Current Biology, Volume *32*

Supplemental Information

Genomic analysis reveals cryptic diversity in aphelids and sheds light on the emergence of Fungi

Kirill V. Mikhailov, Sergey A. Karpov, Peter M. Letcher, Philip A. Lee, Maria D. Logacheva, Aleksey A. Penin, Maksim A. Nesterenko, Igor R. Pozdnyakov, Evgenii V. Potapenko, Dmitry Y. Sherbakov, Yuri V. Panchin, and Vladimir V. Aleoshin

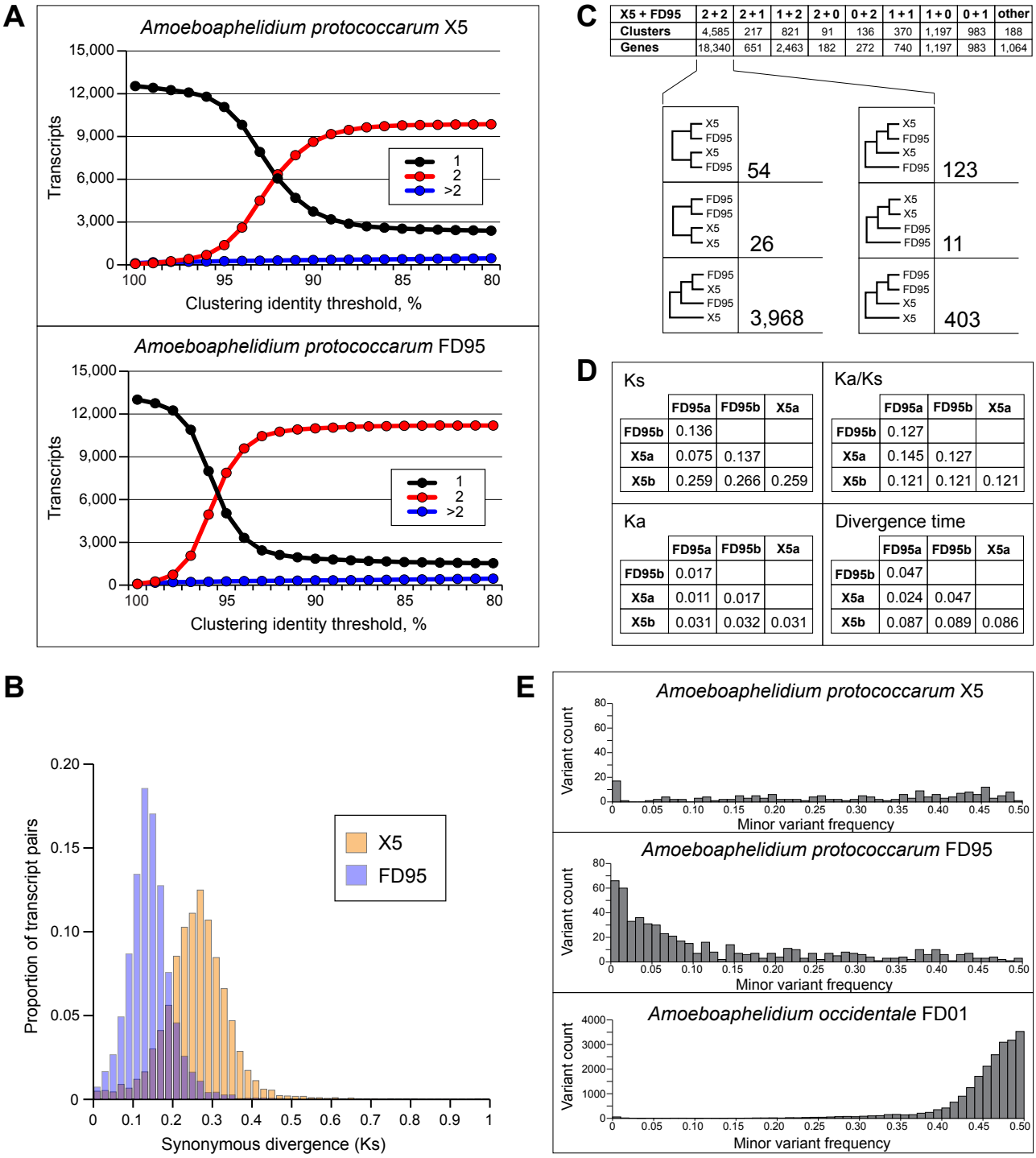

Figure S1. Characteristics of hybrid genomes of *A. protococcarum*. Related to Figure 1. (**A**) Clustering of transcripts in the genomes of *A. protococcarum* strains X5 and FD95. The cd-hit clustering was performed with a varying identity threshold (from 100% to 80% with a step of 1%), for each value of the threshold the graphs show the number of genes that fall in clusters of size 2 (red), clusters with over 2 members (blue) or remain singular (black). (**B**) Distributions of per-site synonymous divergence (Ks) values between the gene pairs in the genomes of *A. protococcarum* strains X5 and FD95. (**C**) Similarity clustering of pooled transcripts from the genomes of *A. protococcarum* strains X5 and FD95, and UPGMA tree inference for clusters containing a gene pair from each of the strains. The cd-hit clusters of all predicted transcripts in the genomes of X5 and FD95 were classified into categories according to the number genes from each strain that formed the cluster; sequences in clusters with a pair of genes from each strain ("2+2" clusters) were aligned and their phylogenetic relationship was inferred using the UPGMA method – the resulting trees were classified into the six topologies depicted in the diagram. (**D**) Estimates of synonymous divergence (Ks), nonsynonymous divergence (Ka), and divergence time in substitutions per site in the concatenated alignments of "2+2" cluster sequences following subgenome assignments of homoeologous genes. (**E**) Histograms of variant frequencies (minor variant to read depth ratios) in the mappings of paired-end reads to genome assemblies of *A. protococcarum* strains X5 and FD95 and *A. occidentale* strain FD01.

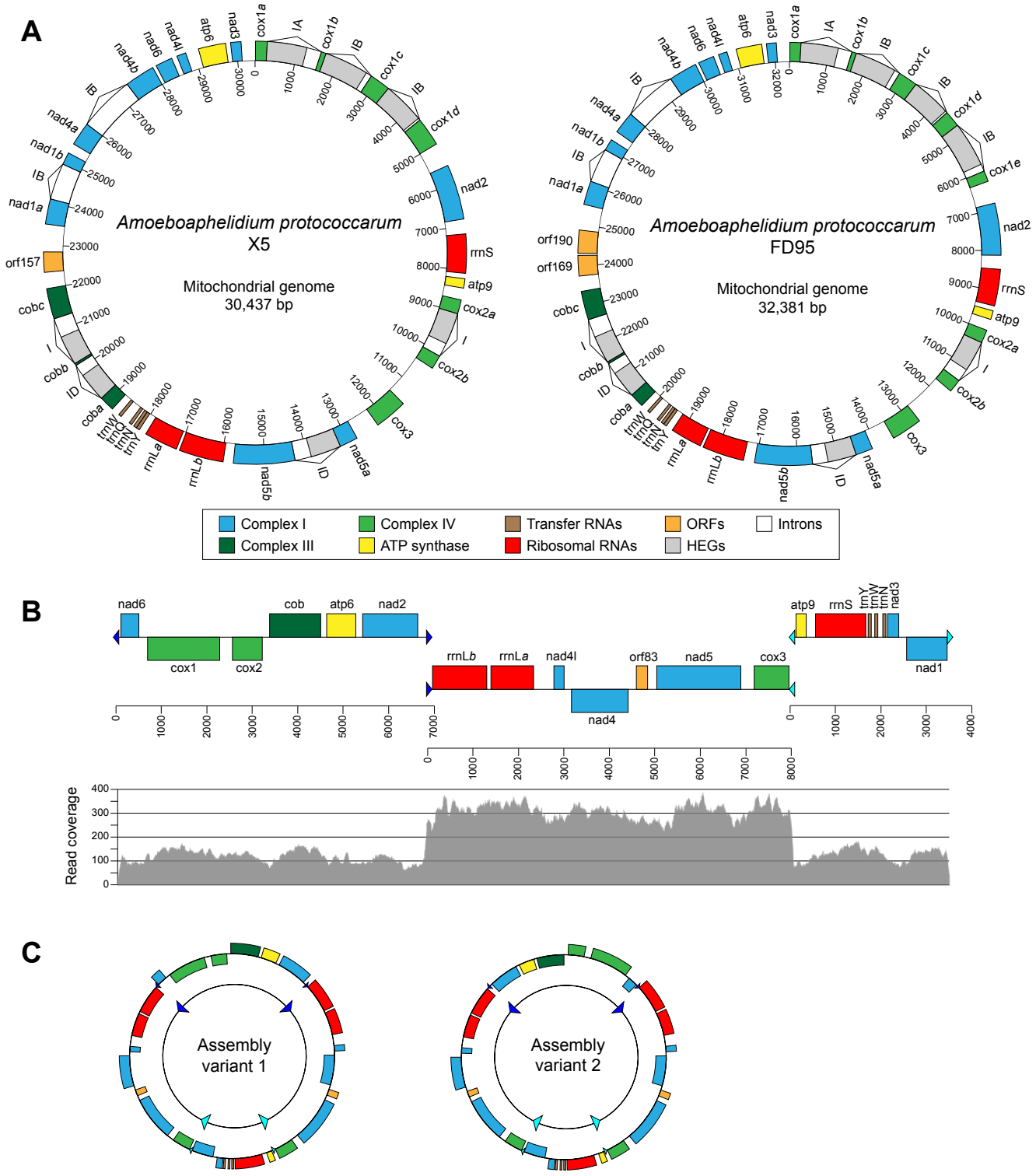

Figure S2. Mitochondrial genome assemblies of *A. protococcarum* and *A. occidentale*. Related to Figure 1. (**A**) Mitochondrial genome maps of *A. protococcarum* strains X5 and FD95. (**B**) Mitochondrial contigs of *A. occidentale* strain FD01. All genes in the mitochondrial genomes of *A. protococcarum* are encoded on a single strand; they include 13 respiratory complex components, small and large subunits of ribosomal RNA, with the latter one split into 2 parts (*rrnLa* and *rrnLb*), 4 transfer RNAs, and homing endonuclease genes (HEGs) – 7 in X5 and 8 in FD95. The mitochondrial genome in *A. occidentale* is found in 3 contigs and contains no introns or HEGs; the average read coverage of the 8 Kb contig is approximately 2.5 times higher than the coverage of the shorter contigs. The overlaps between contigs of *A. occidentale* are represented with light blue and dark blue arrowheads. (**C**) Two possible models for the assembly of a 26,157 bp circular mitochondrial genome for *A. occidentale* with the 8 Kb contig incorporated as an inverted repeat.

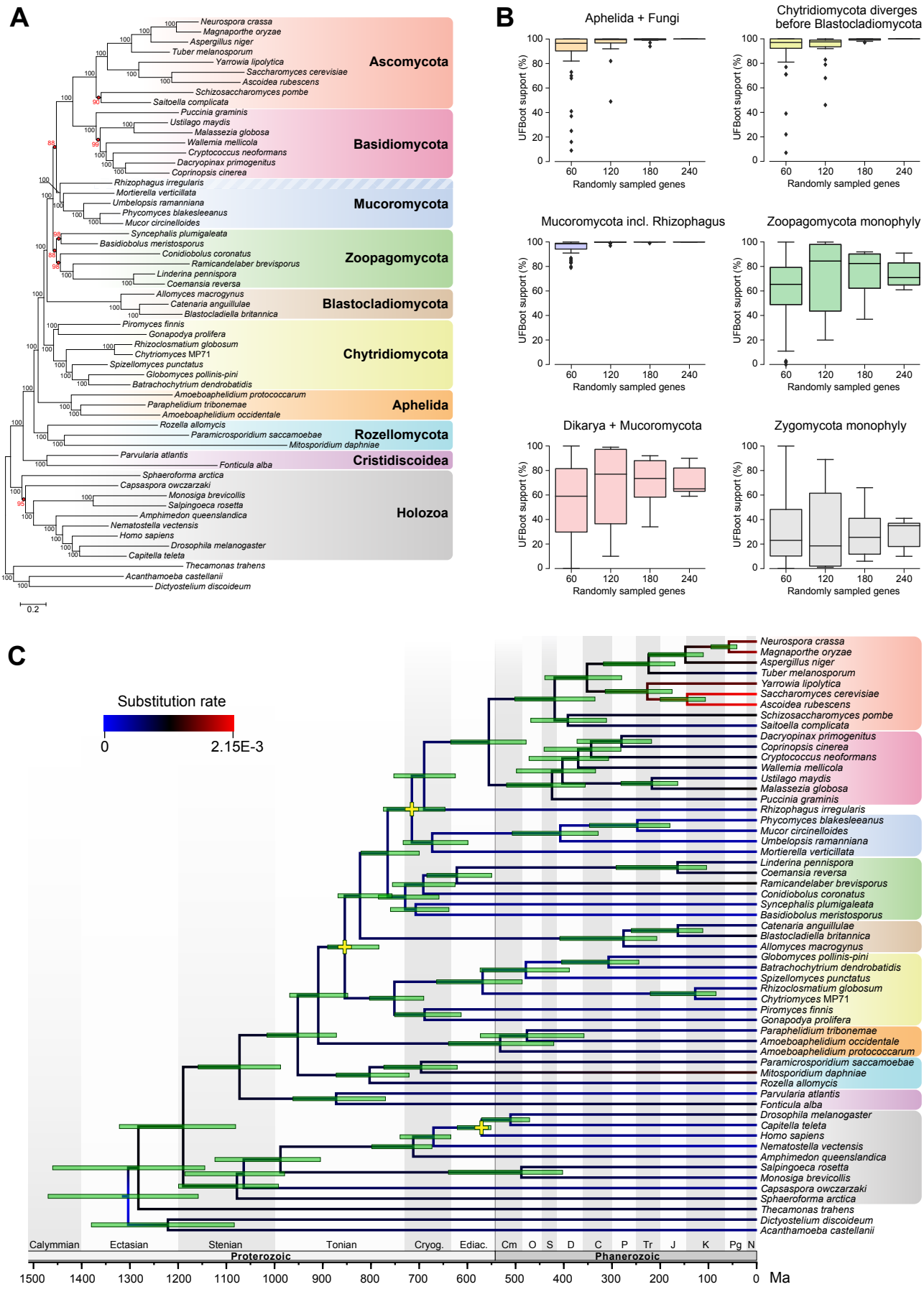

Figure S3. Assessment of the influence of gene subsampling on the stability of phylogeny and divergence date estimates using time-calibrated analysis. Related to Figure 2. (**A**) IQ-TREE maximum likelihood reconstruction with the 300-gene alignment using the LG+C60+F+G4 evolutionary model; node support was calculated using the ultrafast bootstrap approximation with 1000 replicates; nodes with support values below 100% are marked in red. (**B**) Impact of gene subsampling on the bipartitions of interest; the 300-gene dataset was used to randomly sample 20%, 40%, 60%, and 80% of

genes, which were concatenated and analyzed with IQ-TREE; 40 replicates were generated for the 60-gene dataset, 20 replicates for the 120-gene dataset, 10 replicates for the 180-gene dataset, and 5 replicates for the 240-gene dataset; ultrafast bootstrap support values for bipartitions across replicates are presented using box plots with a 1.5 interquartile range threshold to specify outliers. (**C**) Time-calibrated phylogeny inferred by PhyloBayes under the CAT-GTR model using a 30-gene dataset – a subset of the 300-gene dataset, selected for most clock-like behavior; green bars at the tree nodes represent 95% confidence intervals for posterior probability estimates of divergence times; the analysis was performed under a lognormal autocorrelated relaxed clock model with three calibration points (nodes marked with yellow crosses): setting the maximal age of true fungi at 890 Ma, the minimal age for the divergence of mycorrhizal fungi at 470 Ma, and confining the divergence of the bilaterian lineage within the 550-636 Ma time interval; the maximal and minimal ages for fungal divergences are motivated by the proposed links between the evolution of streptophytes and fungi[S1], and the previously estimated bounds on the emergence of streptophytes and land plants[S2]; the estimated evolutionary rates of branches (in substitutions per site per million years) are presented using a color gradient.

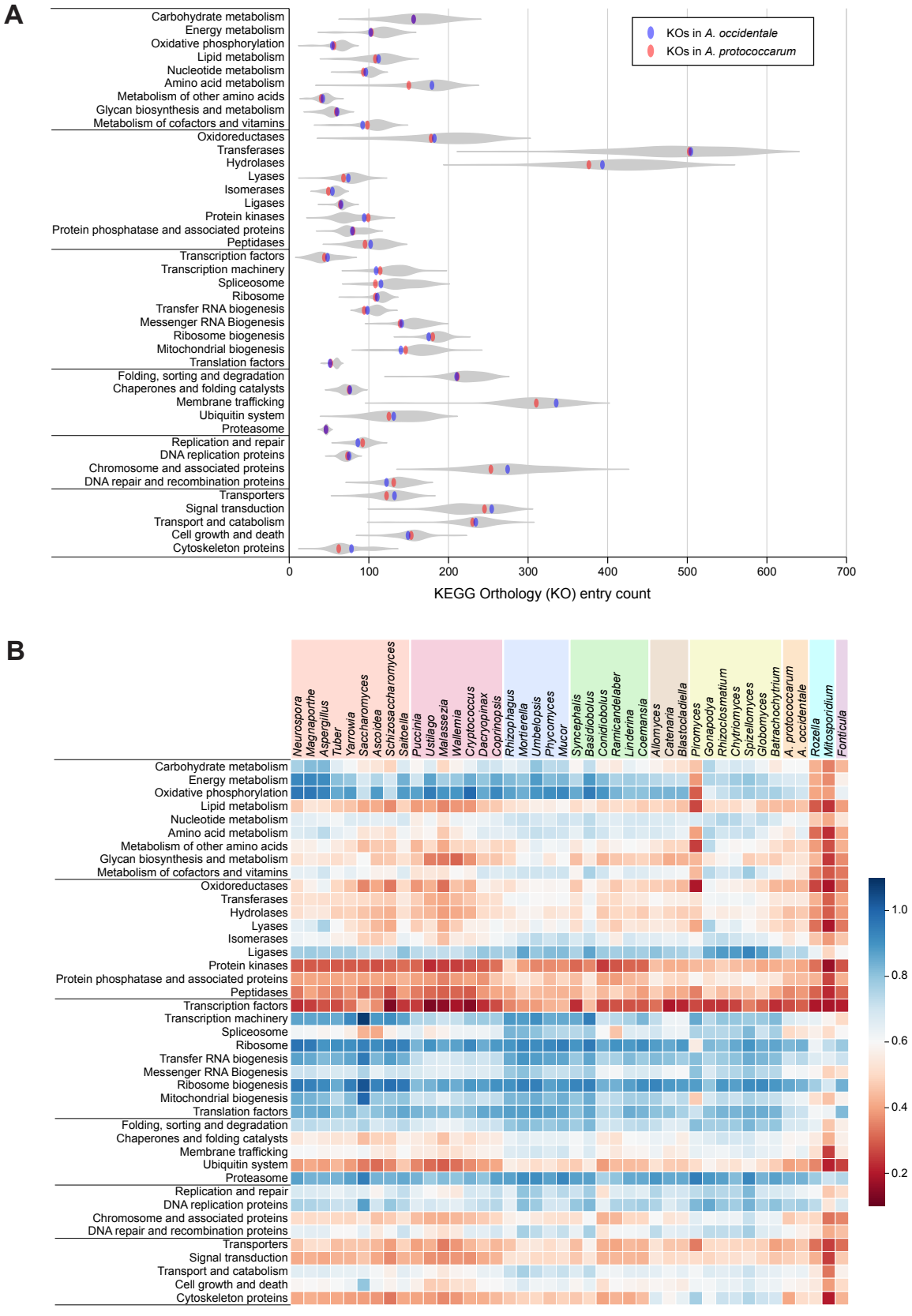

Figure S4. Comparative analysis of functional annotations for the genomes in Holomycota using KEGG. Related to Figure 3. (**A**) Distributions of annotated KEGG orthology (KO) counts in holomycotan genomes by functional categories, according to the KEGG BRITE classification; the distributions, shown as grey violin plots, were constructed from the genomic data of 40 holomycotan species, annotated using KAAS; KO annotations for each genome were reduced to KO presence/absence data; KO counts for the genomes of *A. occidentale* and *A. protococcarum* are shown as blue and red data points, respectively. (**B**) Heatmap of KO entries (presence/absence data) for BRITE functional categories in the genomes of holomycotan species, normalized to the inferred counts of unique KOs in the last common ancestor of Holomycota.

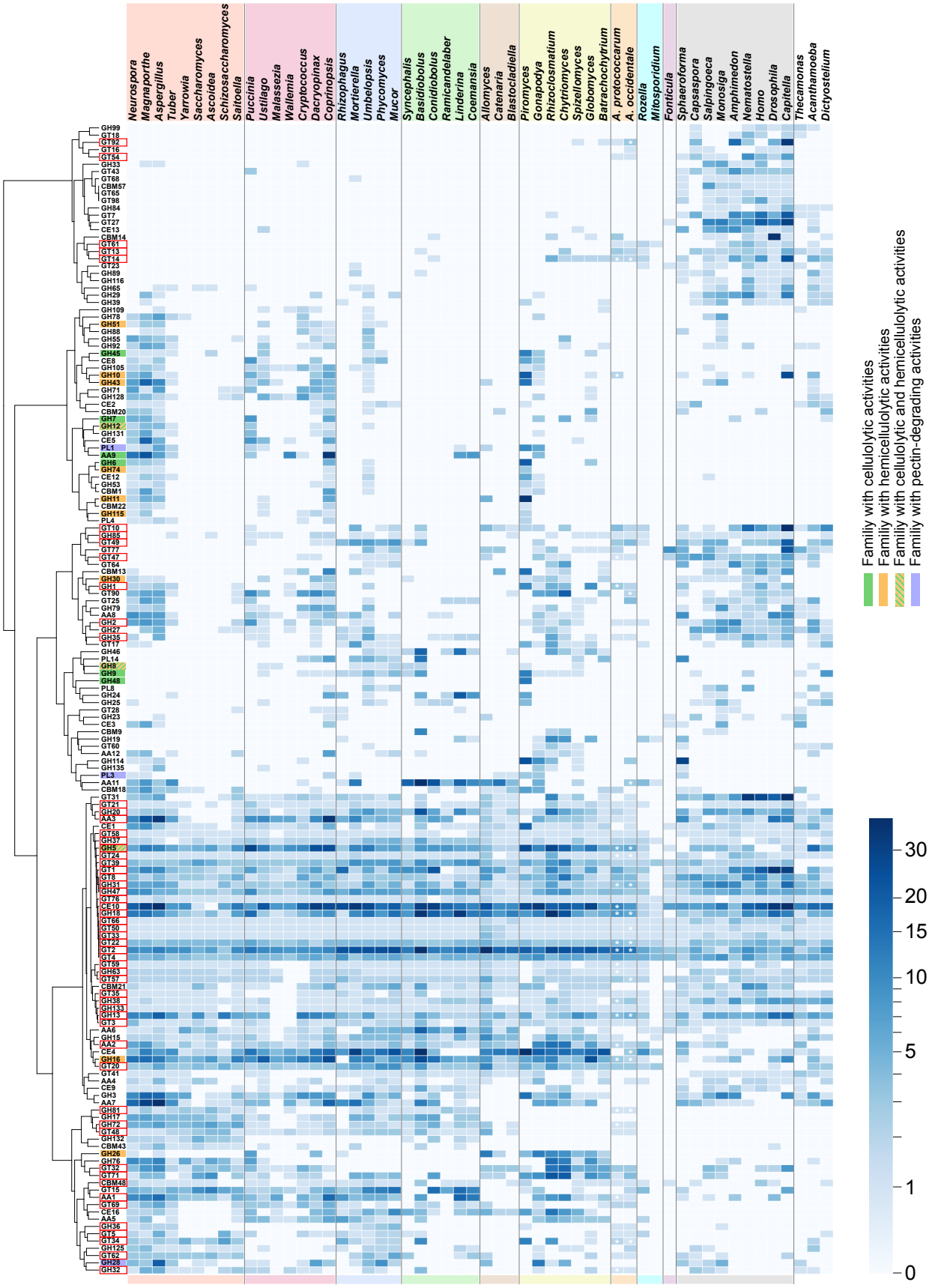

Figure S5. Heatmap of CAZy family member counts in the genomes of 54 eukaryotic species. Related to Figure 4. The coloring intensity is scaled logarithmically (scale bar to the right); CAZy families are clustered on the basis of binary presence/absence profiles in the inspected genomes using the Ward's method; families containing characterized enzymes with plant cell wall degrading activities are marked with colors: cellulolytic – green, hemicellulolytic – orange, pectin-degrading – blue; families were the two *Amoeboaphelidium* species have shared orthologous groups are marked with a red frame; white asterisks for aphelid families denote presence of sequences with a predicted secretory signal peptide.

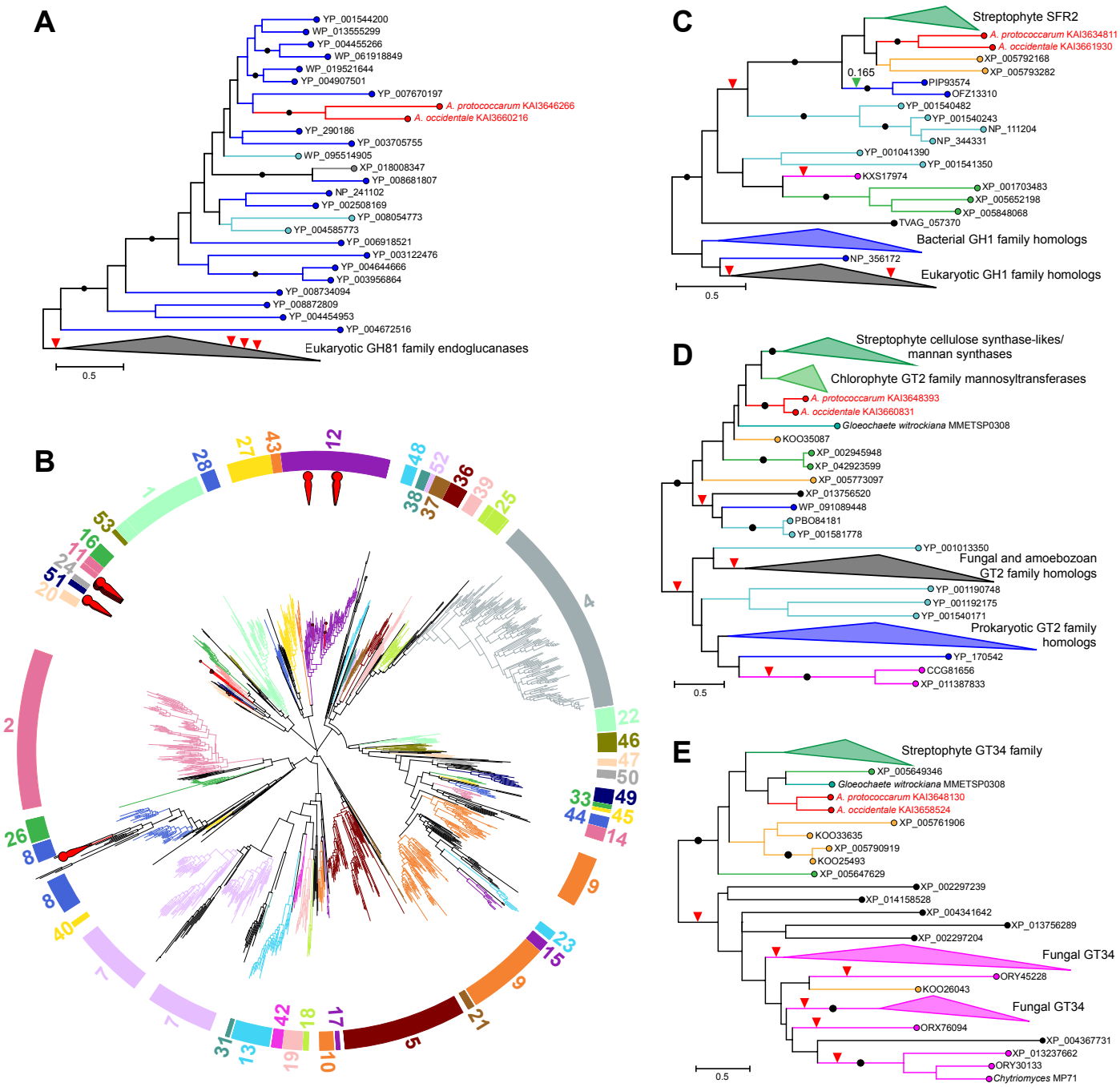

Figure S6. Maximum likelihood phylogenetic trees for GH5 family sequences and horizontal gene transfer candidates among the aphelid CAZymes. Related to Figure 4. (**A**) Phylogenetic tree with GH81 family endoglucanases; aphelid sequences are marked in red, bacterial sequences are in dark blue, and archaeal in light blue; eukaryotic cluster includes fungal, streptophyte and algal sequences; black circles on tree branches correspond to 100% UFboot support in the analysis; alternative placements for aphelid sequences, which were examined using the AU test, are labeled with triangles on tree branches: the placements rejected by the AU test at the 5% level are labeled with red triangles; for GH81 family we tested three alternative positions for aphelids within the cluster of eukaryotic sequences, by placing them at the bases of three fungal clusters – all alternatives were rejected by the test. (**B**) Phylogenetic tree with GH5 family sequences; the tree was reconstructed using the LG+C20+F+G4 model with an alignment of 1,431 Cellulase GH5 domain (PF00150) sequences; GH5 subfamilies and the corresponding subtrees are indicated in color and labeled on the outer rim of the diagram; subfamily classification and the initial sequence set are based on Aspeborg *et al*., 2012[S3] – the sequence set from the 2012 study was expanded with PF00150 domain hits from aphelid, rozellid, and fungal GH5 family enzymes; aphelid sequences in the tree are shown in red and highlighted on the rim of the diagram with red pins. (**C**) Phylogenetic tree with GH1 family SFR2 homologs; the tree colors and labels are as in (A),

additionally, green marks chlorophyte and streptophyte sequences/clusters, hapthophytes are marked in orange, and fungal sequences with magenta; a green triangle on the bacterial branch marks an alternative position for aphelid sequences that was not rejected by the AU test (p-value 0.165). (**D**) Phylogenetic tree with GT2 family putative mannosyltransferases; using the same color scheme and labels as in (A) and (C). (**E**) Phylogenetic tree with GT34 family putative xylosyltransferases; using the same color scheme and labels as in (A) and (C). The putative enzyme activities for (C-E) are based on the characterized enzymes of *Arabidopsis thaliana* most closely related to the aphelid sequences.

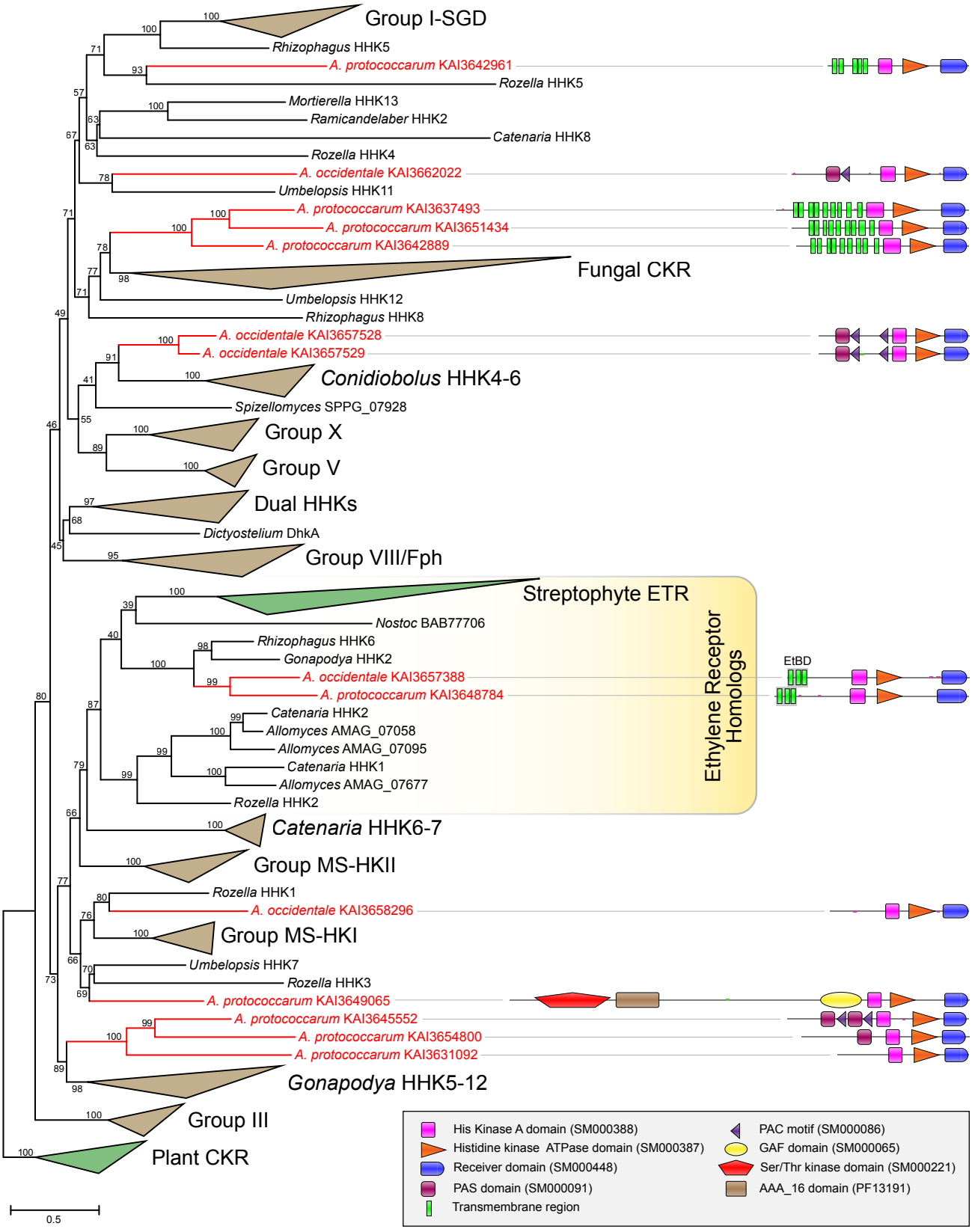

Figure S7. Phylogeny of hybrid histidine kinases with aphelid sequences. Related to Figure 5. Maximum likelihood phylogenetic tree was reconstructed by IQ-TREE with the LG+R6 model using an alignment of hybrid histidine kinase regions spanning the conserved Histidine Kinase A, ATPase, and the Receiver domains; the dataset of hybrid histidine kinases along with the sequence names and group designations are based on the Herivaux *et al*., 2017[S4]; branch support was evaluated using UF bootstrap with 1,000 replicates; highly-supported groups are collapsed in the tree; aphelid sequences are marked with red color and the corresponding protein domain architectures are depicted on the right; EtBD – ethylene binding domain.

Supplemental References

S1. Berbee, M.L., Strullu-Derrien, C., Delaux, P.M., Strother, P.K., Kenrick, P., Selosse, M.A., and Taylor, J.W. (2020). Genomic and fossil windows into the secret lives of the most ancient fungi. Nature reviews. Microbiology *18*, 717-730.

S2. Morris, J.L., Puttick, M.N., Clark, J.W., Edwards, D., Kenrick, P., Pressel, S., Wellman, C.H., Yang, Z., Schneider, H., and Donoghue, P.C.J. (2018). The timescale of early land plant evolution. Proc Natl Acad Sci U S A *115*, E2274-E2283.

S3. Aspeborg, H., Coutinho, P.M., Wang, Y., Brumer, H., 3rd, and Henrissat, B. (2012). Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5 (GH5). BMC Evol Biol *12*, 186.

S4. Herivaux, A., Duge de Bernonville, T., Roux, C., Clastre, M., Courdavault, V., Gastebois, A., Bouchara, J.P., James, T.Y., Latge, J.P., Martin, F., and Papon, N. (2017). The Identification of Phytohormone Receptor Homologs in Early Diverging Fungi Suggests a Role for Plant Sensing in Land Colonization by Fungi. mBio *8*.