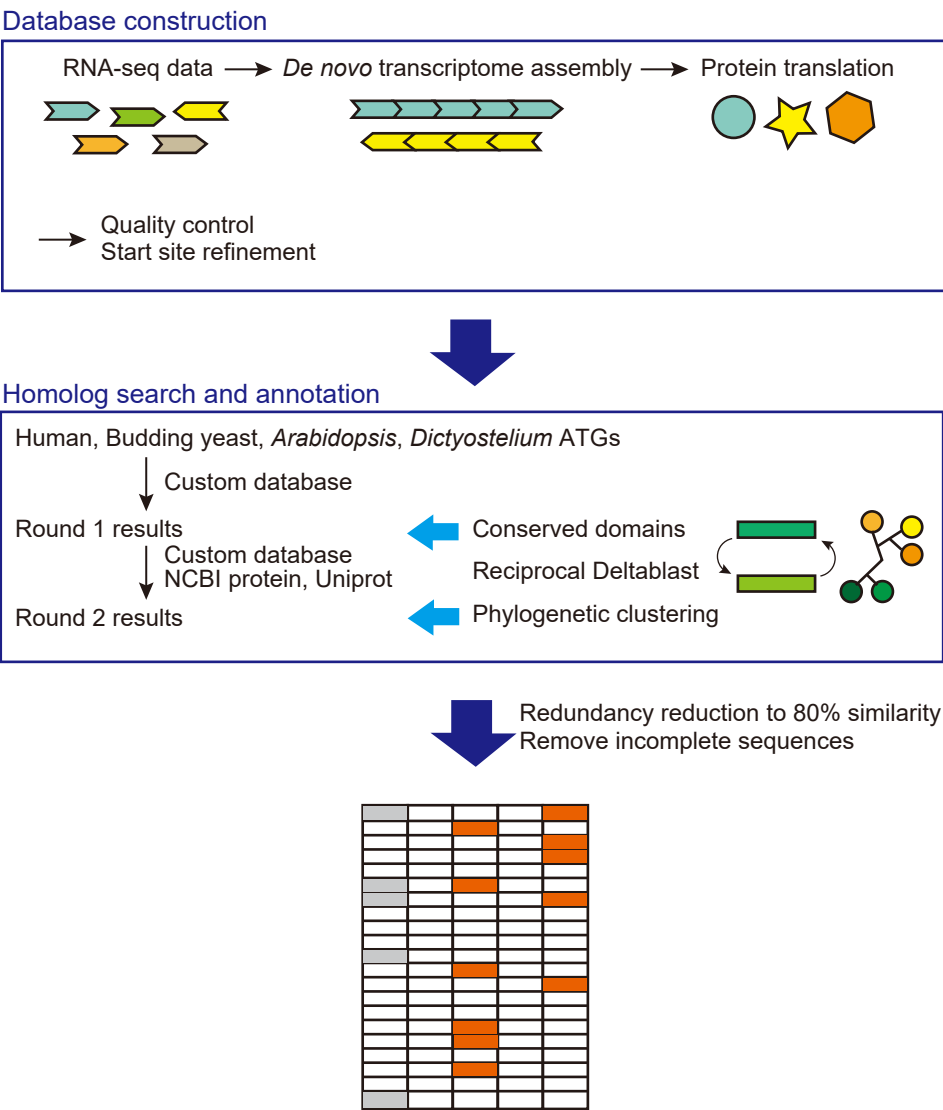


Figure S2. The workflow of our study. There are two main steps: first, a custom transcriptome database was constructed via de novo transcriptome assembly from RNA-sequencing data. Construction of the custom database was necessary as we included many species without high-quality reference genomes or gene annotations in order to cover the diversity within eukaryotes. Second, a pipeline was developed to systematically search for and annotate components of the ATG conjugation systems. Because some assembled transcriptomes have low BUSCO scores, we supplemented the custom database with public data from NCBI protein and Uniprot if available. Because the ATG sequences may have diverged during evolution, using the same query sequences for all 94 species may not be sufficient. Instead, we carried out two rounds of analysis, where the output of the first round, which used query sequences from four model organisms, were used as the new queries for the second round. Additional rounds were carried out for ATG16. In each round, a combination of conserved domain search and reciprocal DELTA-BLAST were used to annotate the sequences, and phylogenetic clustering to resolve the remaining ambiguity.

Figure S3

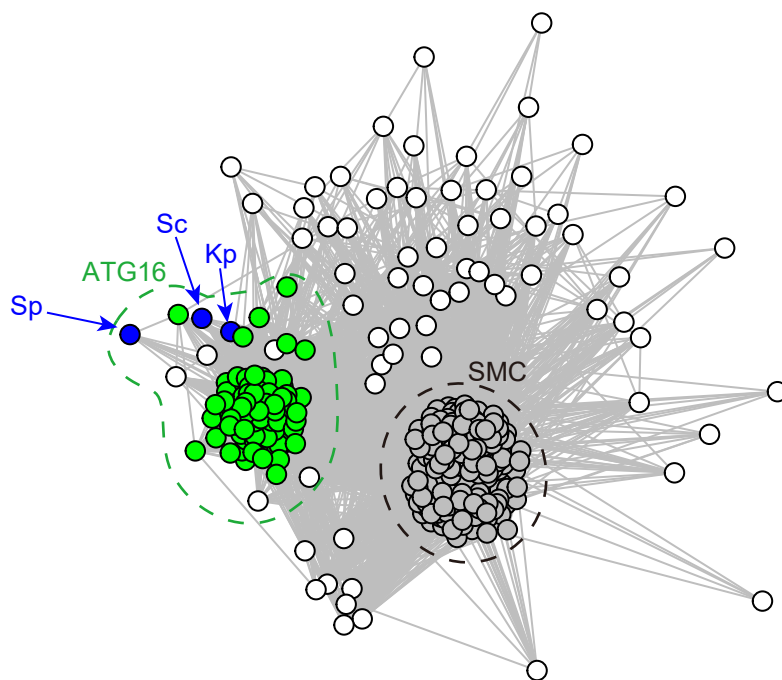


Figure S3. Graph Splitting is able to separate ATG16 homologs from non-homologs. Homologs are represented by green circles surrounded by the green dashed line, while non-homologs are represented by white and gray circles. The non-homologous sequences included in this plot are either sequences with some level of evidence of being a homolog (mostly from conserved domain search, which in the case of ATG16 does not distinguish true ATG16 from other coiled-coil-containing proteins very well in our experience) (white circles), or SMC proteins (gray circles) included as reference because they are a major family of coiled-coil-containing proteins. Five sequences were removed (white circles within the region surrounded by the green dashed line) because more complete sequence exists for the same species. ATG16 homologs without the WD40 domain, e.g. those of *Saccharomyces*, *Komagataella* and *Schizosaccharomyces*, are in the peripheral of the ATG16 cluster. SMC, structural maintenance of chromosomes. Sc, *Saccharomyces cerevisiae*; Kp, *Komagataella pastoris*; Sp, *Schizosaccharomyces pombe*.

Figure S4

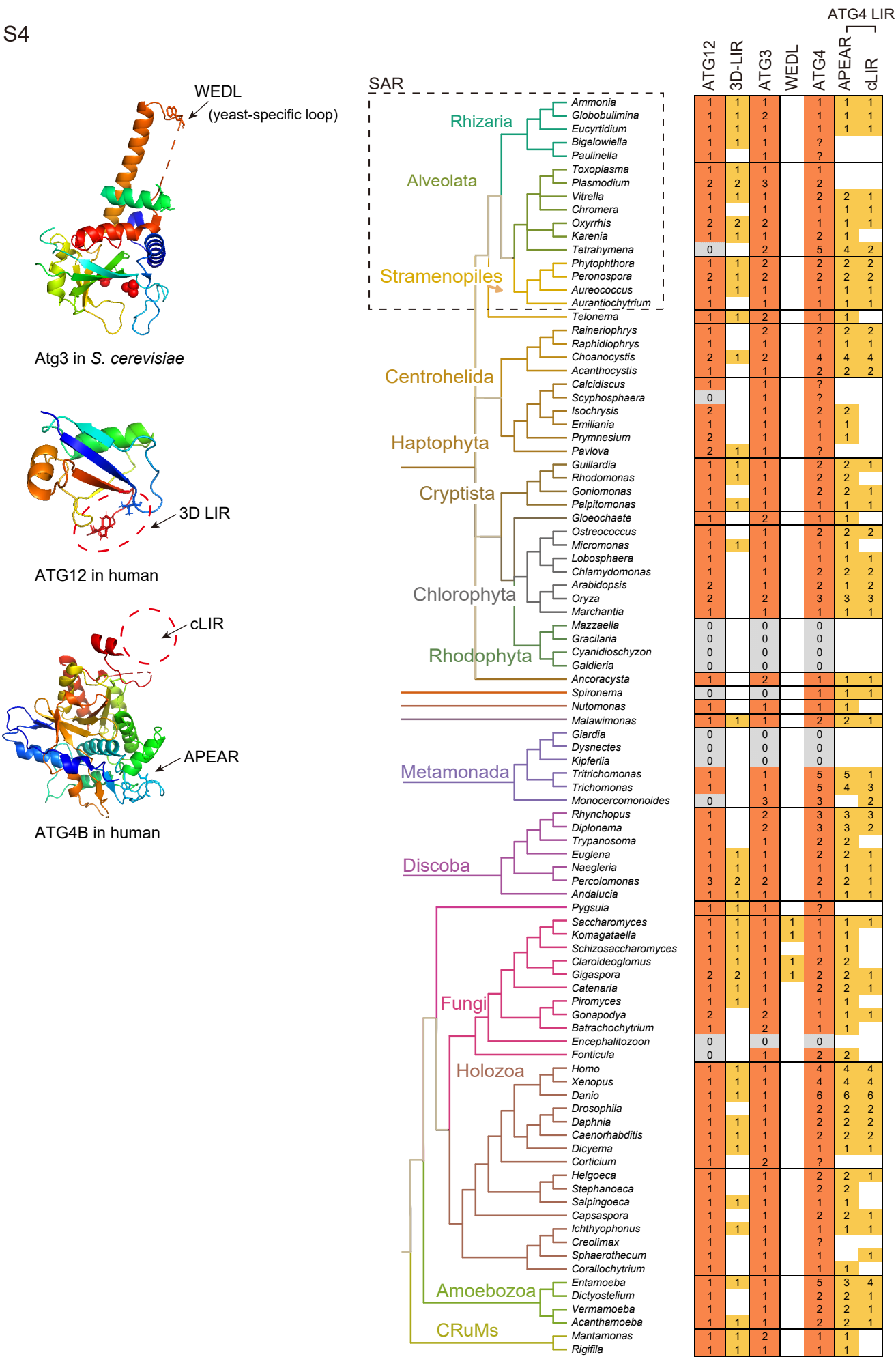
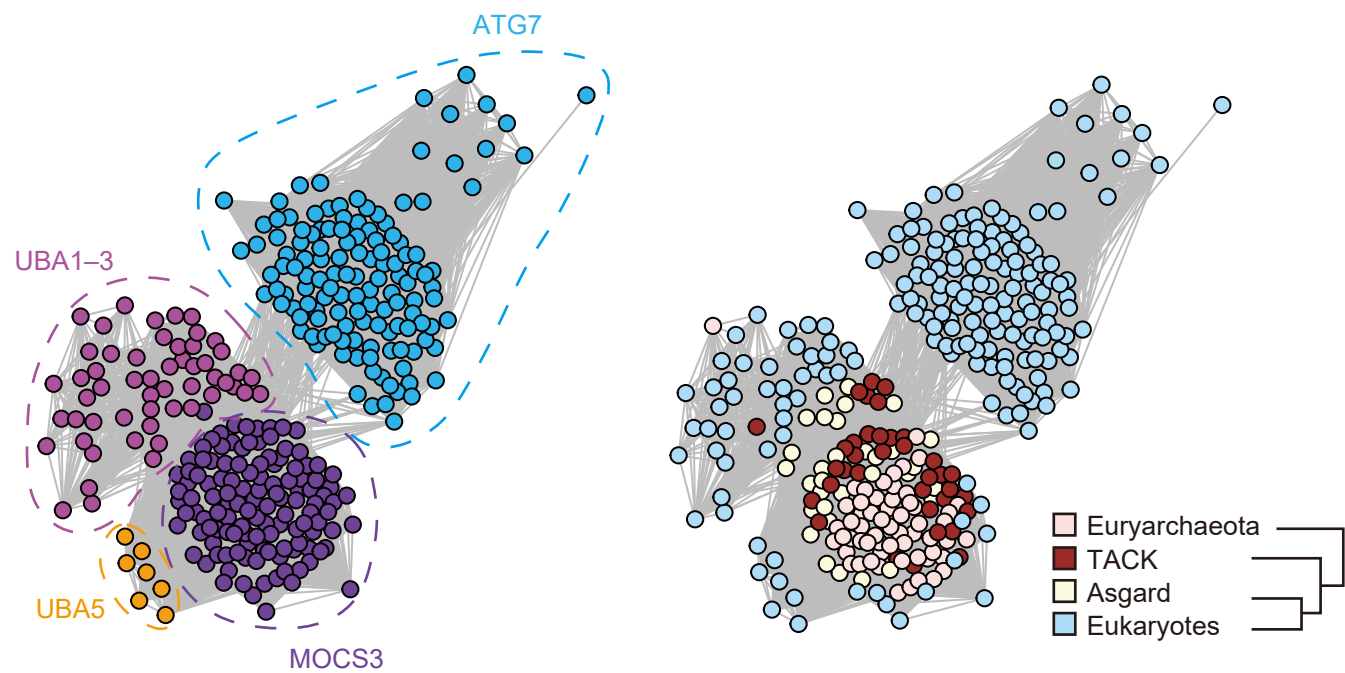


Figure S4. The conservation of LIR motifs in ATG12, ATG3 and ATG4. Both the 3D-LIR in ATG12 and the LIR motifs (cLIR and APEAR) in ATG4 are at least partially conserved in a wide range of species. The LIR motif in Atg3, on the other hand, exists on a mostly fungi-specific loop. The positions of the LIR motifs on the respective proteins are shown on the left (PDB: 2dyt, 4naw, 2cy7). LIR, LC3-interacting region; APEAR, ATG8-PE association region.

Figure S5

A



B

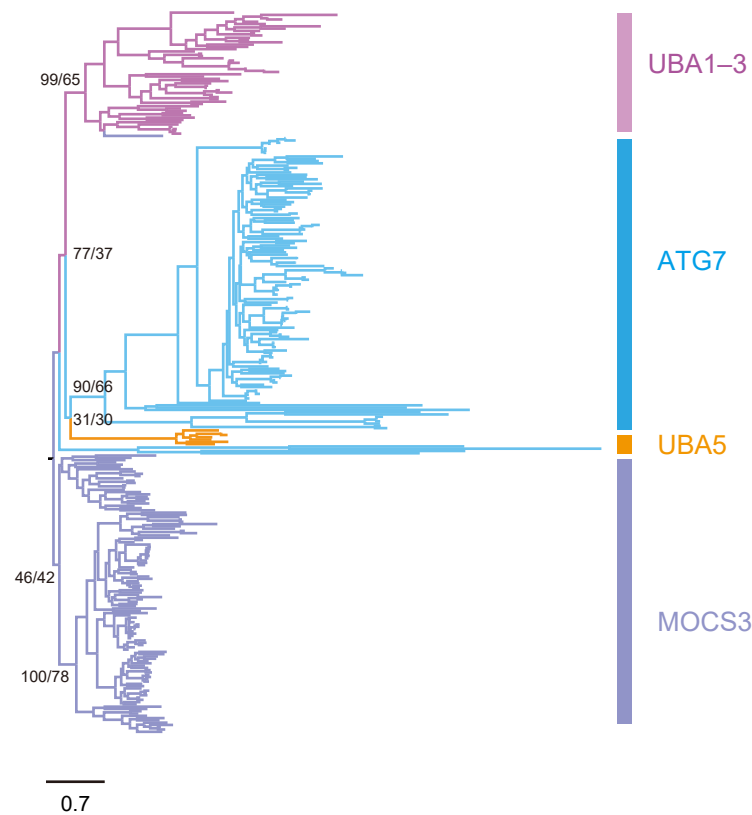


Figure S5. Phylogenetic analysis of the E1-like enzymes. UBA1, UBA2, UBA3, MOCS3, UBA5 and ATG7, in eukaryotes and three archaeal clades were analyzed by Graph Splitting (**A**) and IQ-TREE (**B**). The E1-like enzymes are divided into four groups, the MOCS3 group, the UBA1-3 group, the UBA5 group, and the ATG7 group. The (unrooted) maximum likelihood phylogenetic tree (**B**) is manually rooted by the MOCS3 sequences, and statistical support (SH-aLRT/ultrafast bootstrap) is shown for the major branches.