An ancestral signalling pathway is conserved in plant lineages forming intracellular symbioses

Supplementary methods

OneKP transcriptome annotation

A total of 264 transcriptomes from non-angiosperm species were collected from the 1KP project (https://db.cngb.org/onekp/). For each transcriptome, ORFs were predicted using TransDecoder v5.5.0 (http://transdecoder.github.io) with default parameters. Then, predicted ORFs were subjected to BLASTp v2.7.1+[66] and PFAM searches against the Uniprot (accessed 11/2018) and PFAM v32[67] database. Results of BLAST and PFAM searches were used to improve final annotation and coding-sequence prediction.

Phylogeny and sequence analysis

Phylogenetic analysis of candidate genes. Homologs of each candidate genes (Supplementary Table 2) were retrieved from the SymDB database with the tBLASTn algorithm v2.9.0+[66] with e-value no greater than 1e-10 (e-value threshold was adapted for highly duplicated gene families) and using the protein sequence from *Medicago truncatula* as query. To remove short and partial transcript that are abundant in the transcriptomes from the 1KP project, shorter transcript than 60% of the *M. truncatula* query were removed prior alignment. Sequences were aligned using MAFFT v7.407[68] with default parameter and the resulting alignments trimmed using trimAl v1.2rev57[69] to remove all positions with more than 20% of gaps. Cleaned alignments were subjected to Maximum Likelihood (ML) approach using the IQ-TREE software v1.6.7[70]. Prior to ML analysis, best-fitted evolutionary model was tested for each alignment using ModelFinder[71] as implemented in IQ-TREE. Branch support was tested with 10,000 replicates of SH-alrt[72]. From these initial phylogenies, clade containing the candidate genes were extracted, removing distant homologs, and a new phylogenetic analysis performed as described above. Trees were visualized and annotated using the iTOL v4.4.2[73]. Following gene phylogenies, for genes missing from sequenced genomes, the tBLASTn search was repeated on the genome assembly. This lead to the identification of a number of non-annotated genes that are summarized in Supplementary Table 6, and included in Figure 1 and Supplementary Figure 1.

Synteny analysis. For the genes for which orthologs/homologs were not found in in the genome and transcriptome of *M. polymorpha* ssp. *ruderalis*, more detailed searches in the genomes of *M.*

polymorpha ssp. ruderalis, *M. polymorpha* ssp. polymorpha and *M. polymorpha* ssp. monitvagans were conducted. To do this, the coding sequences of the respective *M. paleacea* genes were used as queries for a BLASTn search against the genome assemblies of all three *M. polymorpha* subspecies. The genomic regions were then aligned with the corresponding *M. paleacea* genomic regions for comparisons using Mauve 2.3.1 and visualized in Geneious. Following this, gene syntenic analysis was performed by mapping the respective transcriptome onto the *M. polymorpha* ssp. *ruderalis* and *M. paleacea* genomic regions upstream and downstream of each gene using Geneious. The identities of the preceding and subsequent genes were confirmed by aligning the *M. paleacea* and *M. polymorpha* ssp. *ruderalis* sequences using MAFFT[68]. Their functional identity was confirmed using BLAST through searches against the SWISSPROT database and using Pfam.

Analysis of the Nelumbo CCaMK locus. Reference coding sequence of *Medicago truncatula CCaMK* was used as query to search for homolog sequence in three different *Nelumbo nucifera* genome assemblies[74,75]. Search was performed using the BLASTn+, with an e-value threshold set at 1e-10 and default parameters. For each hit in the BLAST results, 5 kb of upstream and downstream sequences were extracted using custom Python script. Extracted sequences were then aligned using MAFFT[68] with the genomic reference and the coding sequences of *M. truncatula* and *Oryza sativa*.

Signature of selection on CCaMK and CYCLOPS in mosses. With the exception of *Takakia lepidozioides*, mosses are not mycorrhizal[76]. We investigated the selection acting on *CCaMK* and *CYCLOPS* on the branch before the radiation of mosses but after the divergence of *T. lepidozioides* (foreground branch). We adopted branch models[77] to identify differential signatures of selective pressure acting on the sequences between the foreground branch and the rest of the tree (i.e. background branches) and branch-site models[78] to identify differential signatures of selective pressures acting on specific sites between foreground and background branches. These models are implemented in the codeml program[79] using ete-evol program[80]. In addition to the branch model, we also used the RELAX program[81] to look for relaxation (K<1) or intensification (K>1) of the strength of selection acting on the non-mycorrhizal mosses (mosses clade without *T. lepidozioides*) compared to the rest of the tree. These methods calculate different synonymous and nonsynonymous substitution rates ($\omega = \frac{d_N}{d_S}$) using the phylogenetic tree topology for both foreground and background branches. Because (i) the programs used do not accept gap in codon sequences and (ii) there is a negative correlation between the number of sequences and the number of ungapped positions, we used different number of sequences for each analysis. Protein

sequences from CCaMK and CYCLOPS orthologs were aligned using MUSCLE v3.8.3[82]. Short sequences were excluded to maximize sequence number while limiting gapped positions compared to *CCaMK/CYCLOPS* sequences of *T. lepidozioides* using a custom R script[83]. We opted for 124 sequences for *CCaMK* and 121 sequences for *CYCLOPS* corresponding to 1092 and 552 nucleotide positions, respectively (Supplementary Table 3). We used likelihood-ratio test (LRT) to assess significance between models ran with codeml. For the branch model, we compare likelihoods from the "b_free" and "M0" models to determine if the ratios are different between background and foreground branches[77] (*p-val*>0.05 : $\omega_{foreground}$ and $\omega_{background}$ are not different, *p-val*<0.05 : $\omega_{foreground}$ and $\omega_{background}$ are different). We also compared likelihoods from the "b_free" and "b_neut" models to determine if the ratio on the foreground branch is not neutral[77] (*p-val*>0.05: no signature of selection on foreground (neutral), *p-val*<0.05 and $\omega_{foreground}$<1: signature of negative/relaxed selection on foreground, *p-val*<0.05 and $\omega_{foreground}$>1: signature of positive selection on foreground). For the branch-site model, we compare likelihoods from the "bsA" and "M1" models to determine if the selection pressure is relaxed on sites on the foreground branch[78] (*p-val*<0.05: signature of relaxed selection on sites on foreground). We also compared likelihoods from the "bsA" and "bsA1" models to determine if there are sites under positive selection on the foreground branch[78] (*p-val*<0.05: signature of positive selection on specific sites on foreground).

MpoCCaMK and MpoCYCLOPS pseudogenes in wild populations

Plant material and DNA extraction. Samples from a total of 35 accessions of *Marchantia polymorpha* (Supplementary Table 4) were collected. Approximately 0.1 g of cleaned young thalli was reduced to a fine powder in liquid nitrogen and cleaned as follow. First, 1 ml of cleaning STE buffer (Sucrose 0.25 M, Tris-HCl 0.03 M at pH=7.5 and EDTA 0.05M at pH=7.5) was added to the powder and the mixture was thoroughly vortexed. After 5 min of centrifugation at 14, 000 rpm for 5 min, the supernatant was removed. Washes with STE buffer were repeated until the red coloration was completely gone. Subsequent DNA extraction was performed using a CTAB method[58].

PCR and sequencing. Putative *MpoCCaMK* and *MpoCYCLOPS* genomic sequences (1 kb) were amplified by PCR using GoTaq polymerase (Promega) according to manufacturer's protocol. Primers used were (*MpoCCAMK* fwd GGTCATCTTGTACATTCTCCT, rev CTTGCTACTGAAGATACTTGCA; *MpoCYCLOPS* fwd CAGCACCCTCAAAACTTAGAGT, *MpoCYCLOPS* rev CCCAGCTGTTACCTTAAGAAG). After amplification, PCR products were sent for Sanger sequencing at Eurofins (Germany) using the PCR forward and reverse primers of the two genes. Sanger sequences were assembled using the Tracy software v0.5.5 and aligned

using MUSCLE[82] v3.8.3 with genomic *CCaMK* and *CYCLOPS* sequences from the three *Marchantia polymorpha* subspecies.

Mycorrhization assays

Material and growth conditions. Thalli of *Marchantia paleacea*, *Marchantia polymorpha* ssp. ruderalis (TAK1), *Marchantia polymorpha* ssp. *polymorpha* (MppBR5) and *Marchantia polymorpha* ssp. *montivagans* (MpmSA2) were grown on a substrate containing 50% Oil-Dri UK (Damolin, Etrechy, France) and 50% sand (0.7–1.3 mm) in 9x9x8 cm pots (5 thalli per pot, 5 pots per species/subspecies). Each pot was inoculated with 1000 sterile spores of *Rhizophagus irregularis* DAOM 197198 (Agronutrition Labège, France), grown at a 16h/8h photoperiod at 22°C/20°C and watered once a week with ½ Long Ashton medium containing 7.5 µM of phosphate[84].

Microscopic phenotyping. After 8 weeks of colonization, thalli were harvested Sections of the thalli were made by hand or using a Leica vt1000s vibratome (100µm sections) on thalli included in 6% agarose. Sections were cleared in 10% KOH overnight and ink colored one hour in 5% Sheaffer ink, 5% acetic acid. Sections were observed under a microscope and pictures of representative sections were taken with a Zeiss Axiozoom V16 microscope.

Molecular phenotyping. After 8 weeks of colonization, portions of 5 thalli per condition were harvested and frozen in liquid nitrogen. DNA was extracted from frozen thalli material as described in Edwards[85]. PCR was performed using GoTaq polymerase (Promega) following manufacturer instructions on the *M. polymorpha Elongation factor-1 alpha* (Mapoly0024s0116, fwd CGACCACTGGTCACCTTATC, rev AACCTCAGTGGTCAGACCG) and on the *Rhizophagus irregularis Elongation factor-1 alpha* (Rhiir2_1_GeneCatalog_transcripts_20160502.nt, fwd TGTTGCTTTCGTCCCAATATC, rev CTCAACACACATCGGTTTGG)

Medicago truncatula root assays

Plasmid construction. The Golden Gate modular cloning system was used to prepare the plasmids as described in Schiessl *et al*[86]. Levels 0 and 1 used in this study are listed in Supplemental Table 6 and held for distribution in the ENSA project core collection (https://www.ensa.ac. uk/). We used L1 plasmid piCH47811[87], L1 construct L1M-R1-pAtUBI-DSred-t35S[23] and L2 acceptor backbone EC50507. Sequences were domesticated (listed in Supplemental Table 7) synthesized and cloned into pMS (GeneArt, Thermo Fisher Scientific, Waltham, USA).

Activation of pENOD11:GUS by CCaMK orthologs. Constructs containing CCaMK-K were transformed in *A. rhizogenes* A4TC24 by electroporation. Transformed strains were grown at 28°C in Luria-Bertani medium supplemented with rifampicin and kanamycin (20 µg/mL). *M. truncatula pENOD11:GUS* L416 roots[88], were transformed with the different CCaMK-K as described by Boisson-Dernier *et al.*[89], and grown on Fahraeus medium for 2 months, selected with the DsRed marker present in all the constructs and GUS stained as in Vernié *et al*[90].

Trans-complementation assays. The *MedtruCCaMK* null allele *dmi3-1*[91] and *MedtruCYCLOPS* null allele *ipd3-1*[92] were used in this study. Roots were transformed as described above with the various *CCaMK* and *CYCLOPS* constructs respectively (Supplementary Table 7). For nodulation assays, *M. truncatula* plants with DsRed positive roots were transferred in pots containing Zeolite substrate (50% fraction 1.0-2,5mm, 50% fraction 0,5-1.0-mm, Symbiom) and watered with liquid Fahraeus medium. Wild-type *S. meliloti* RCR2011 pXLGD4 (GMI6526) was grown at 28°C in tryptone yeast medium supplemented with 6 mM calcium chloride and 10 µg/mL tetracycline, rinsed with water and diluted at OD_{600}=0.02. Each pot was inoculated with 10 ml of bacterial suspension. Two independent biological replicates were conducted for *ccamk* and *cyclops* complementation assays. For *ccamk*, one biological assay was harvested at 26DPI and scored for the ratio of infected nodules after X-Gal staining, and one at 42DPI scored for the total nodule number. For *cyclops*, one biological assay was harvested at 32DPI and one at 49 DPI, both were scored for infected nodules after X-Gal staining and the ratio of fully infected nodules out of the total number of nodules was calculated and used on the box plot (both biological replicates combined).

References in Supplementary methods

66. Camacho, C. *et al.* BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, (2009).

67. Finn, R. D. *et al.* Pfam: The protein families database. *Nucleic Acids Res.* **42**, D222-30 (2014).

68. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

69. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973 (2009).

70. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

71. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).

72. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307-321 (2010).

73. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* (2019).

74. Ming, R. *et al.* Genome of the long-living sacred lotus (*Nelumbo nucifera* Gaertn.). *Genome Biol.* (2013).

75. Gui, S. *et al.* Improving Nelumbo nucifera genome assemblies using high-resolution genetic maps and BioNano genome mapping reveals ancient chromosome rearrangements. *Plant J.* (2018).

76. Read, D. J., Duckett, J. G., Francis, R., Ligrone, R. & Russell, A. Symbiotic fungal associations in 'lower' land plants. in *Philosophical Transactions of the Royal Society B: Biological Sciences* (2000).

77. Yang, Z. & Nielsent, R. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* (2002).

78. Zhang, J., Nielsen, R. & Yang, Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* (2005).

79. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).

80. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).

81. Wertheim, J. O., Murrell, B., Smith, M. D., Pond, S. L. K. & Scheffler, K. RELAX: Detecting relaxed selection in a phylogenetic framework. *Mol. Biol. Evol.* **32**, 820-832

(2015).

82. Edgar, R. C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

83. Team, R. C. R: A Language and Environment for Statistical Computing. *Vienna, Austria* (2019).

84. Balzergue, C., Puech-Pags, V., Bécard, G. & Rochange, S. F. The regulation of arbuscular mycorrhizal symbiosis by phosphate in pea involves early and systemic signalling events. *J. Exp. Bot.* **62**, 1049–1060 (2011).

85. Edwards, K., Johnstone, C. & Thompson, C. A simple and rapid method for the preparation of plant genomic DNA for PCR analysis. *Nucleic Acids Research* **19**, 1349 (1991).

86. Schiessl, K. *et al.* NODULE INCEPTION Recruits the Lateral Root Developmental Program for Symbiotic Nodule Organogenesis in *Medicago truncatula*. *Curr. Biol.* (2019).

87. Engler, C. *et al.* A Golden Gate modular cloning toolbox for plants. *ACS Synth. Biol.* **3**, 839–843 (2014).

88. Boisson-Dernier, A. *et al. MtENOD11* Gene Activation During Rhizobial Infection and Mycorrhizal Arbuscule Development Requires a Common AT-Rich-Containing Regulatory Sequence. *Mol. Plant-Microbe Interact.* **18**, 1269–1276 (2005).

89. Boisson-Dernier, A. *et al. Agrobacterium rhizogenes* -Transformed Roots of *Medicago truncatula* for the Study of Nitrogen-Fixing and Endomycorrhizal Symbiotic Associations. *Mol. Plant-Microbe Interact.* **14**, 695–700 (2001).

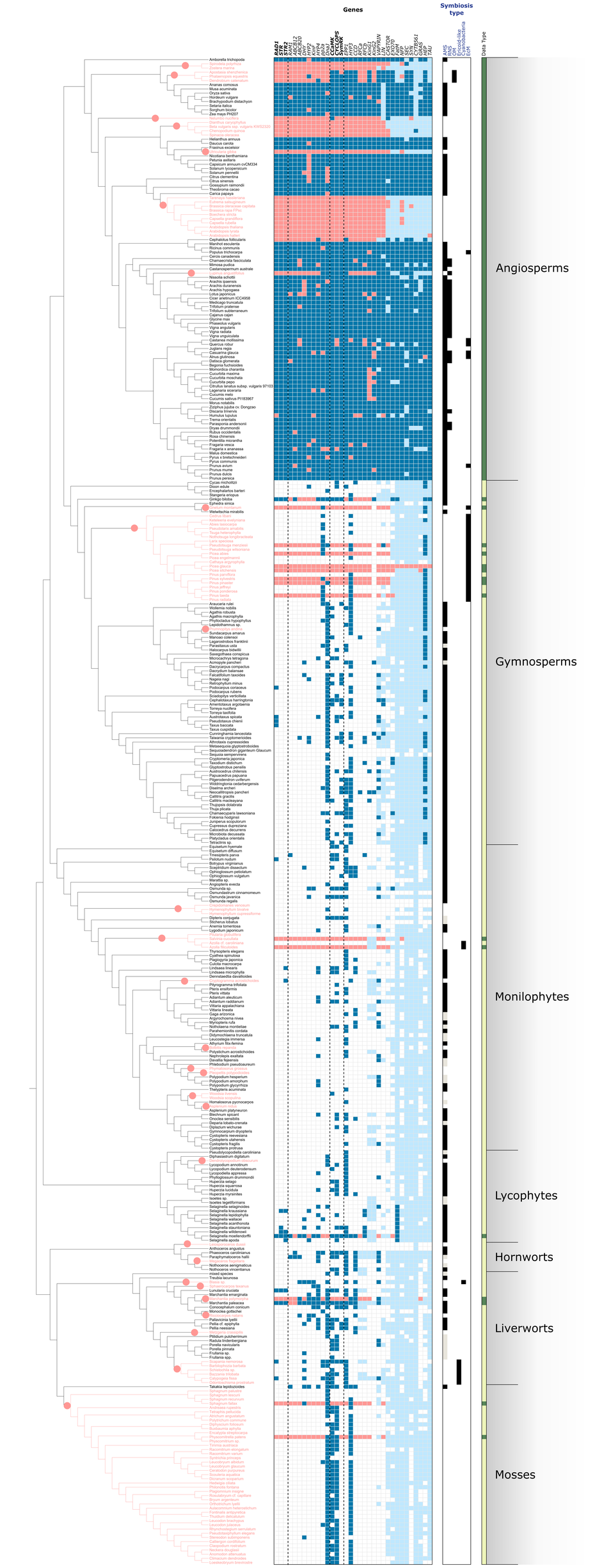
90. Vernié, T. *et al.* The NIN Transcription Factor Coordinates Diverse Nodulation Programs in Different Tissues of the *Medicago truncatula* Root. *Plant Cell* **27**, 3410–3424 (2015).

91. Catoira, R. *et al.* Four genes of Medicago truncatula controlling components of a Nod factor transduction pathway. *Plant Cell* **12**, 1647-1666 (2000).

92. Horváth, B. *et al. Medicago truncatula IPD3* Is a Member of the Common Symbiotic Signaling Pathway Required for Rhizobial and Mycorrhizal Symbioses. *Mol. Plant-Microbe Interact.* **24**, 1345–1358 (2011).

Supplementary Figures

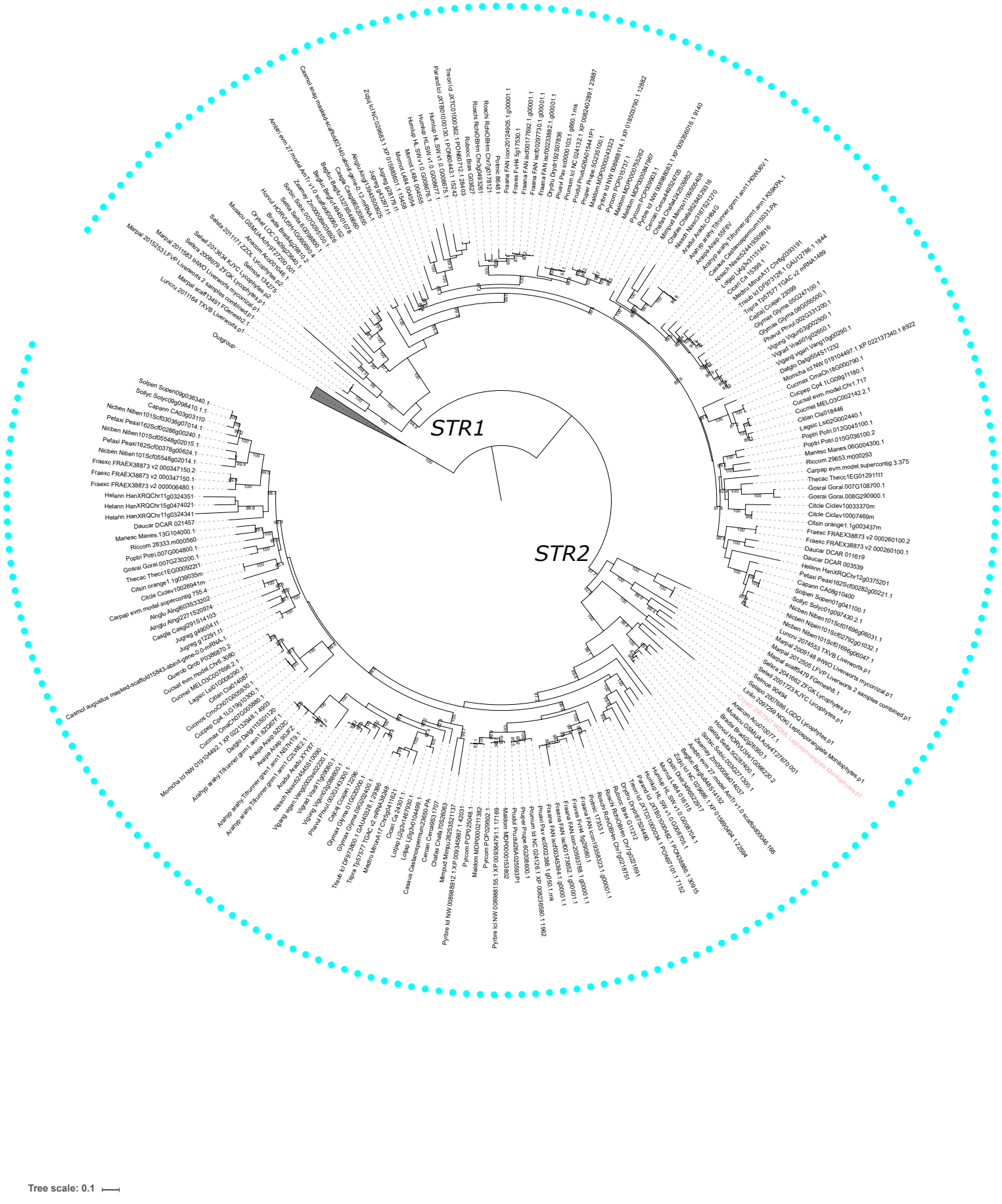
Supplementary Figure 1: Detailed view of Figure 1 with species name added on terminal branches.



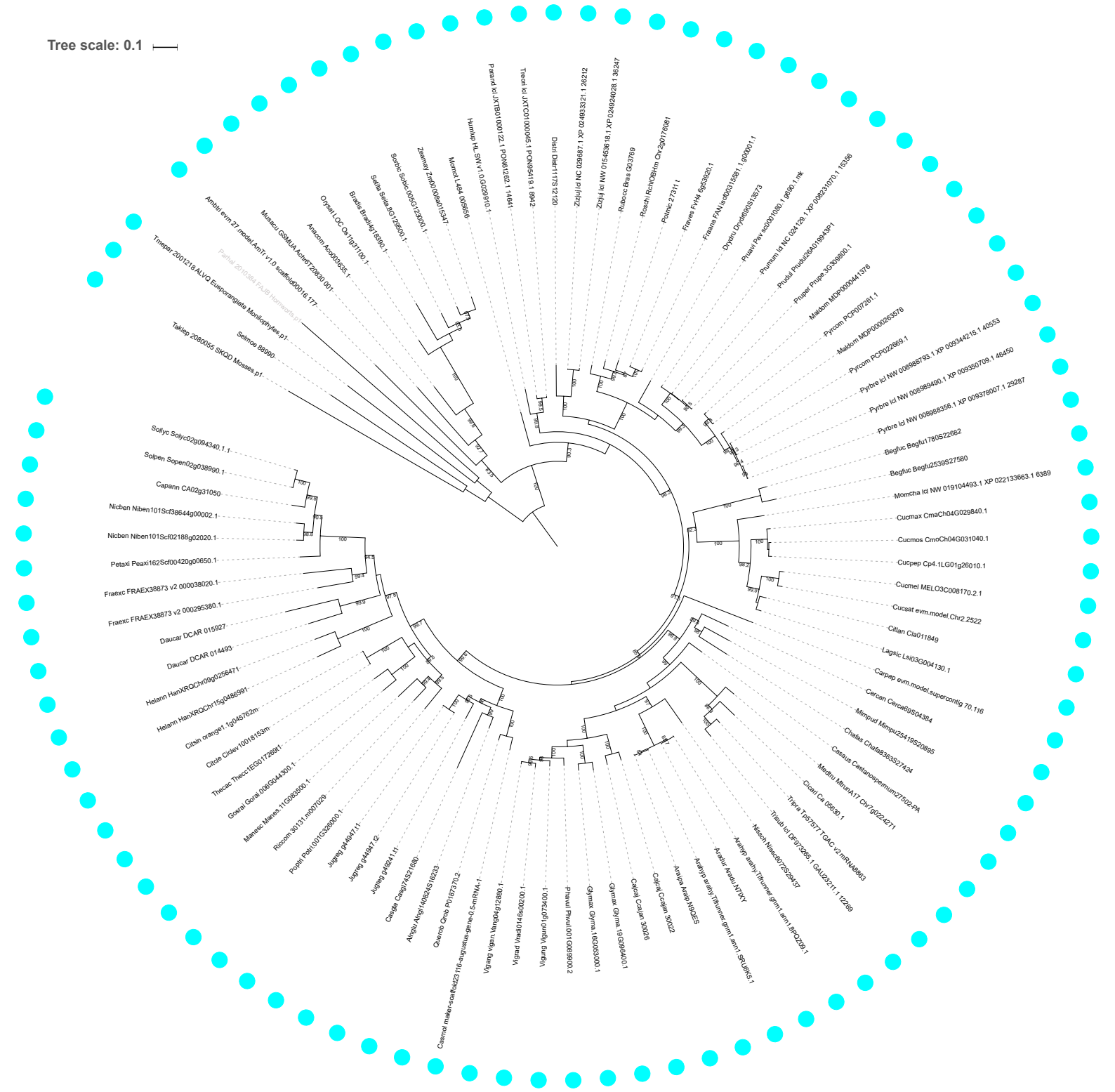
Supplementary Figure 2: Maximum likelihood tree of *RAD1*.

Tree scale: 0.1

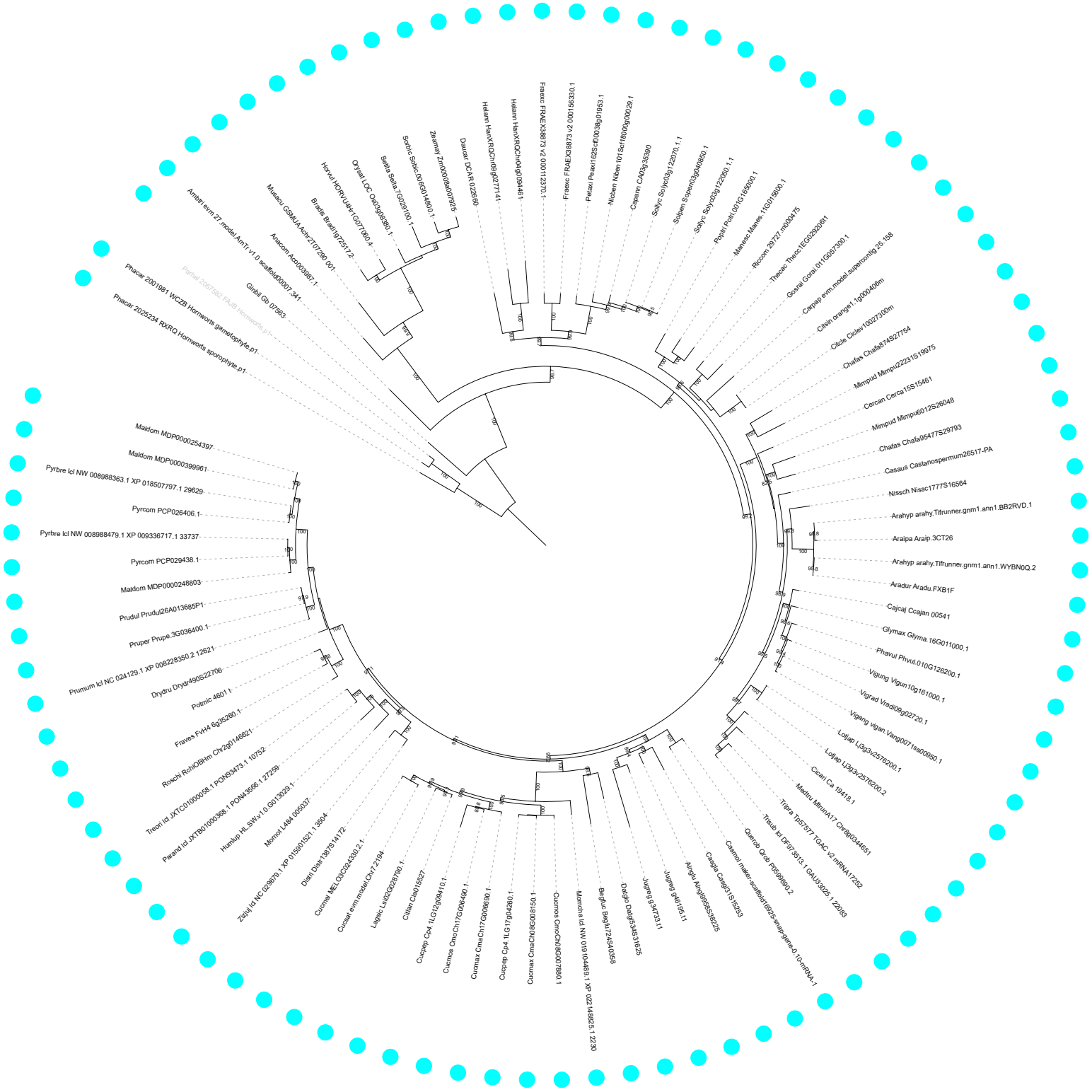
Supplementary Figure 3: Maximum likelihood tree of STR1/STR2.



Supplementary Figure 4: Maximum likelihood tree of *RAM1*.



Supplementary Figure 5: Maximum likelihood tree of *ABCB12*.



Tree scale: 0.1

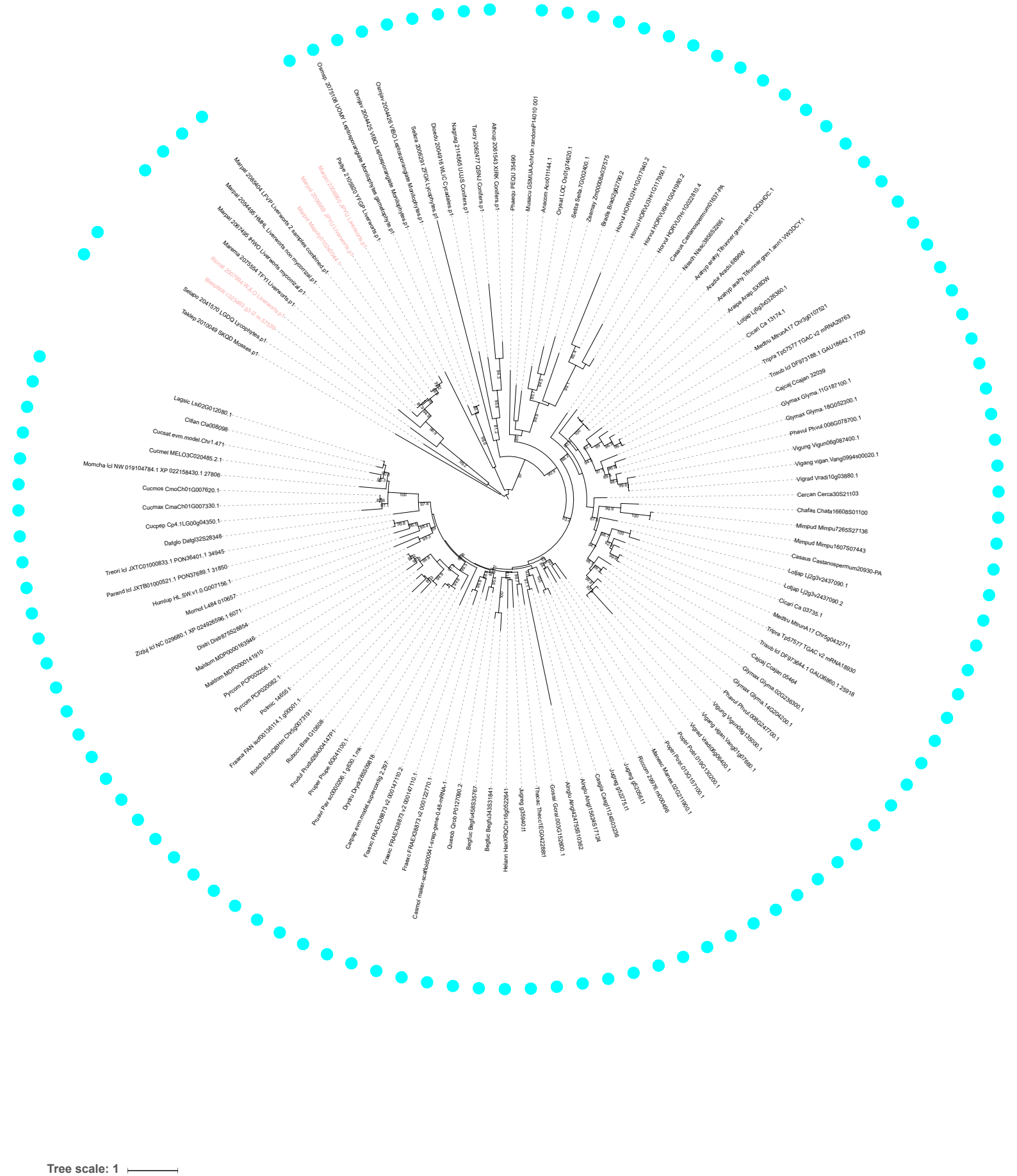
Supplementary Figure 6: Maximum likelihood tree of *ABCB20*.

Tree scale: 0.1

Supplementary Figure 7: Maximum likelihood tree of *DHY*.

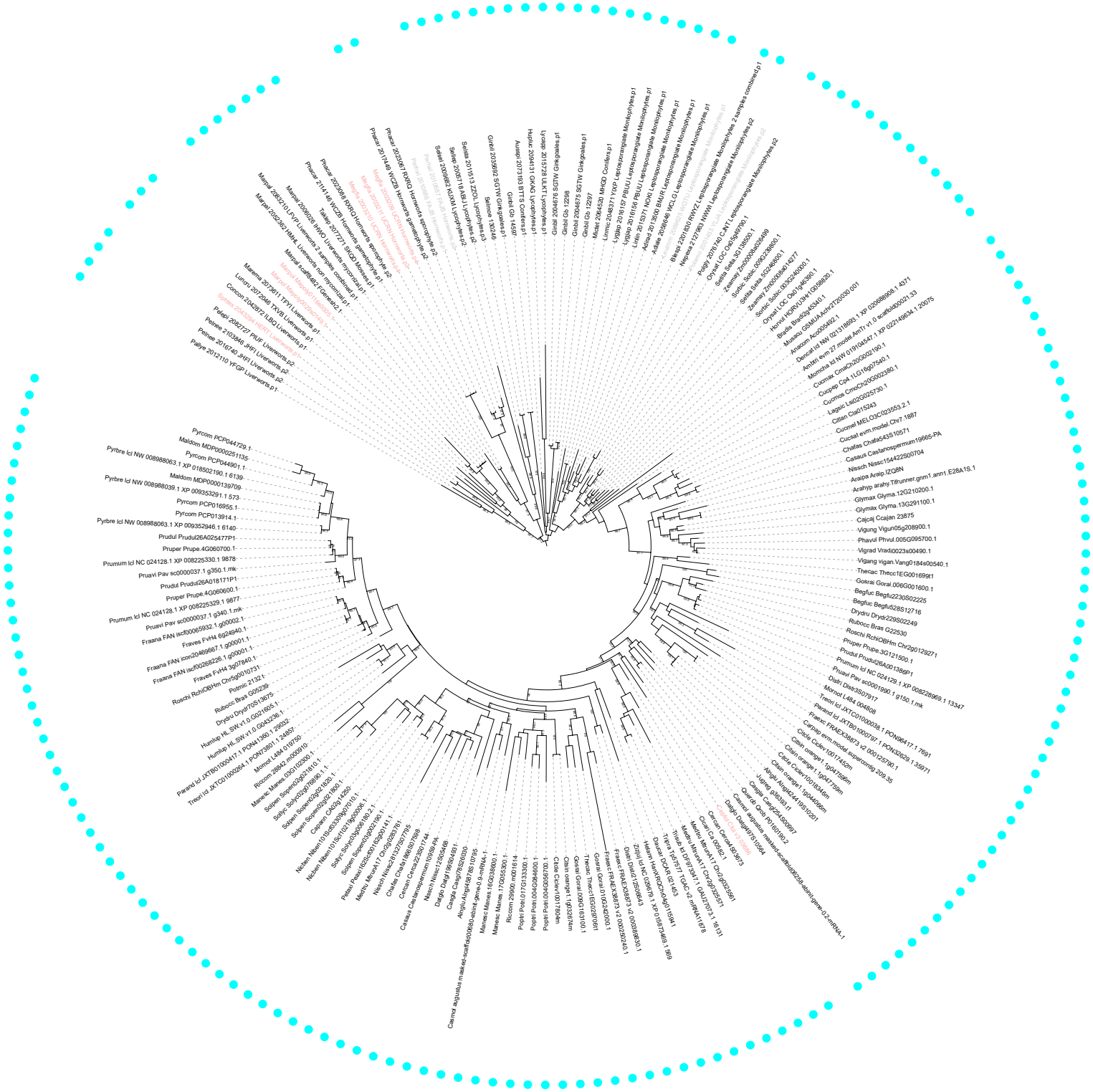
Tree scale: 0.1

Supplementary Figure 8: Maximum likelihood tree of *HYP2*.



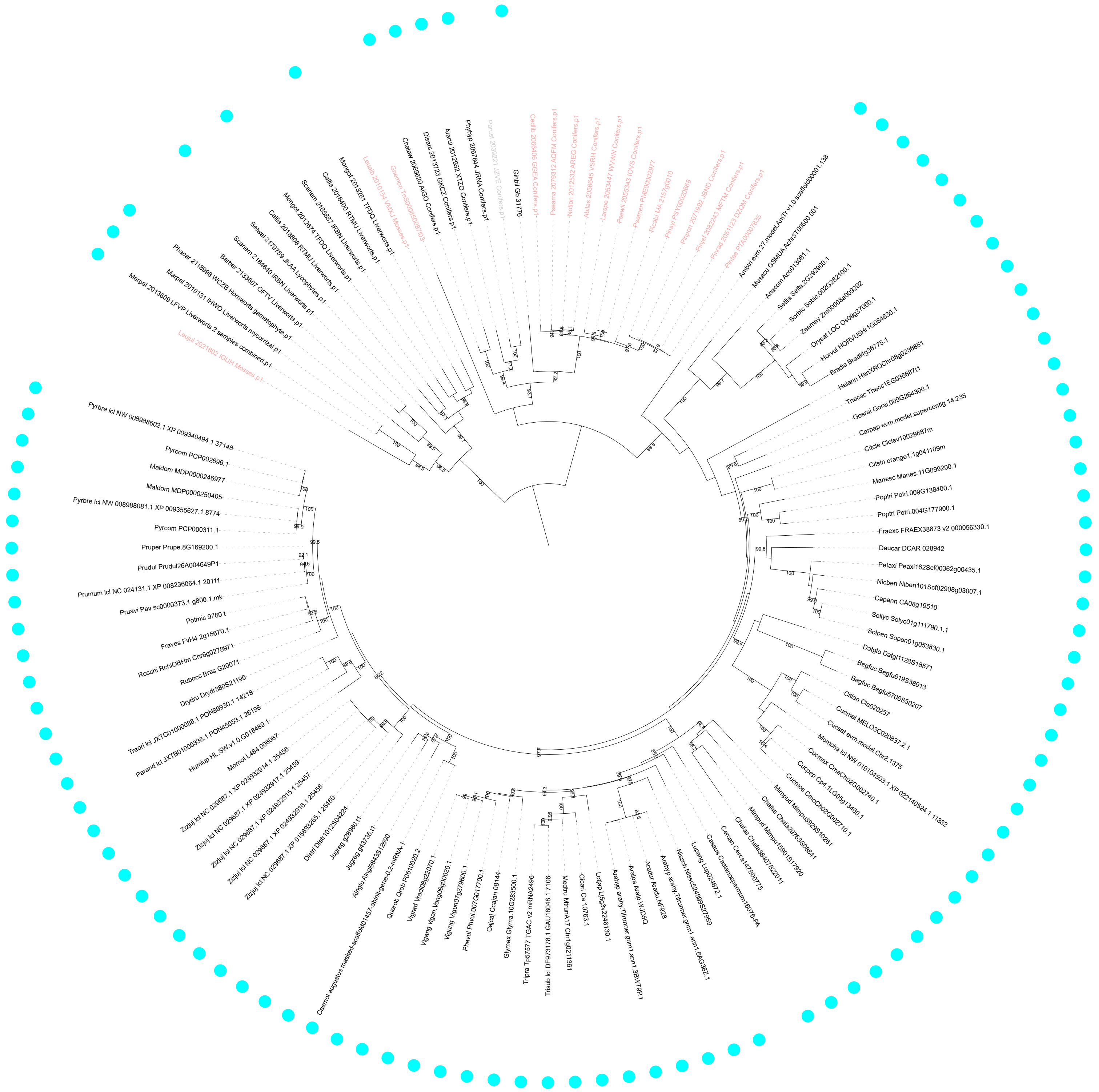
Supplementary Figure 9: Maximum likelihood tree of *KinF*.

Supplementary Figure 10: Maximum likelihood tree of *HYP4*.



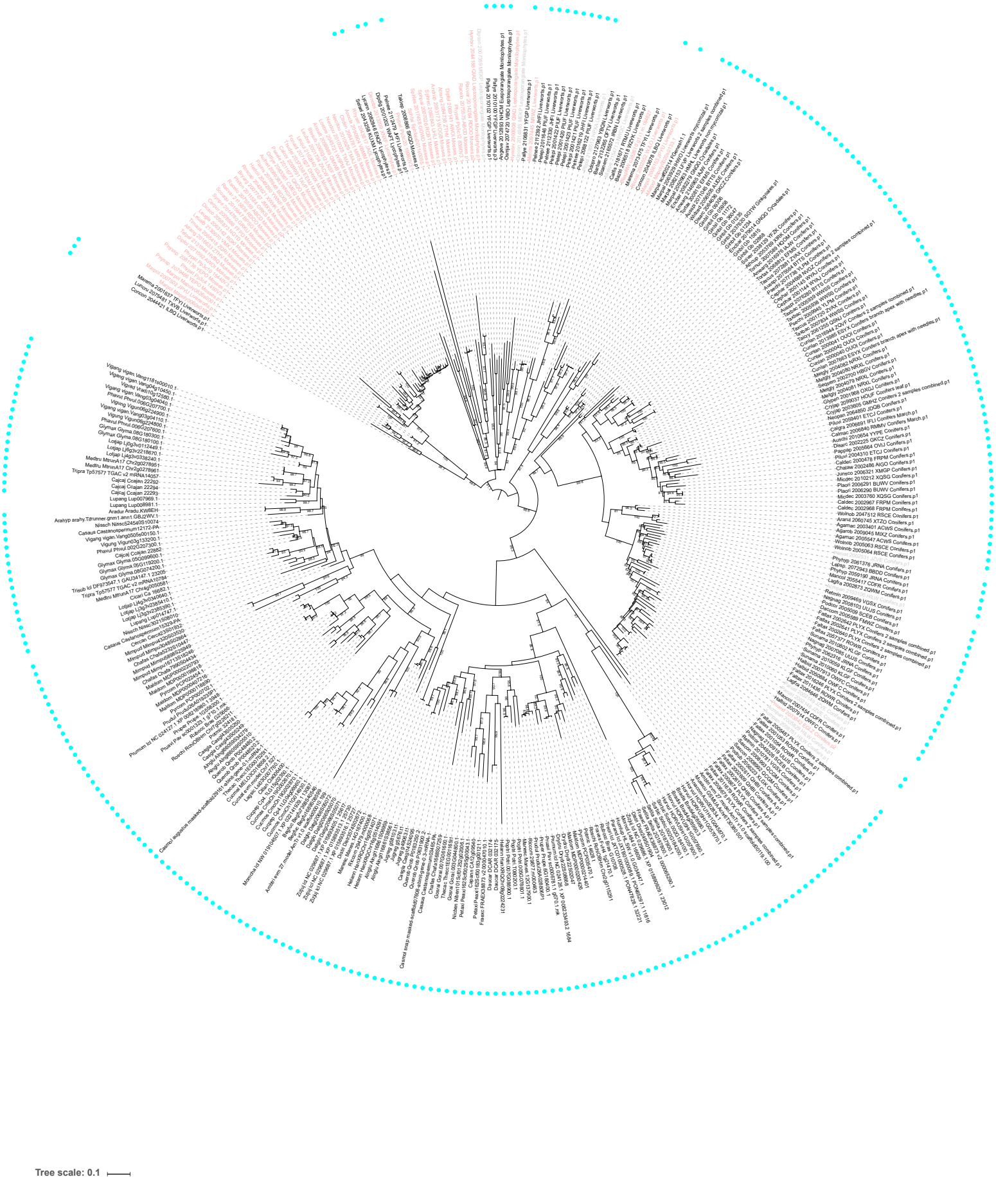
Tree scale: 0.1

Supplementary Figure 11: Maximum likelihood tree of *pp2a*.



Tree scale: 0.1

Supplementary Figure 12: Maximum likelihood tree of *DnaJ*.

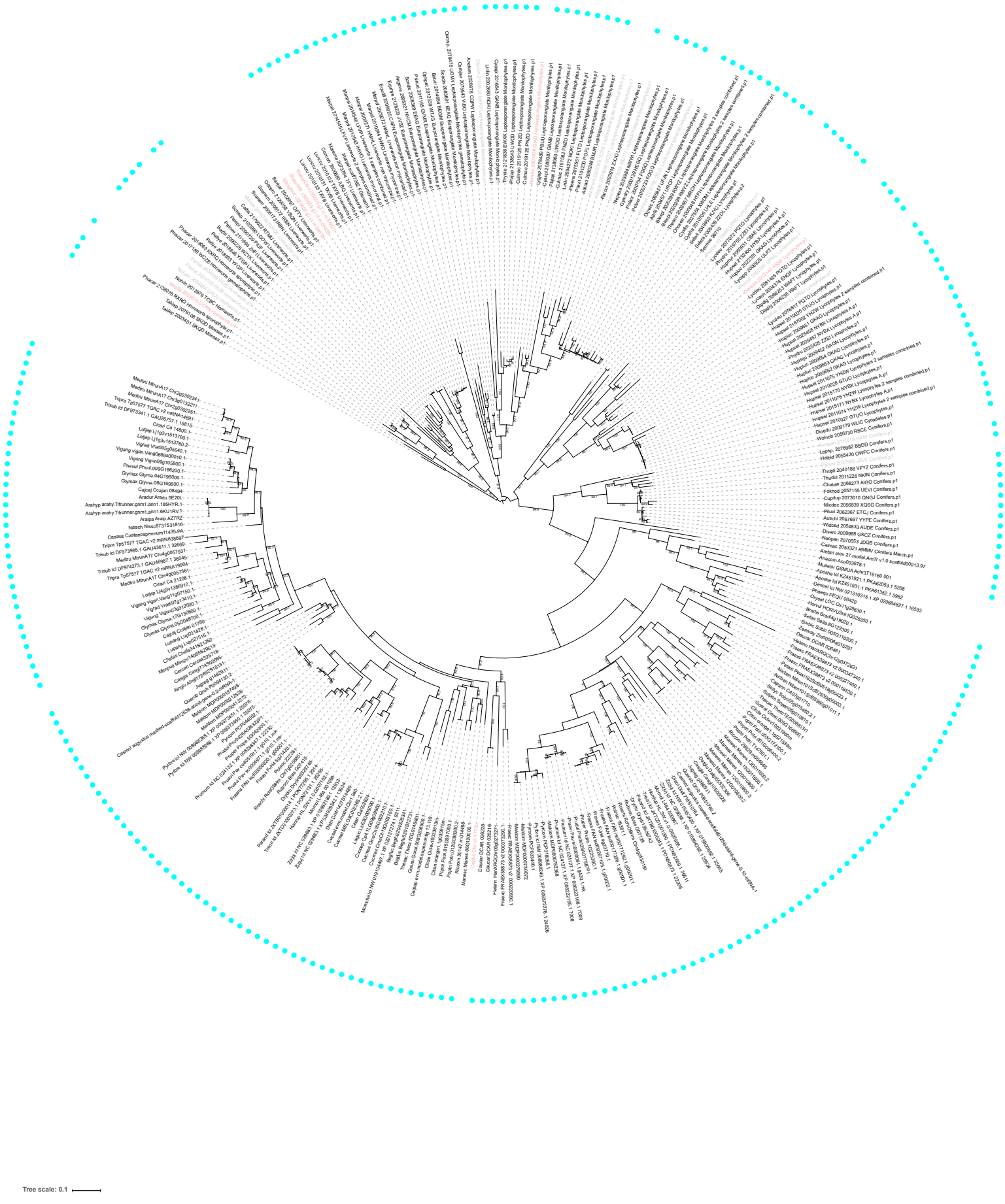


Supplementary Figure 13: Maximum likelihood tree of *CCaMK*.

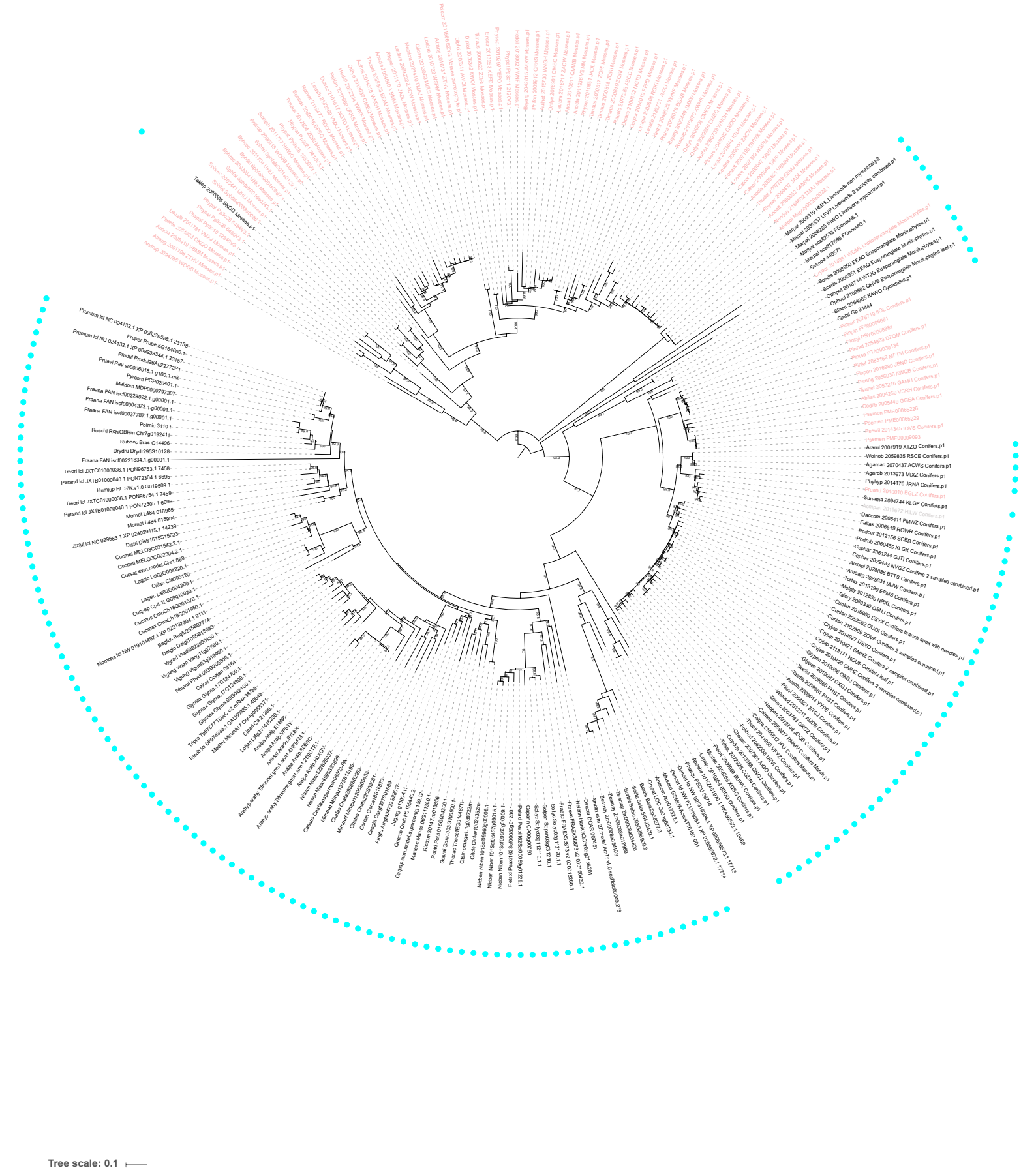
Supplementary Figure 14: Maximum likelihood tree of *CYCLOPS*

Supplementary Figure 15: Maximum likelihood tree of *SymRK*.

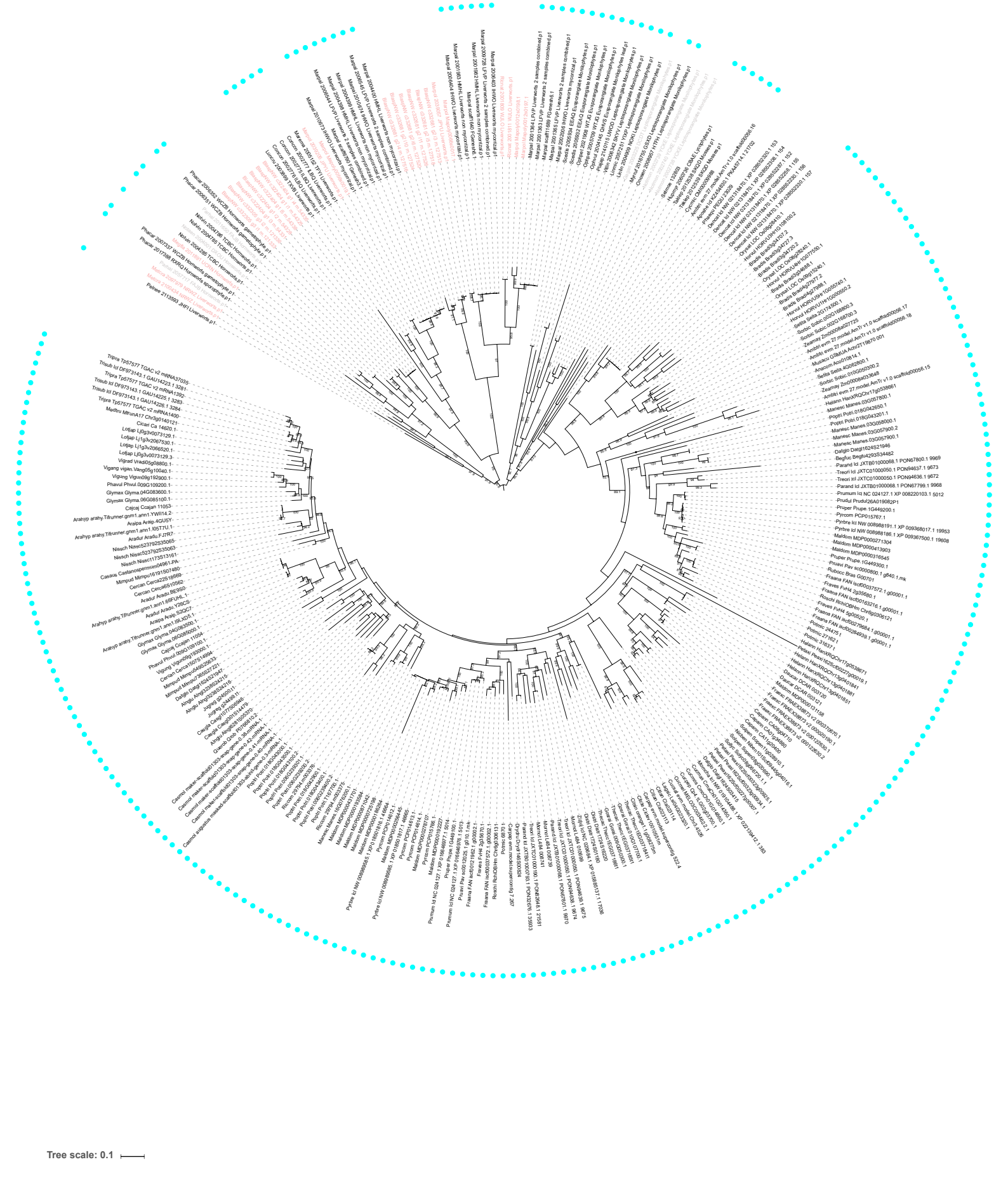
Supplementary Figure 16: Maximum likelihood tree of *HYP*.



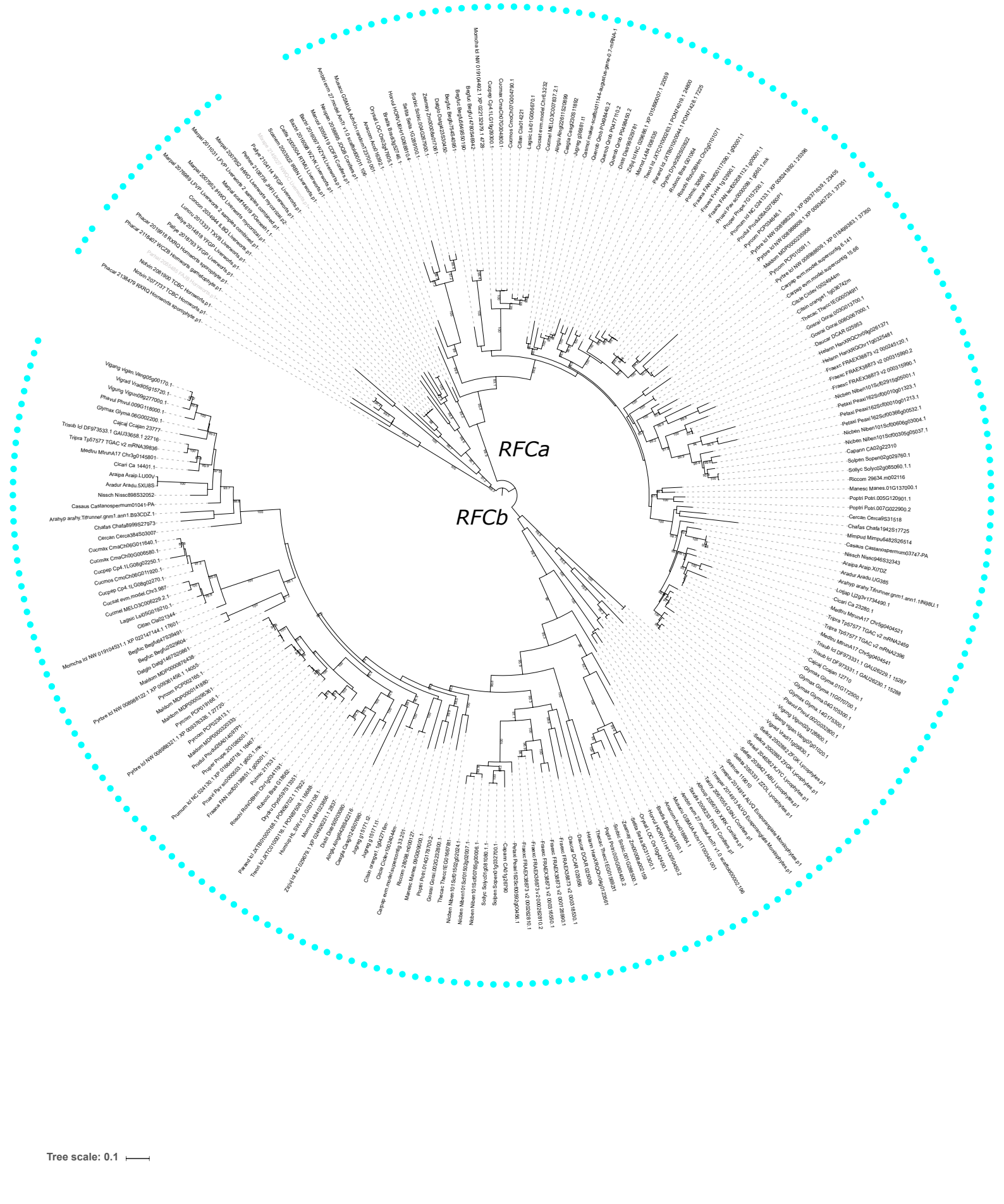
Supplementary Figure 17: Maximum likelihood tree of *HYP3*.



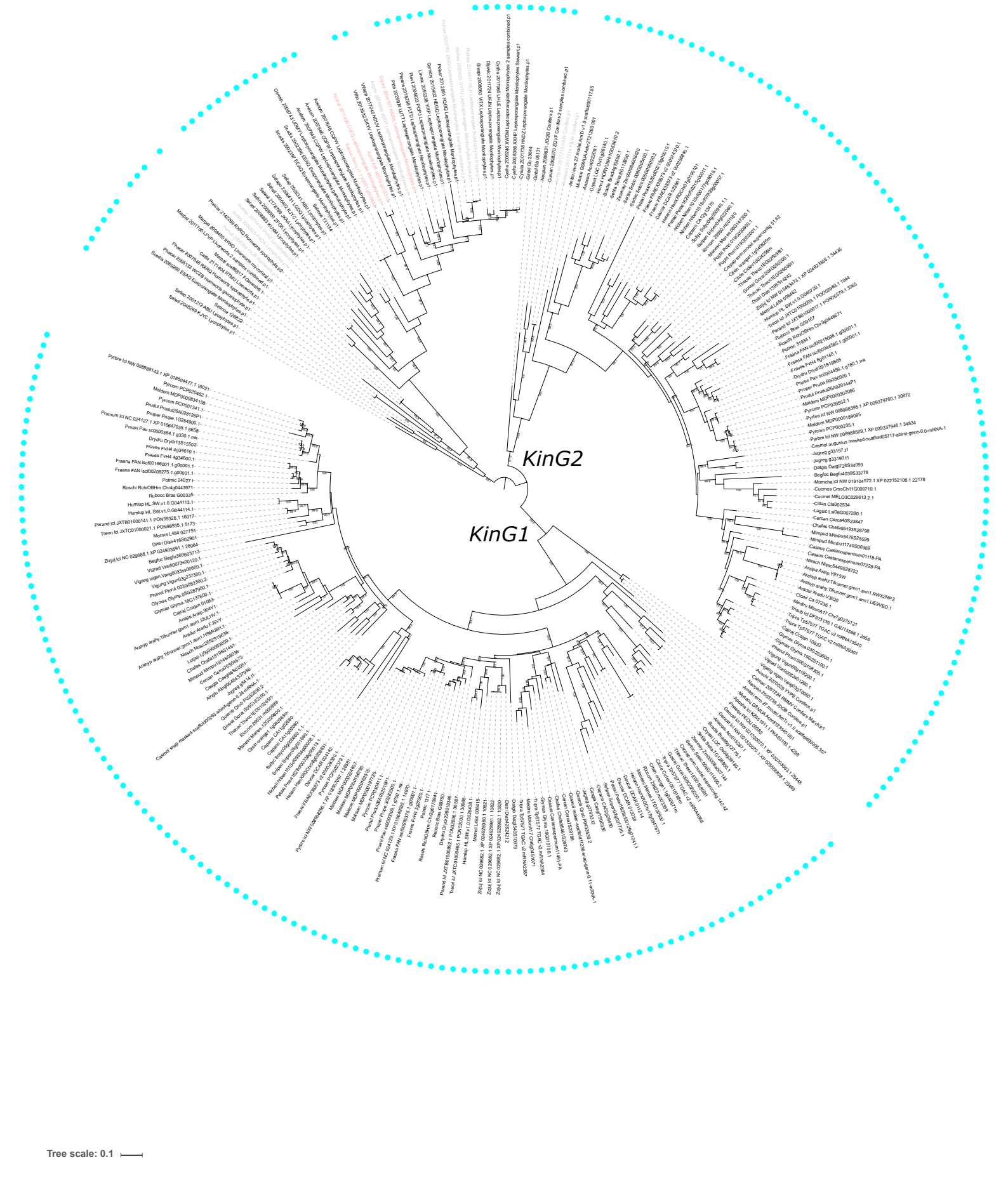
Supplementary Figure 18: Maximum likelihood tree of *CCD*.



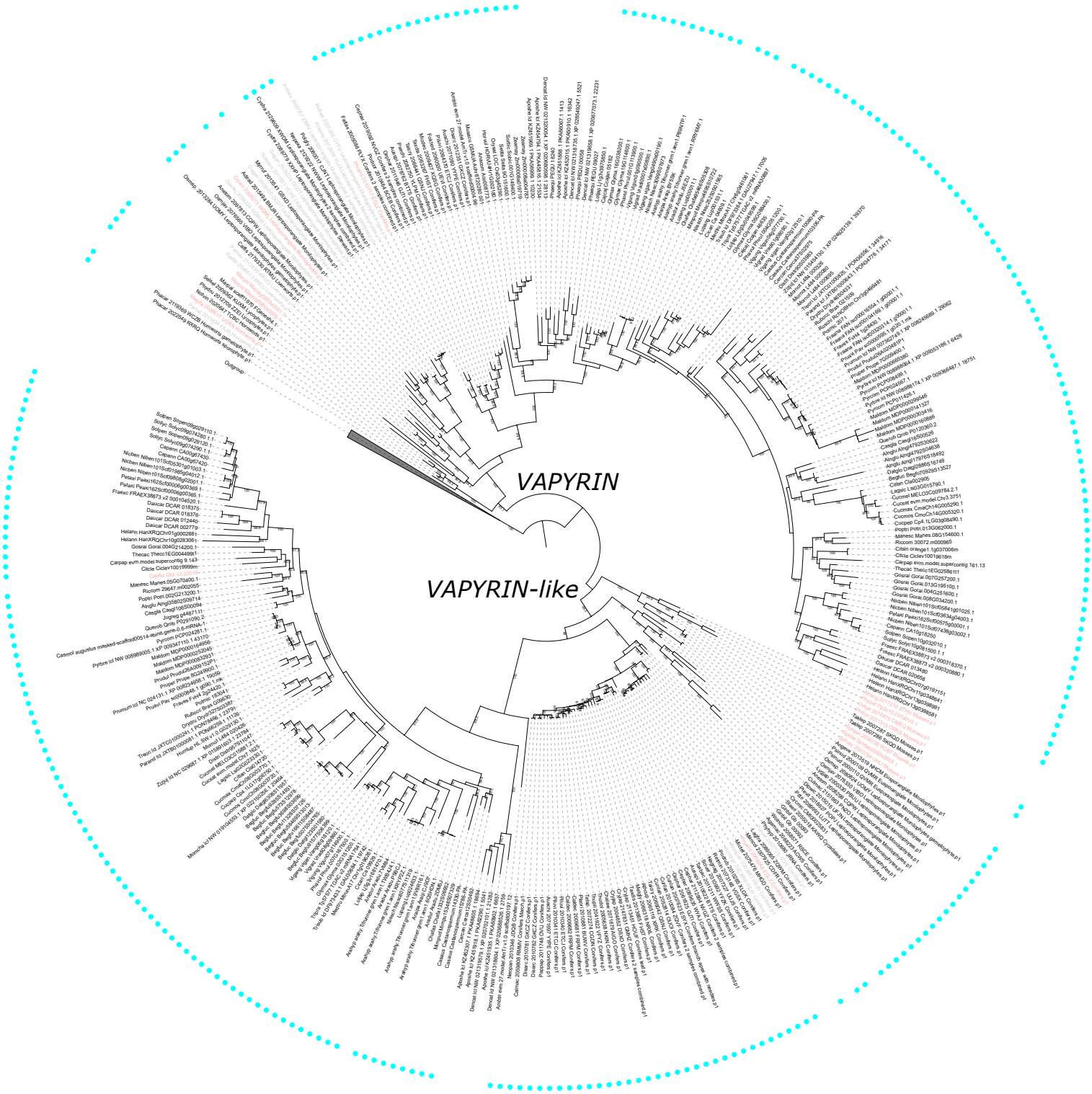
Supplementary Figure 19: Maximum likelihood tree of *RFCa/RFCb*.



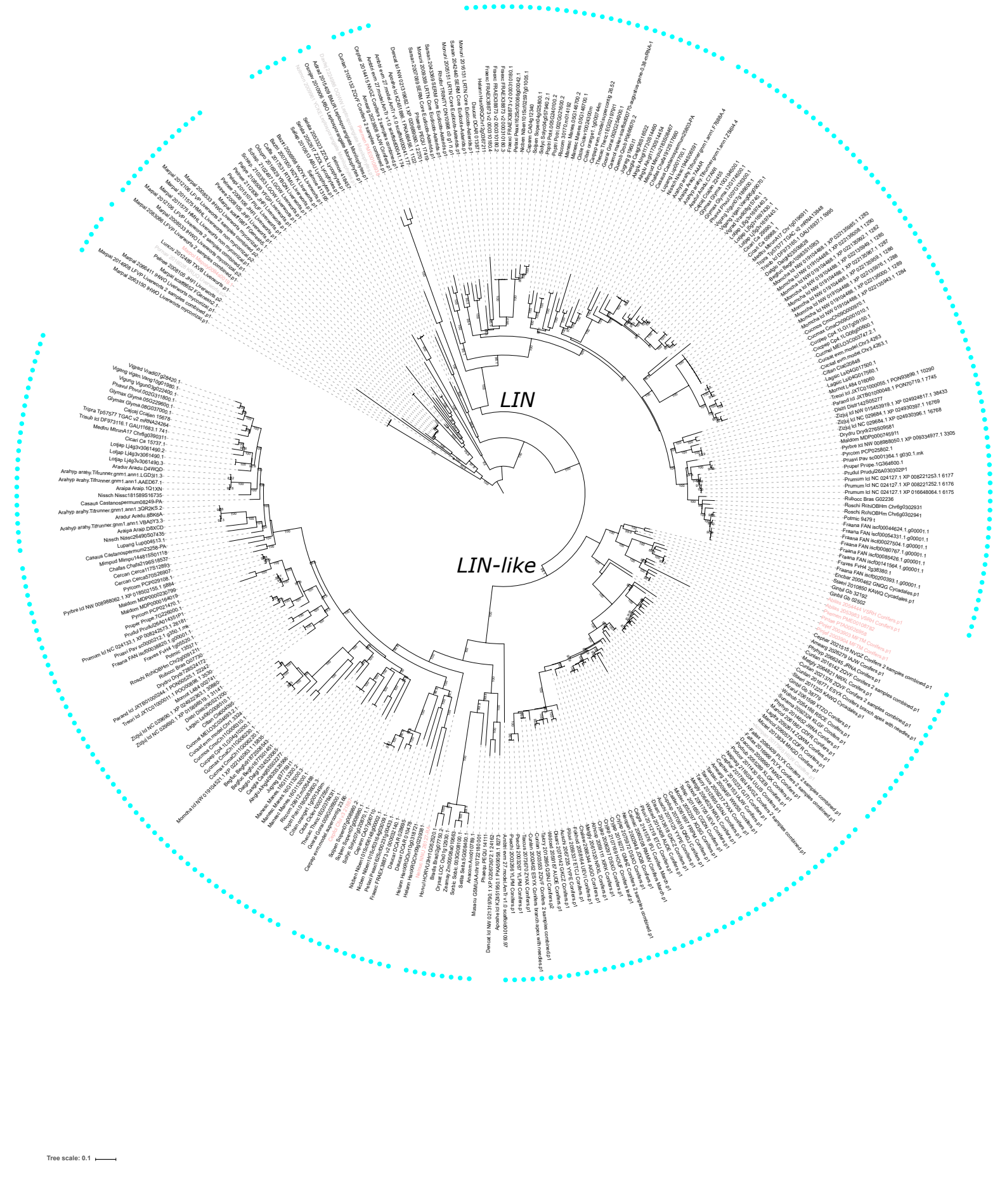
Supplementary Figure 20: Maximum likelihood tree of *KinG1/KinG2*.



Supplementary Figure 21: Maximum likelihood tree of *VAPYRIN/VAPYRIN-like*.

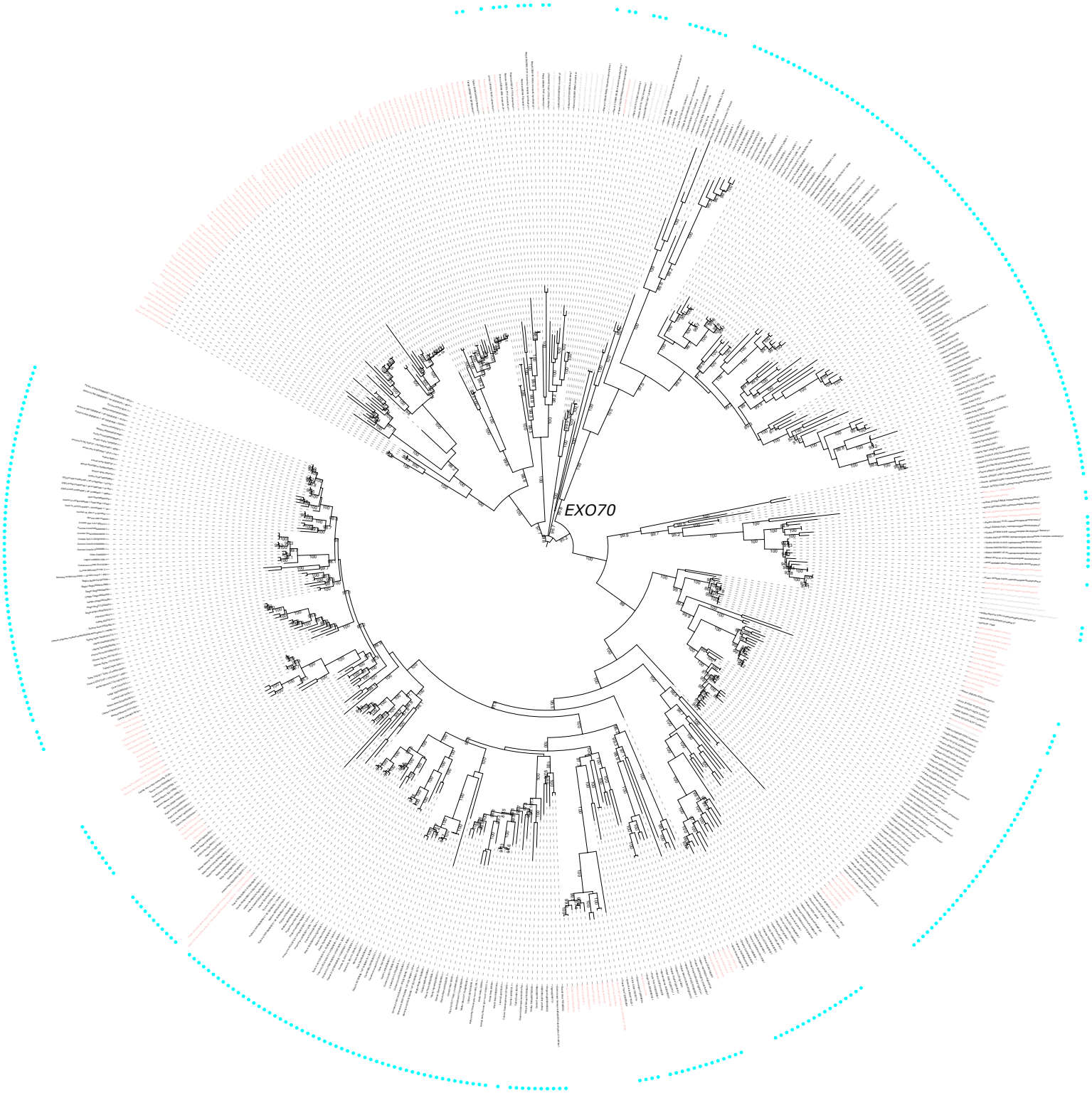


Supplementary Figure 22: Maximum likelihood tree of *LIN/LIN-like*.

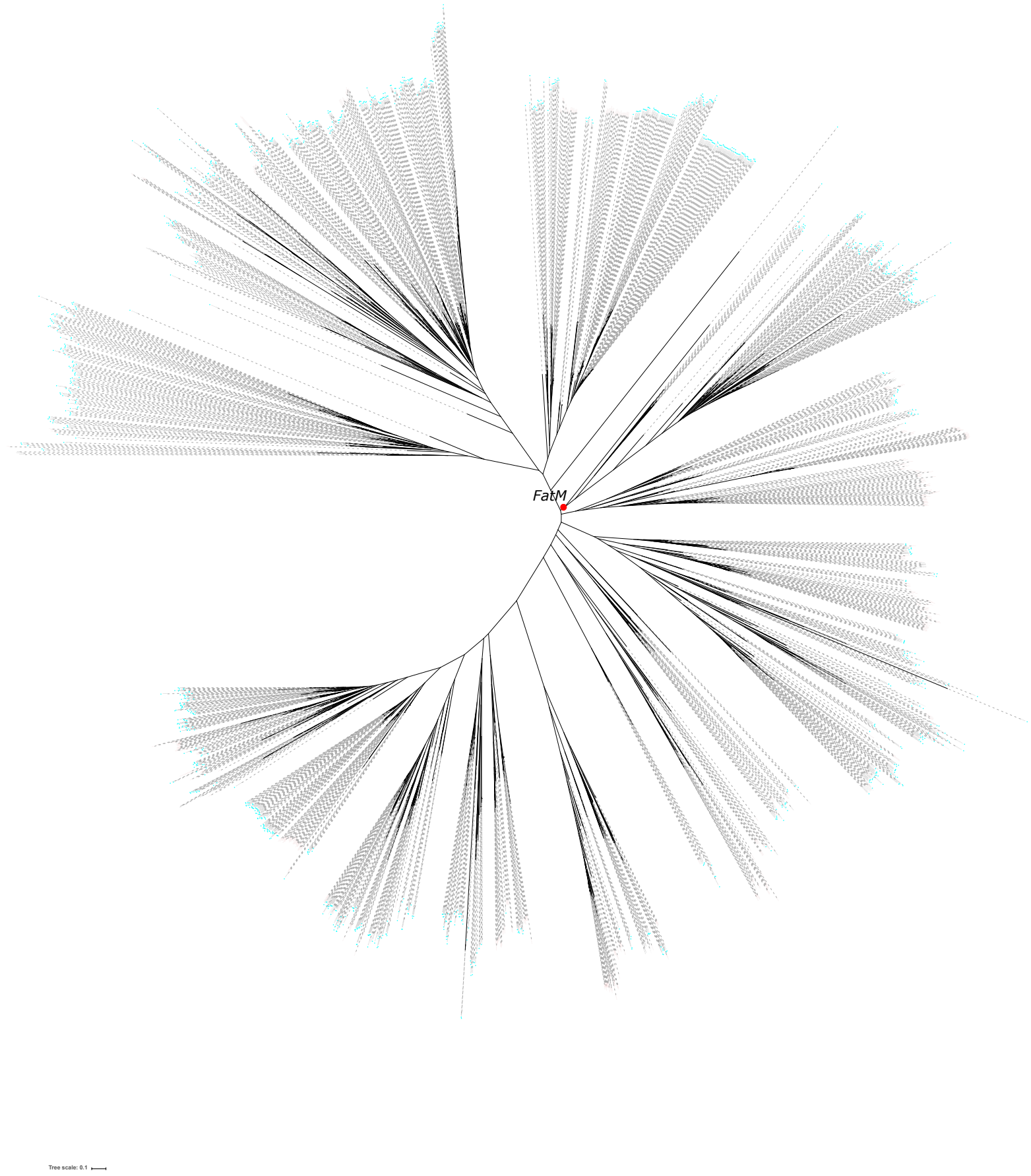


Supplementary Figure 23: Maximum likelihood tree of *CASTOR*.

Supplementary Figure 24: Maximum likelihood tree of *EXO70*.

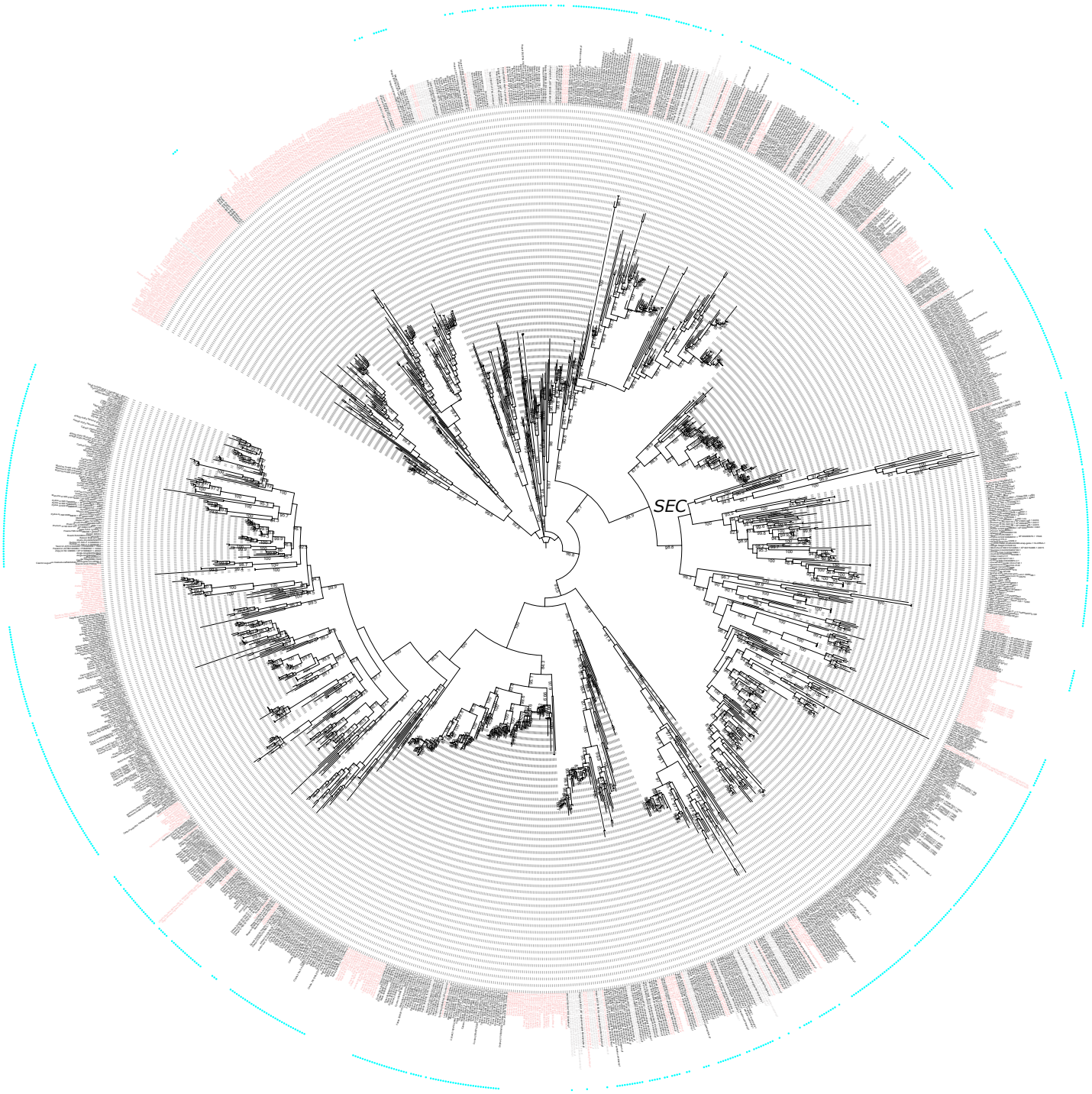


Supplementary Figure 25: Maximum likelihood tree of *FatM*.

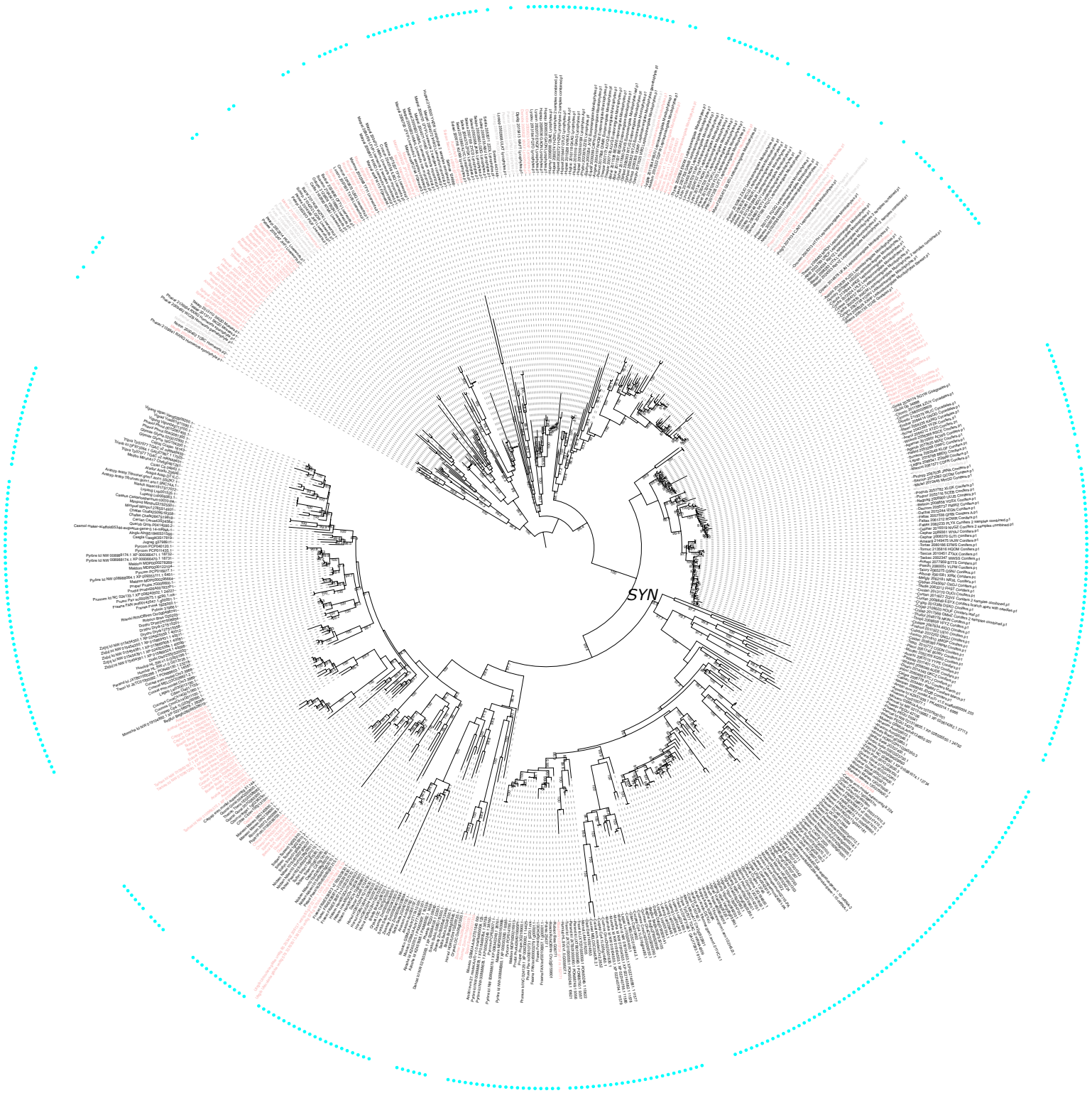


Supplementary Figure 26: Maximum likelihood tree of *NFP*.

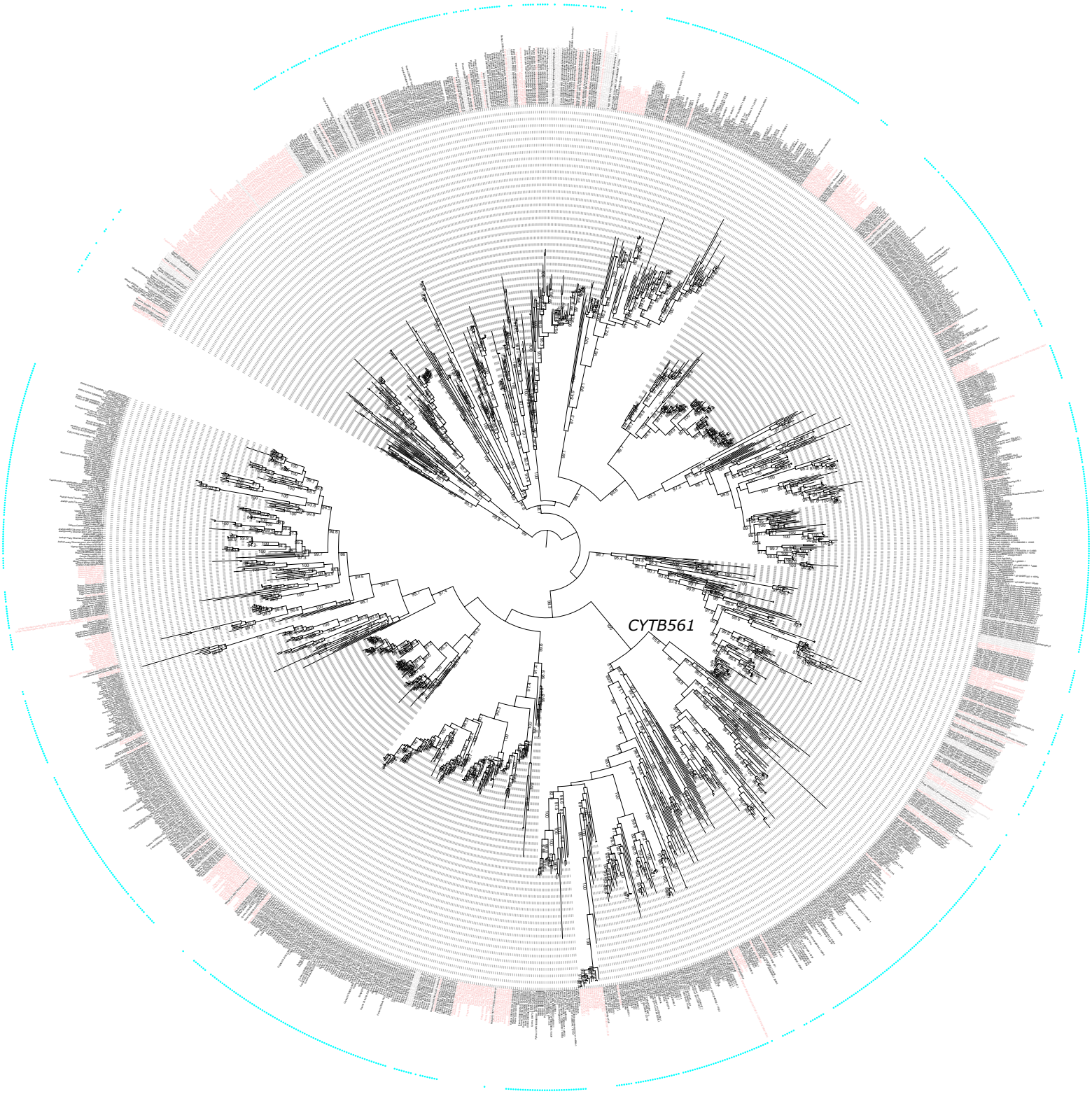
Supplementary Figure 27: Maximum likelihood tree of *SEC*.



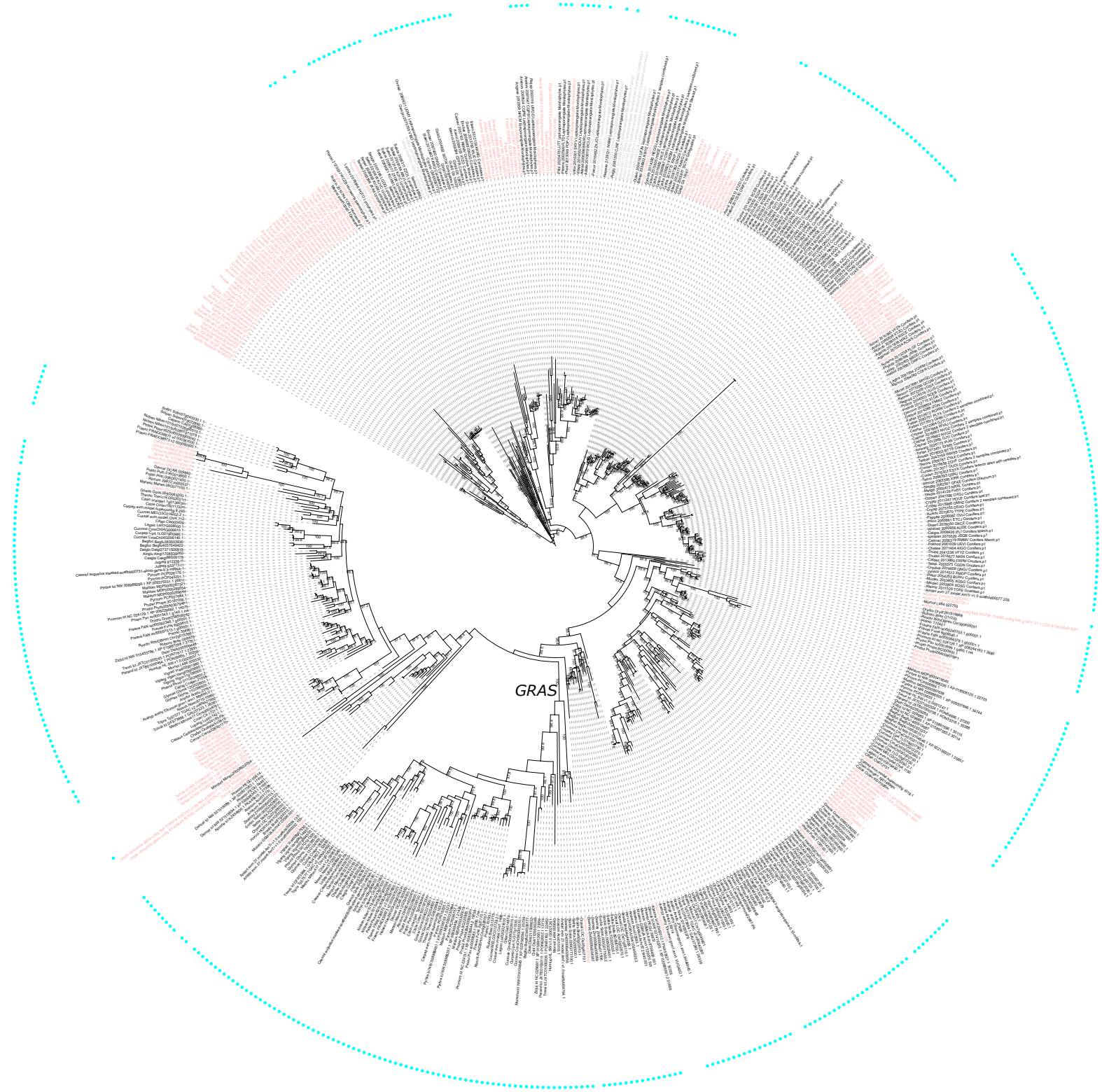
Supplementary Figure 28: Maximum likelihood tree of *SYN*.



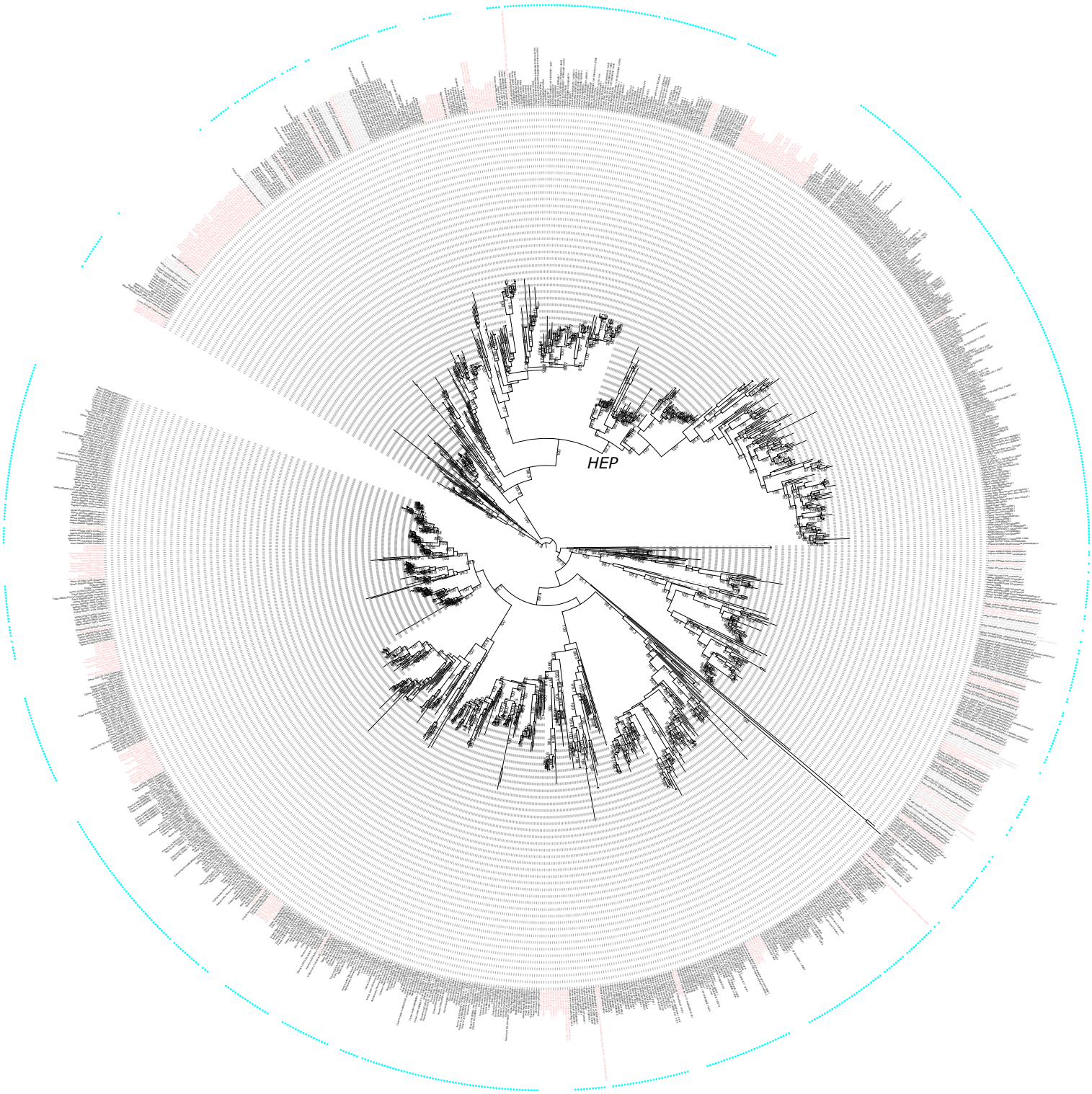
Supplementary Figure 29: Maximum likelihood tree of *CYTB561*.



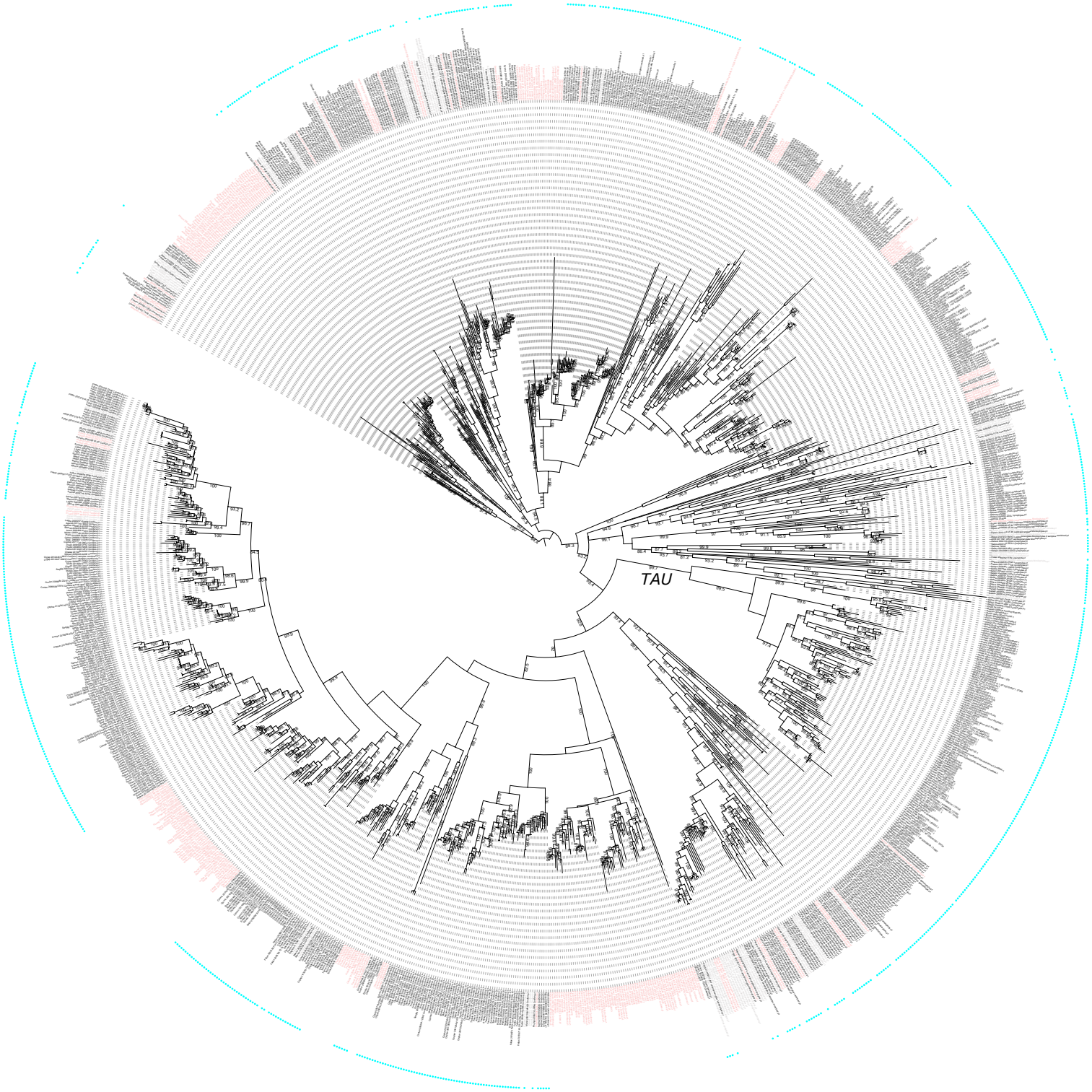
Supplementary Figure 30: Maximum likelihood tree of *GRAS*.



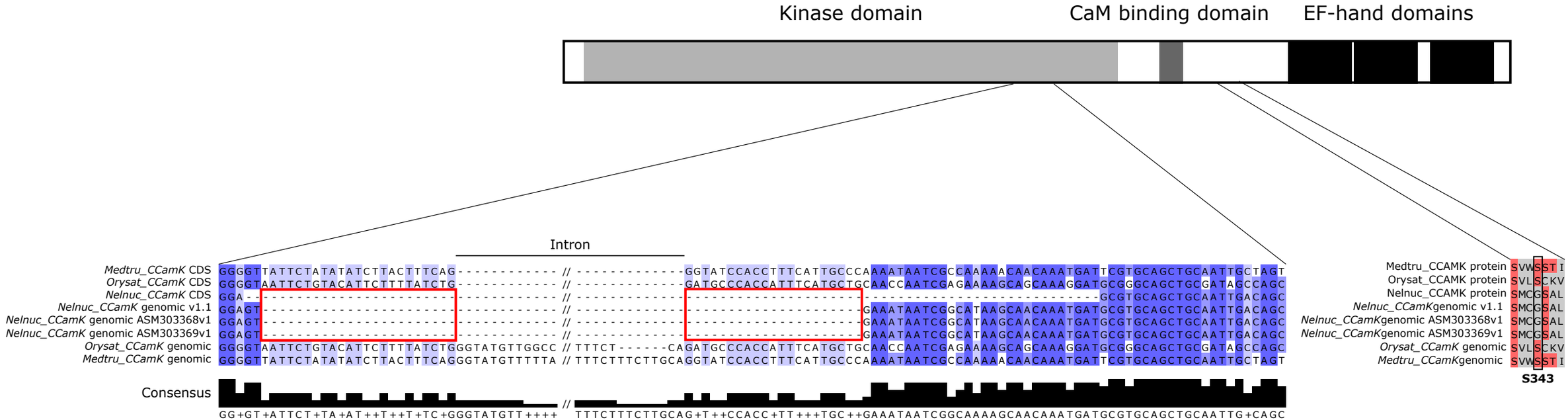
Supplementary Figure 31: Maximum likelihood tree of *HEP*.



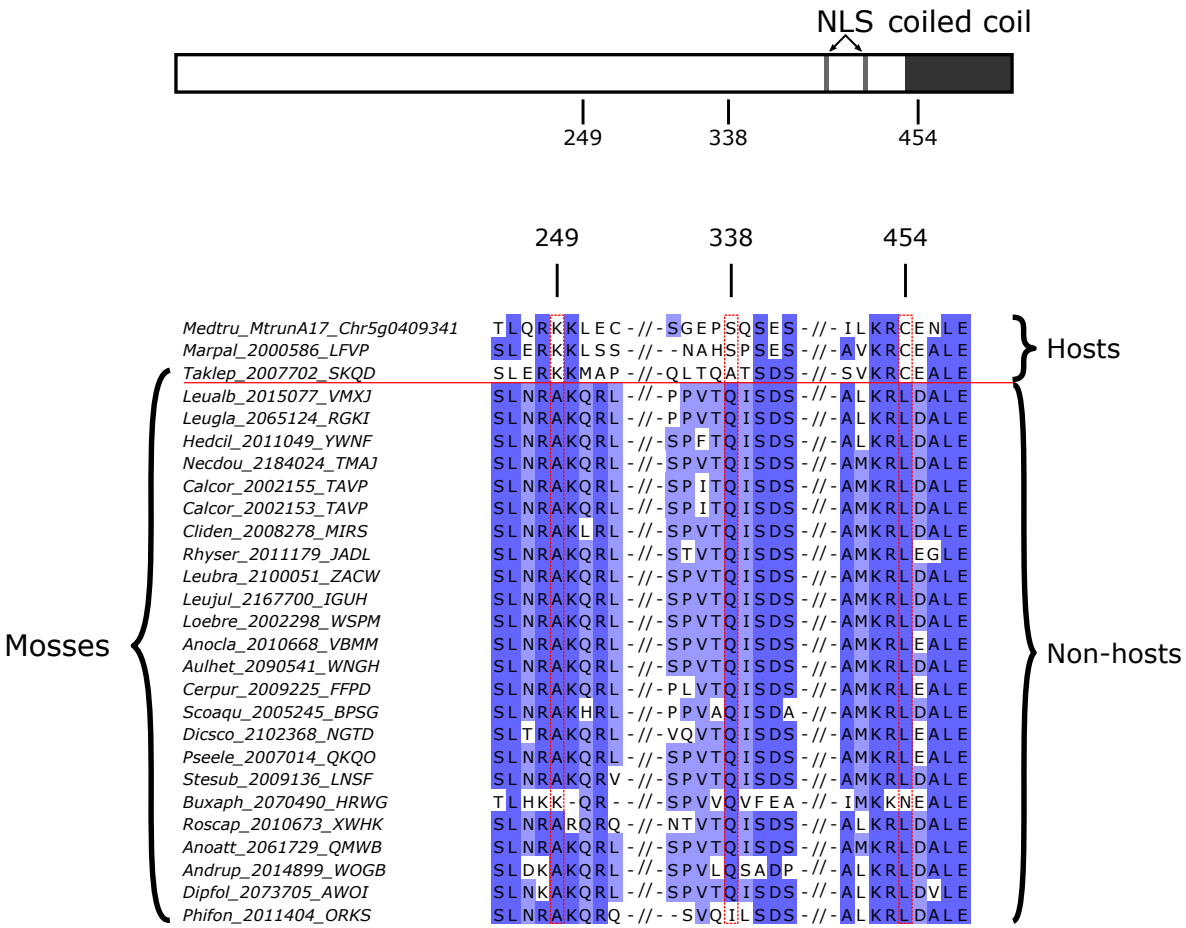
Supplementary Figure 32: Maximum likelihood tree of *TAU*.



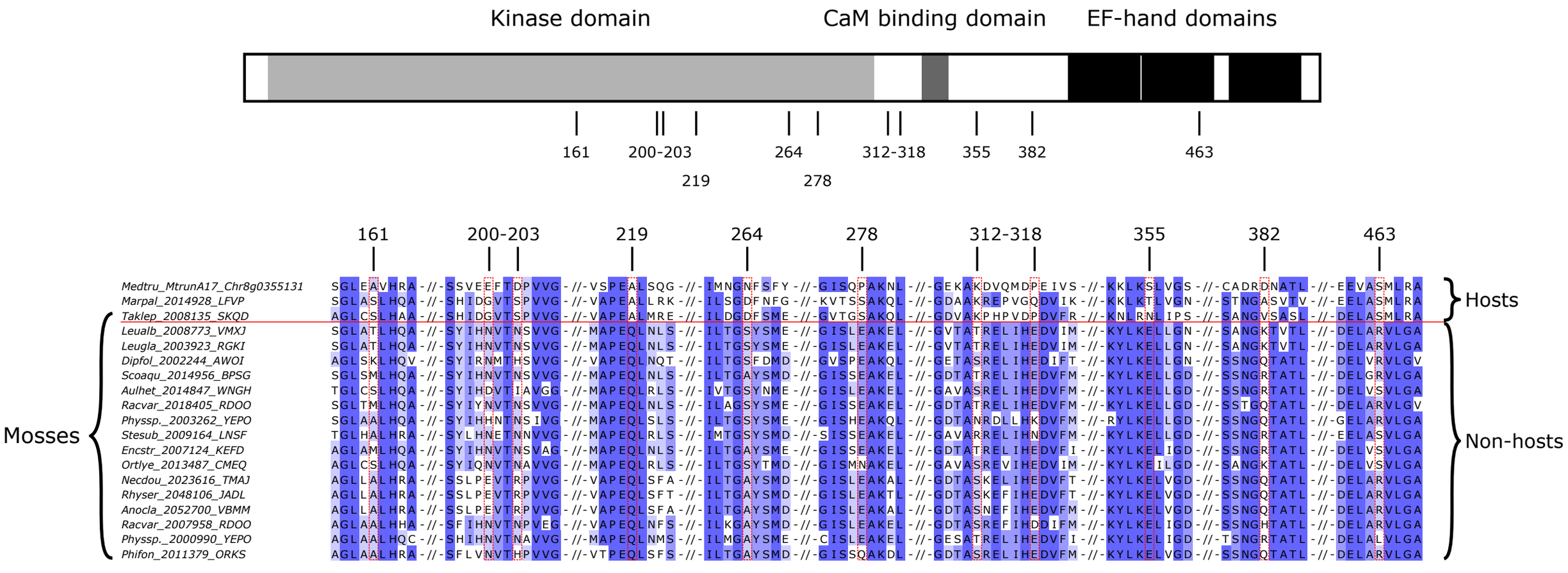
Supplementary Figure 33: Alignment of Nelumbo nucifera CCaMK from 3 assemblies with reference sequences showing an indel in the kinase domain and an amino acid switch on the essential for phosphorylation Serine 343



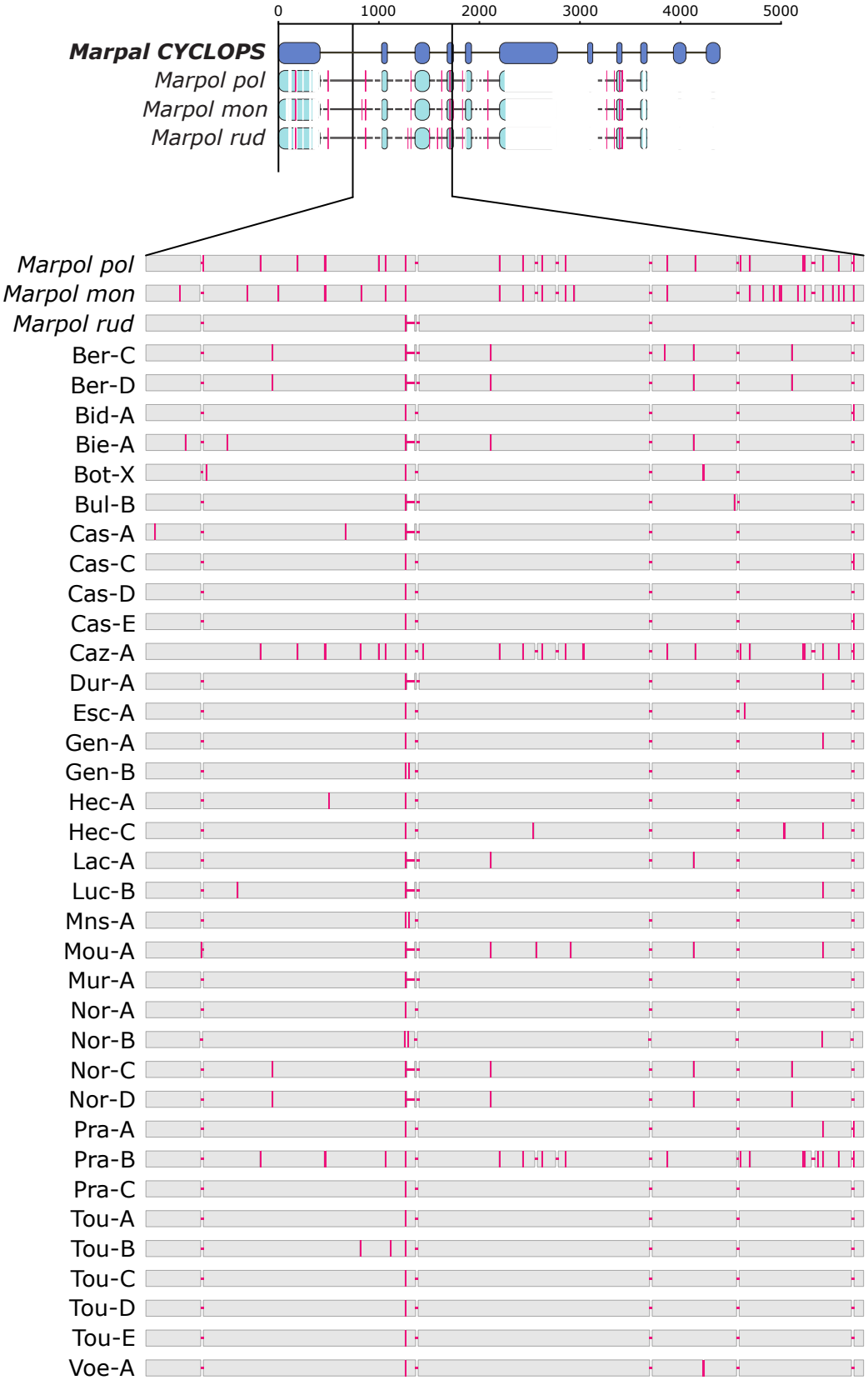
Supplementary Figure 34: Sites under positive selection in *CYCLOPS* in non-mycorrhizal mosses



Supplementary Figure 35: Sites under positive selection in *CCaMK* in non-mycorrhizal mosses



Supplementary Figure 36: Comparative alignment of 1kb from *CYCLOPS* among the three subspecies genomes and 35 accessions of *Marchantia polymorpha*.



Supplementary Figure 37: Presence/absence pattern of 11 genes in *Ericaceae* compared to species with different symbiotic abilities