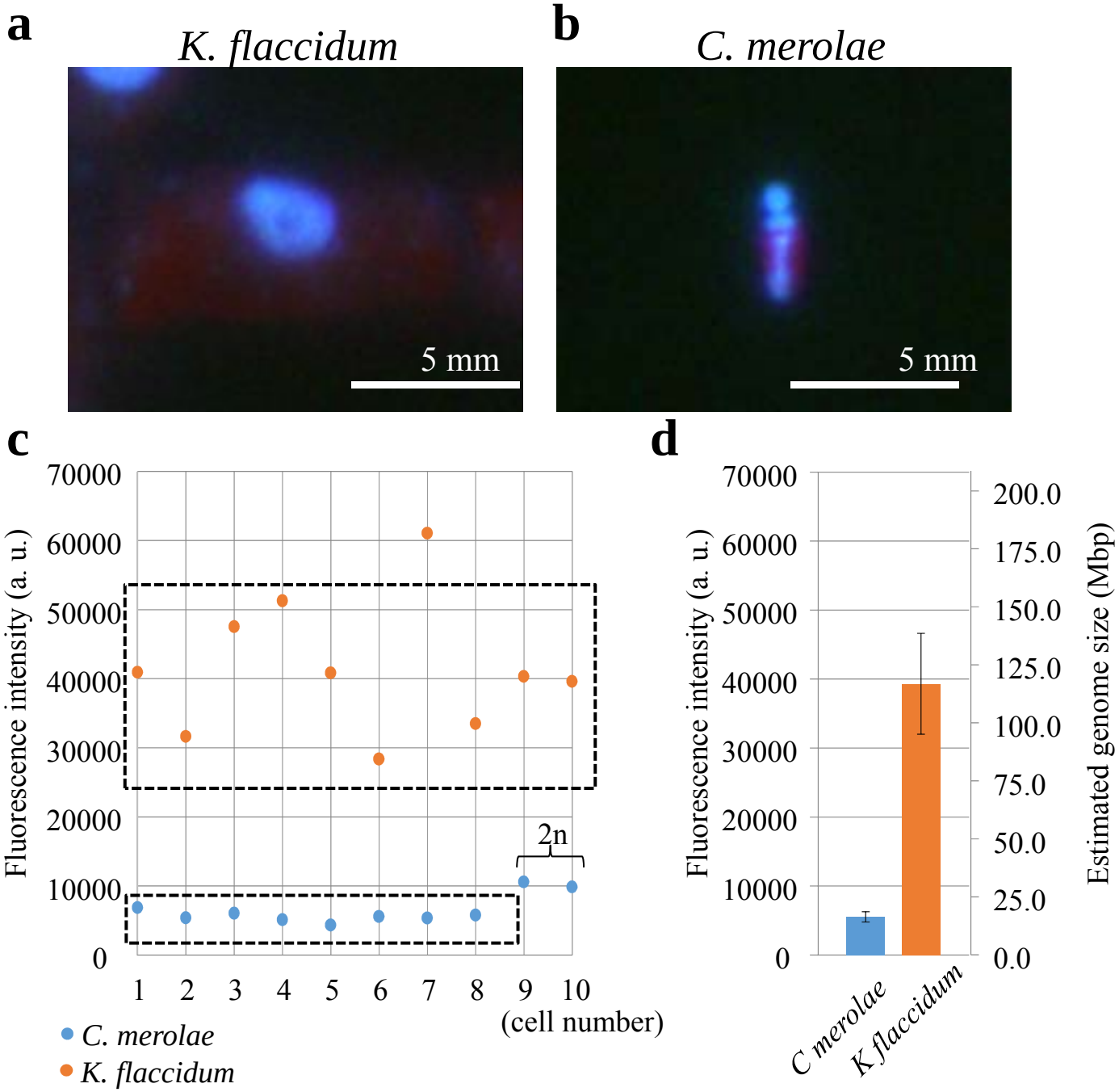
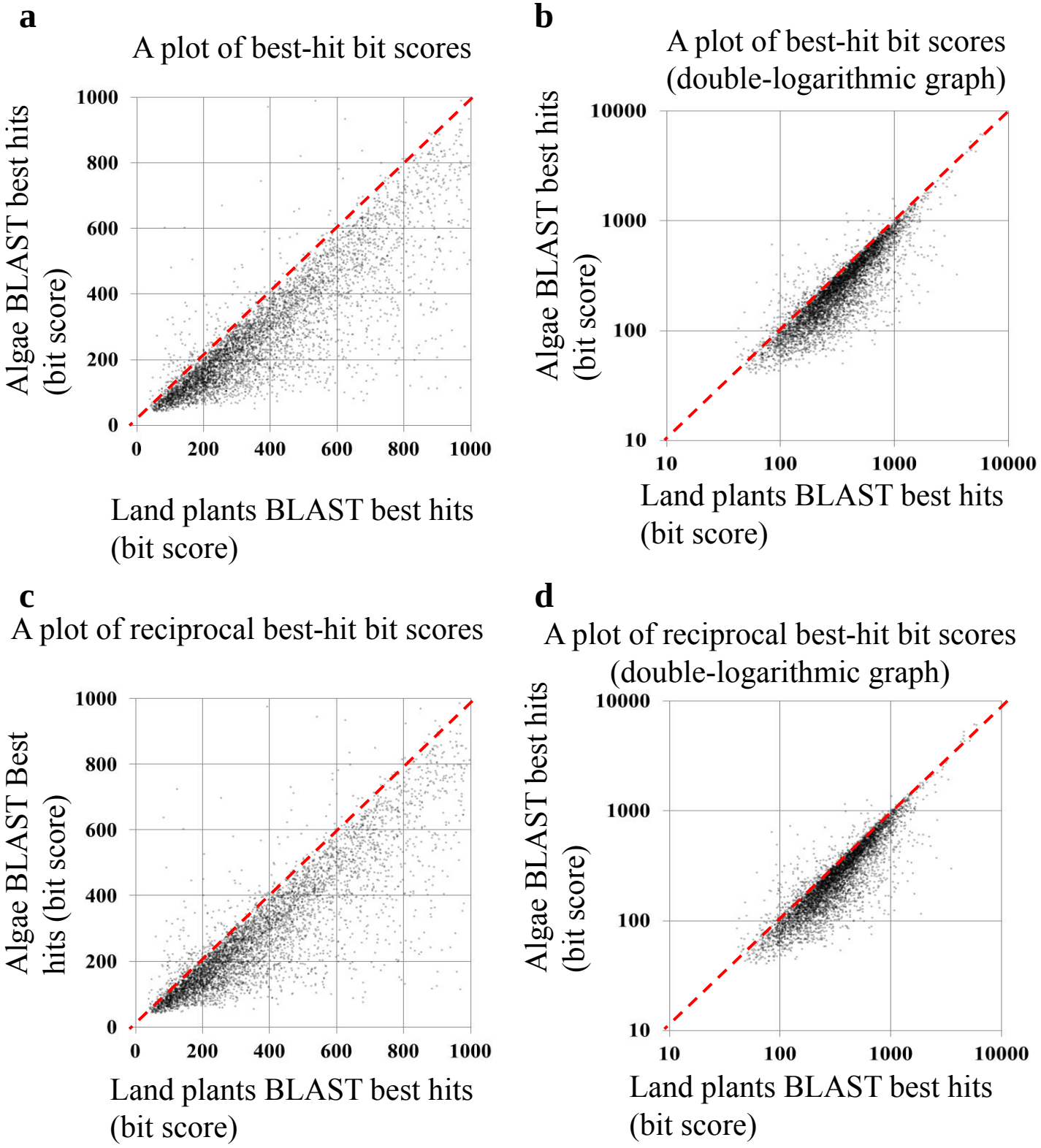


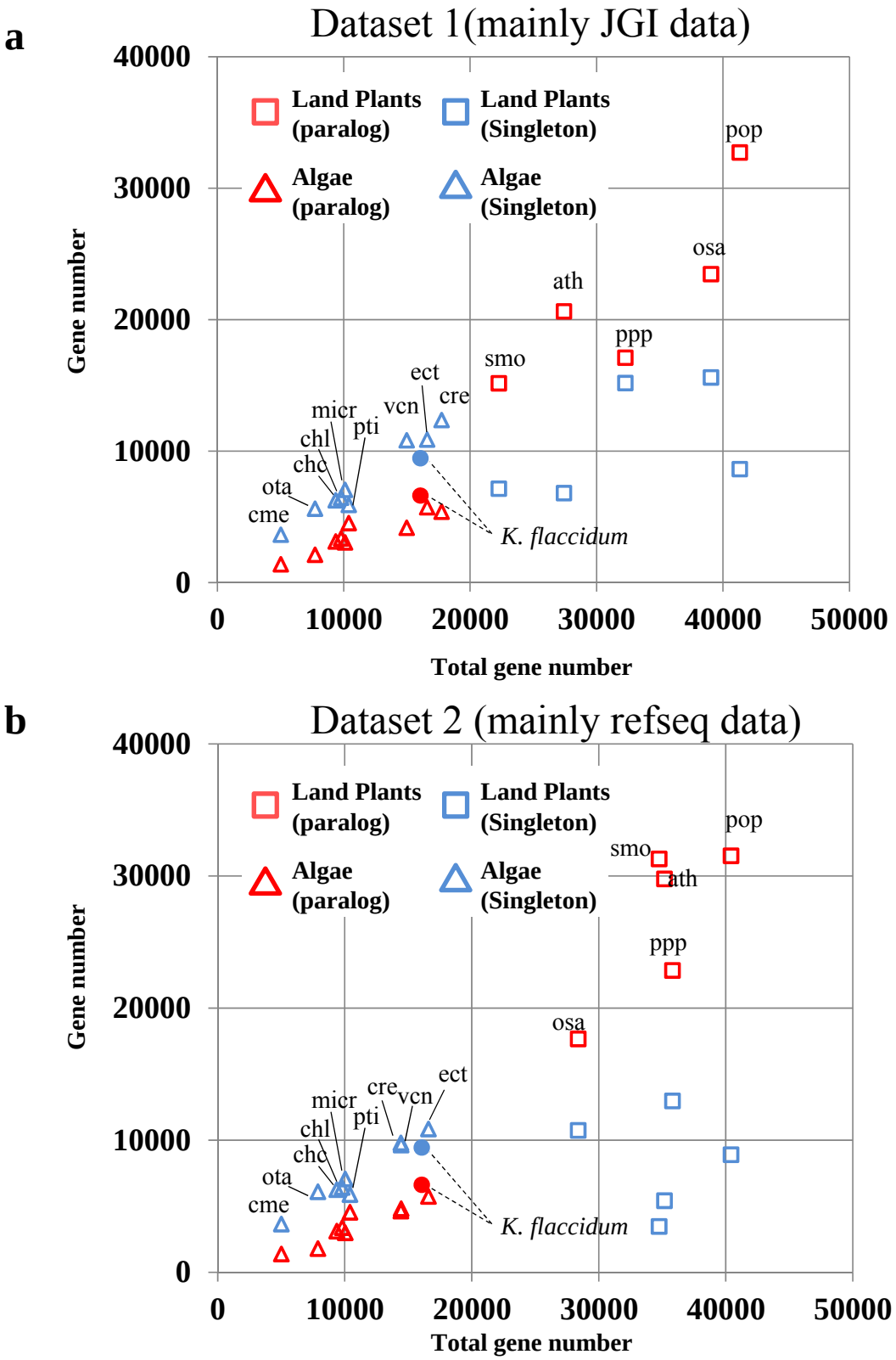
e

	Mean of fluorescence intensity (a. u.)	The estimated genome size
C. merolae	5541 ± 735	(16.5 Mbp)
K. flaccidum	39323 ± 7326	117.1 ± 21.8 Mbp

Supplementary Figure 1. DAPI-stained DNA fluorescence images of (**a**) *K.flaccidum*, and (**b**) *C. merolae*. Fluorescence of chlorophyll within chloroplast is red colour. (**c**) The fluorescence intensity of DAPI-stained nuclei in each cells. Dashed-line indicated data used for genome size estimation. (**d,e**) estimated genome size of *K. flaccidum*. (mean values ±s.d.) *C. merolae* was used as the standard (16.5 Mbp) Error bars depict s.d..



Supplementary Figure 2. Two-dimensional representation of BLASTP and reciprocal BLASTP bit scores. BLASTP bit scores (**a, b**) and reciprocal BLASTP bit scores (**c, d**) of the reciprocal best-hit proteins (5,495 pairs, Supplementary Data 8) of five land plants and nine algae are plotted on the *x* and *y* axes, respectively.



Supplementary Figure 3. Relationship between the numbers of paralogous genes and singletons and the total number of genes in the genomes of 15 species (**a**): dataset1 (**b**) :dataset2 (Supplementary table 3 and Supplementary table 7). The red and blue plot correspond to the data for paralogs and singletons, respectively. The filled circle data points are for *K. flaccidum*, and the triangles and squares data points are for land plants and algae, respectively. Abbreviations for spieces are as follows; ppp, *Physcomitrella patens* subsp. patens; smo, *Selaginella moellendorffii*; osa, *Oryza sativa* subsp. japonica; pop, *Populus trichocarpa*; ath, *Arabidopsis thaliana*; pti, *Phaeodactylum tricornutum*; cme, *Cyanidioschyzon merolae*; micr, *Micromonas* strain RCC299; ota, *Ostreococcus tauri*; chl, *Chlorella variabilis* NC64A; vcn, *Volvox carteri* f. nagariensis; cre, *Chlamydomonas reinhardtii*; ect, *Ectocarpus siliculosus*; chc, *Chondrus crispus*.

280,112 genes — Number of total genes in 15 species (Dataset 1, mainly JGI data : Supplementary Data 7)

↓ OrthoMCL

21,335 families — Number of classified gene families

↓

3,420 families — Numbers of increased gene families in land plants (U-test P<0.05)

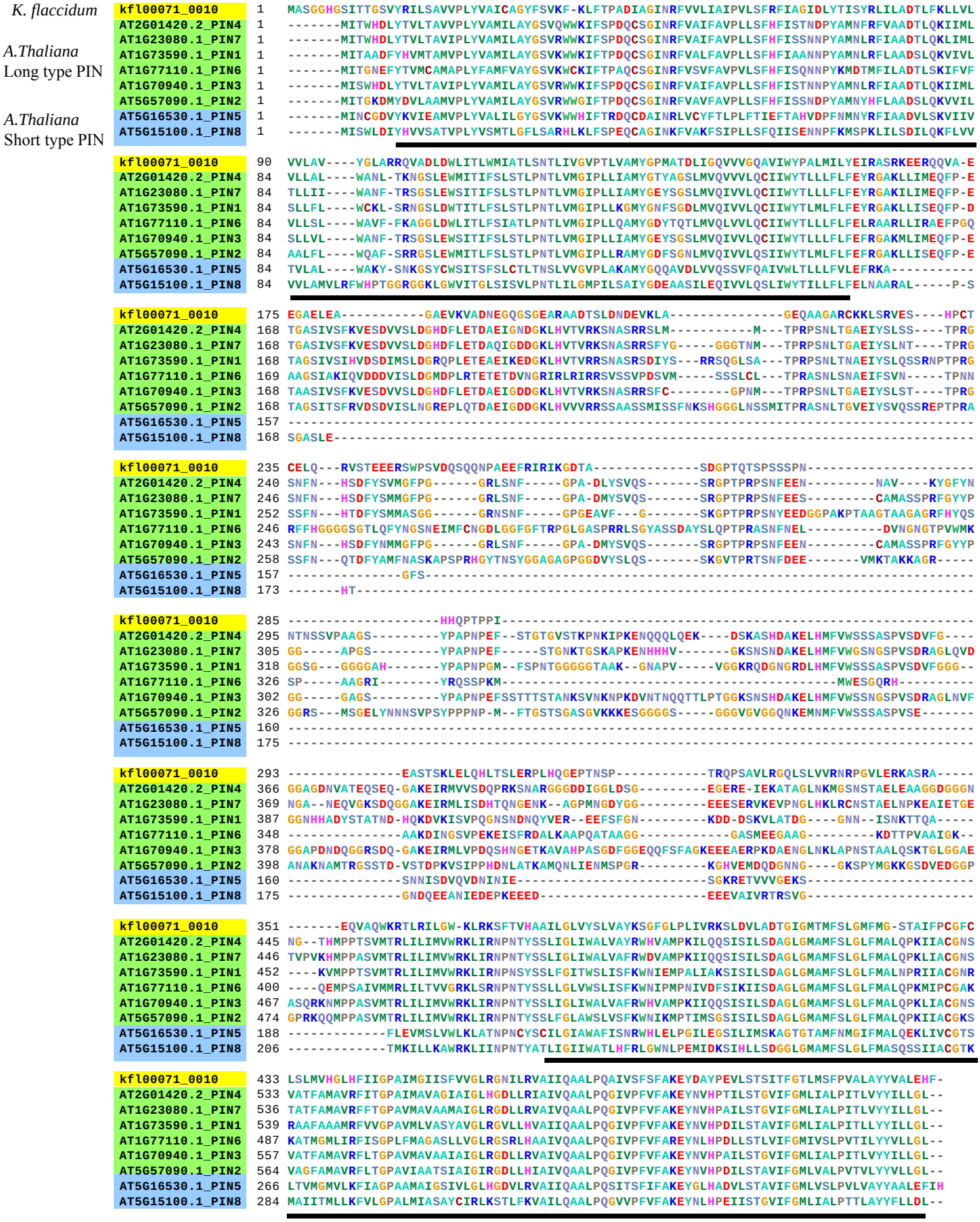
The counterparts of increased paralogs in land plants (in 3,420 families)			
#Taxon	number of gene families	number of singlton	number of the gene families of paralogues
Phaeodactylum tricornutum	805	524	281
Cyanidioschyzon merolae	745	556	189
Chondrus crispus	706	500	206
Ectocarpus siliculosus	853	542	311
Micromonas strain RCC299	1,103	819	284
Ostreococcus tauri	947	712	235
Chlorella variabilis NC64A	1,130	800	330
Volvox carteri f. nagariensis	935	666	269
Chlamydomonas reinhardtii	978	690	288
Klebsormidium flaccidum	2,296	1,700	596

↓

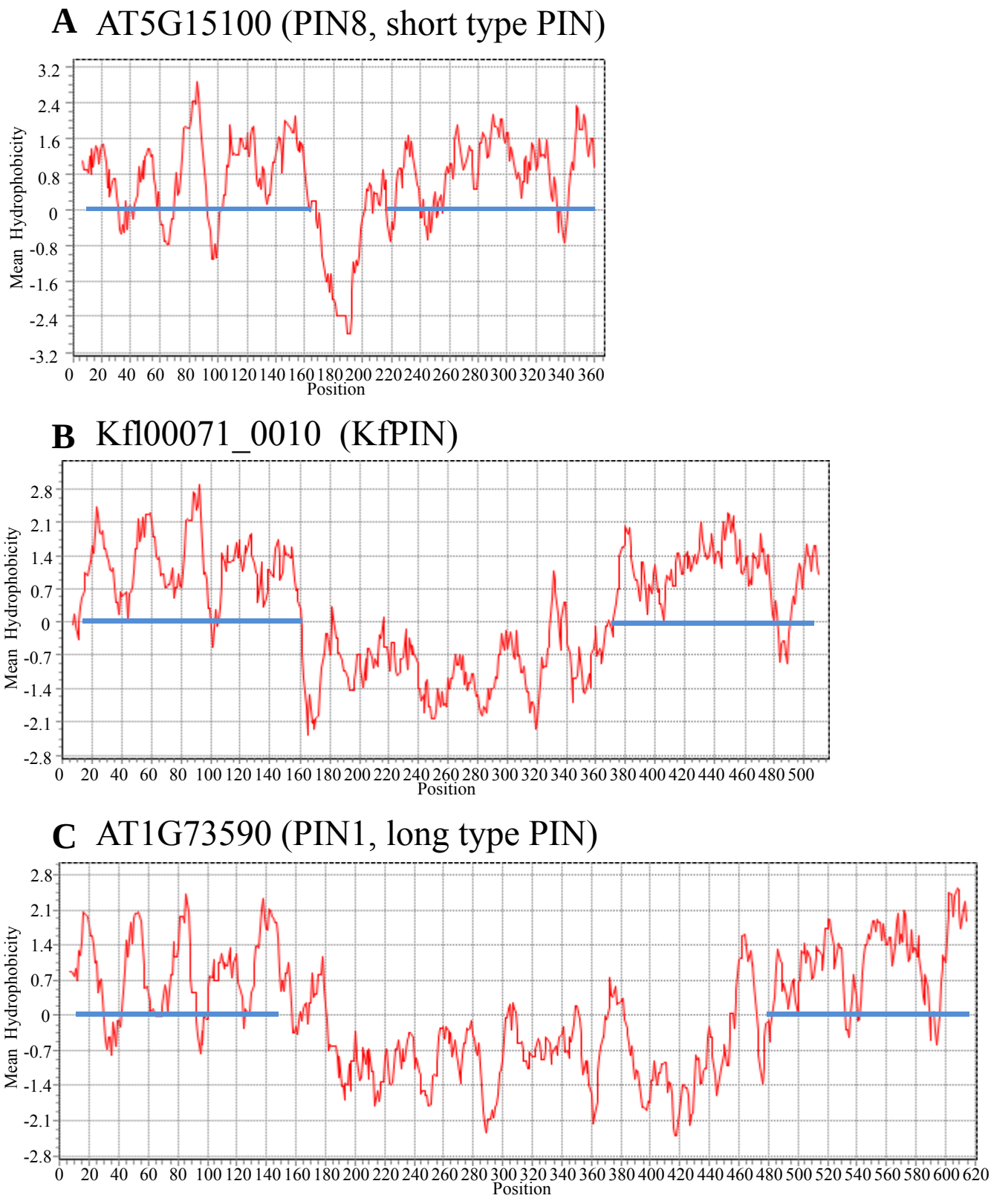
197 families — Numbers of significantly increased gene families in land plants (median land plant gene number / median algae gene number ≥ 10).

The counterparts of significantly increased paralogs in land plants (in 197 families) (median land plant gene number / median algae gene number ≥ 10).			
#Taxon	number of gene families	number of singlton	number of the gene families of paralogues
Phaeodactylum tricornutum	38	19	19
Cyanidioschyzon merolae	24	20	4
Chondrus crispus	34	20	14
Ectocarpus siliculosus	44	19	25
Micromonas strain RCC299	43	24	19
Ostreococcus tauri	39	26	13
Chlorella variabilis NC64A	64	31	33
Volvox carteri f. nagariensis	46	25	21
Chlamydomonas reinhardtii	48	27	21
Klebsormidium flaccidum	131	51	80

Supplementary Figure 4. Numbers of significantly increased gene families in land plants. Increased gene families in land plants were selected by the Mann-Whitney U-test (P<0.05). Significantly increased gene families in land plants were defined as gene families which include median land plant gene number are more than ten times median algae gene number.

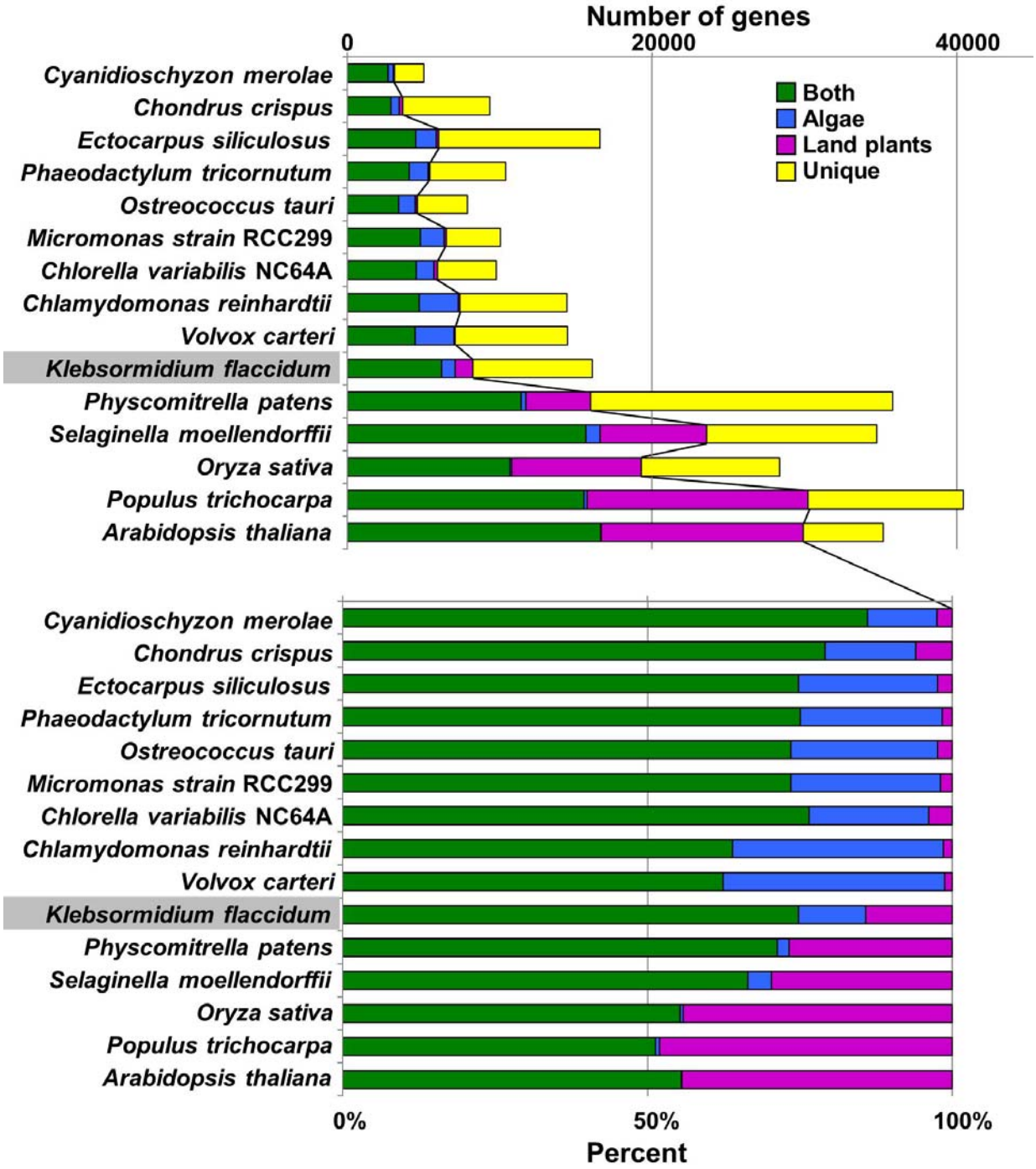


Supplementary Figure 5. Multiple alignment of amino acid sequences of KfPIN (name background-color: yellow), long type AtPINs (green) and short type AtPINs (blue). Black bars under alignments indicate membrane spanning region predicted with Tmpred[59].

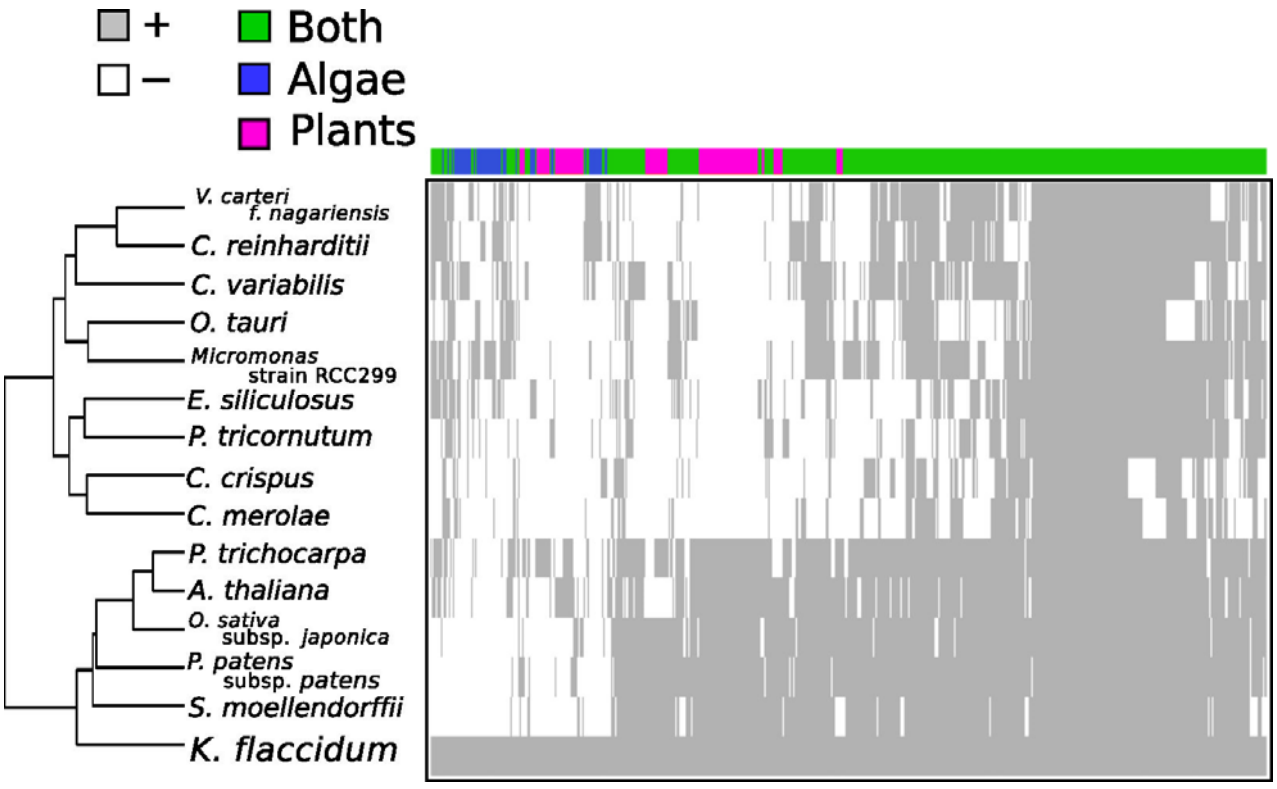


Supplementary Figure 6. Kyte & Doolittle Scale Mean Hydrophobicity Profile (Scan-window size = 13) of a short type AtPIN (PIN8), KfPIN, and a long type AtPIN (PIN1). Blue bars at the center of profile indicate two membrane spanning regions predicted with TMpred.

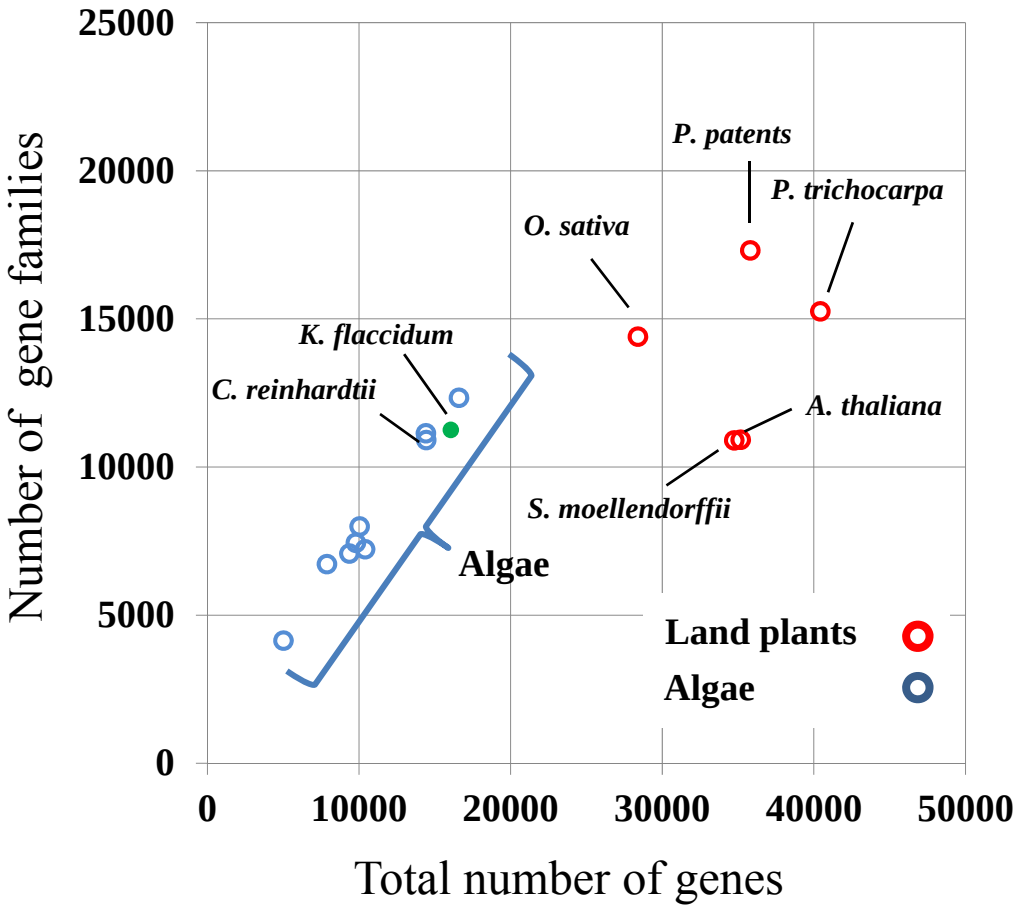
Supplementary Figure 7. Comparison of genes among 15 species of unicellular and multicellular organisms. Numbers of proteins found in both unicellular and multicellular organisms (green), proteins shared by unicellular (blue), proteins shared by multicellular (magenta), and no reciprocal best hit to other species (yellow) with classification by use of OrthoMCL (blue or pink of back ground colors of spieces name indicate unicellular and multicellular organisms respectively, other approach was identical to Fig. 3a).



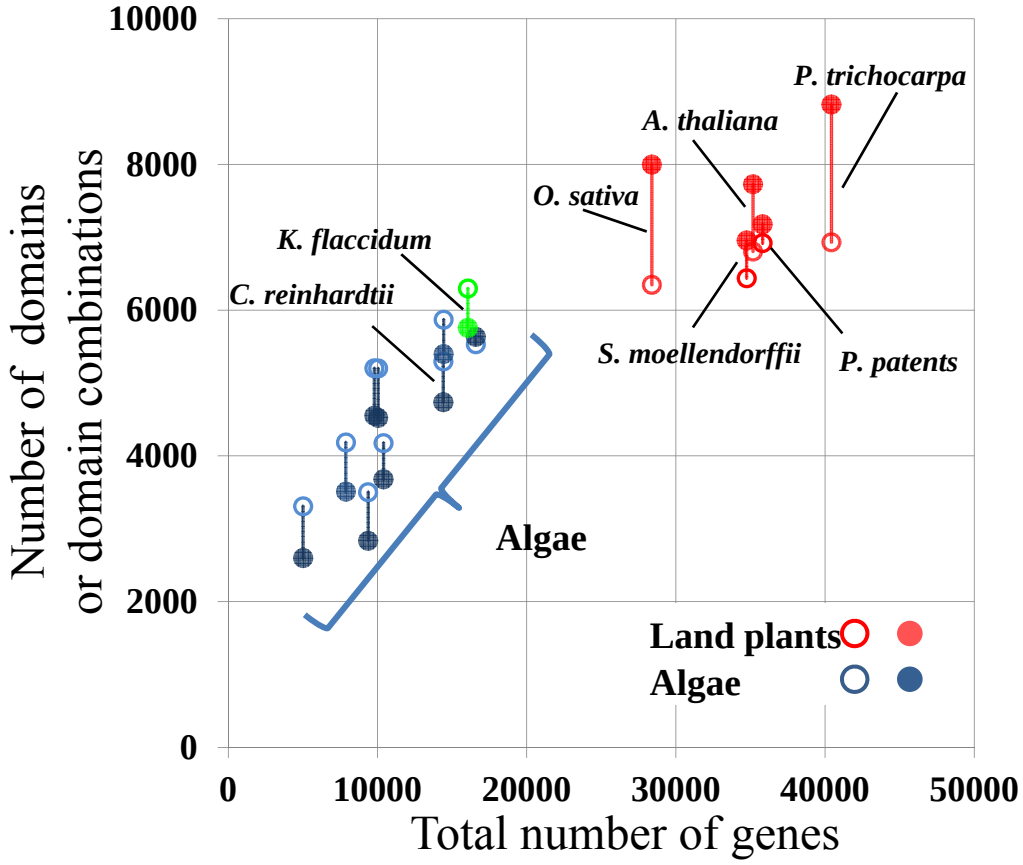
Supplementary Figure 8. Comparison of genes among 15 species of algae and land plants. Numbers of proteins found in both algae and land plants (green), proteins shared by algae (blue), proteins shared by land plants (magenta), and no reciprocal best hit to other species (yellow) with classification by use of OrthoMCL. Dataset2 (mainly refseq data, Supplementary table 2 and Supplementary table 7) was used, and other approach was identical to Fig. 3a.



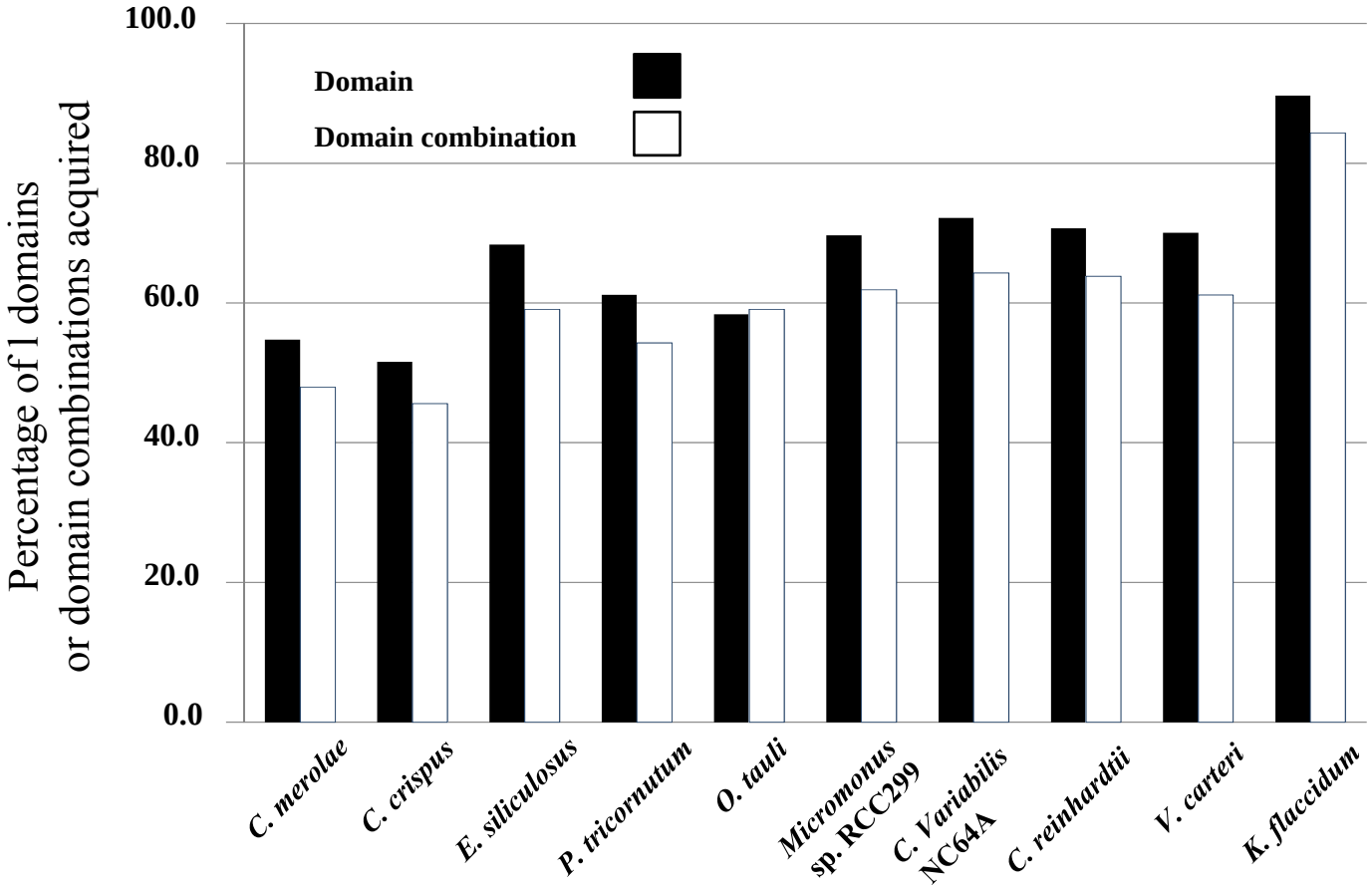
Supplementary Figure 9. Binary heat map of 5,395 groups were extracted as non-unique groups sharing *K. flaccidum* and other 14 organism studied. Grey indicates that the group was similar to at least one gene in each reference organism following OrthoMCL analysis; white indicates no orthologus gene. Dataset2 (mainly refseq data, Supplementary table 7) was used, and other approach was identical to Fig. 3b.



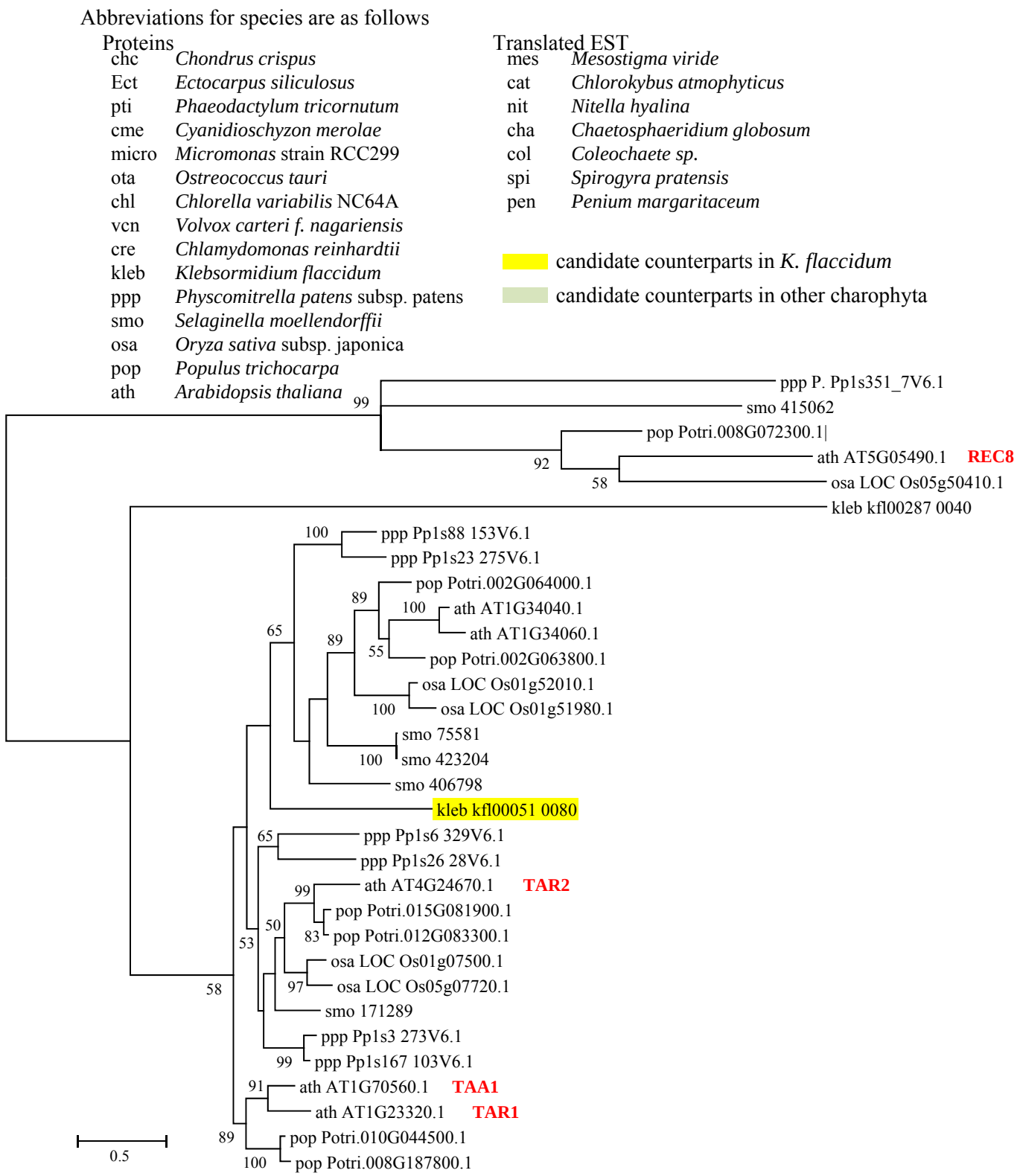
Supplementary Figure 10. Gene families in 15 species of algae and land plants. The green circle denotes the data point for *K. flaccidum*, and red and blue circles denote data points for land plants and algae, respectively. Dataset2 (mainly refseq data, Supplementary table 3 and Supplementary table 7) was used, other approach was identical to Fig. 4a.



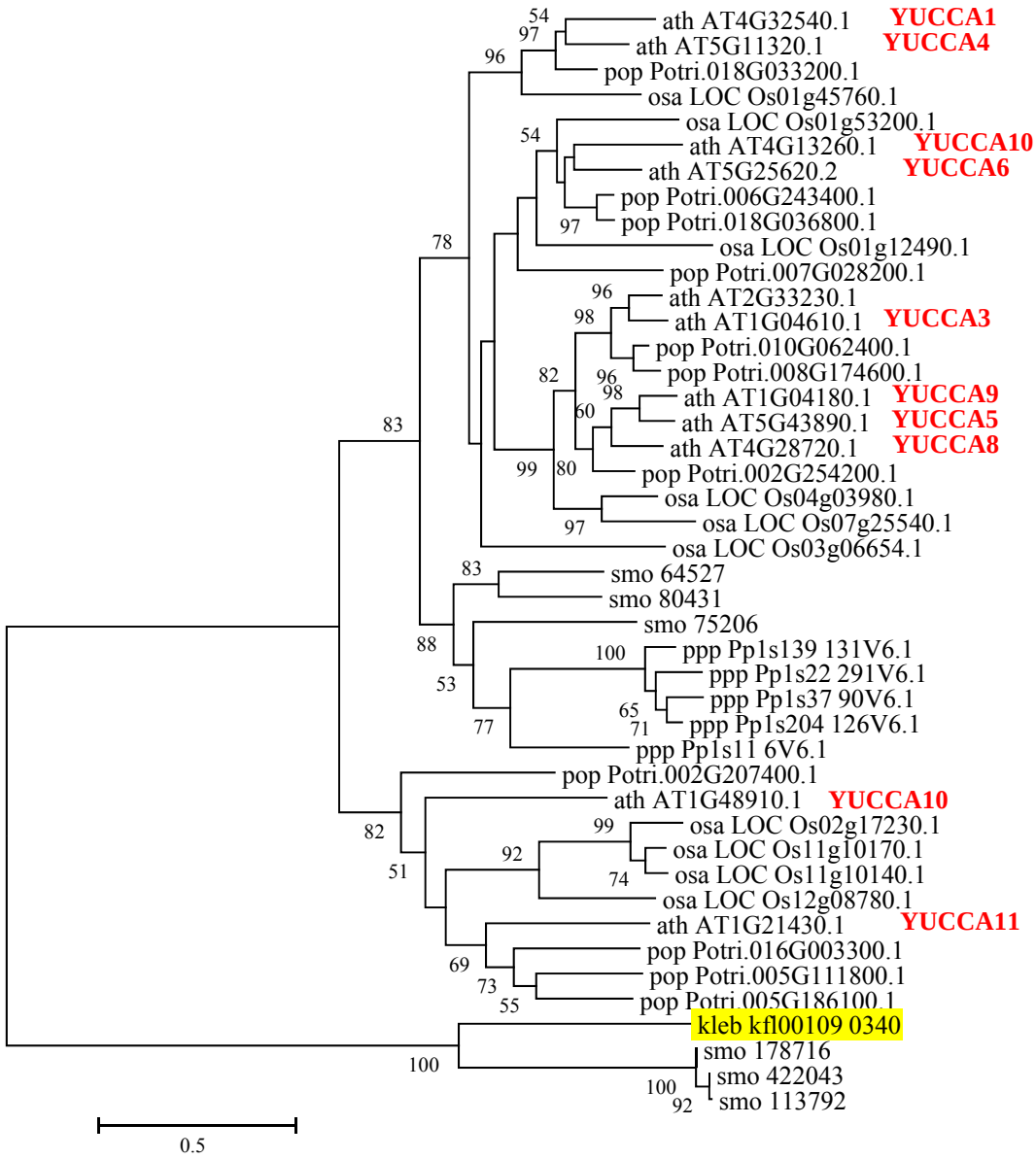
Supplementary Figure 11. Gene domains in 15 species of algae and land plants. Number of domains (open circles) and domain combinations (filled circles) expressed in terms of the total number of genes in each of 15 species. Dataset2 (mainly refseq data, Supplementary table 4 and Supplementary table 7) was used, and other approach was identical to Fig. 4b.



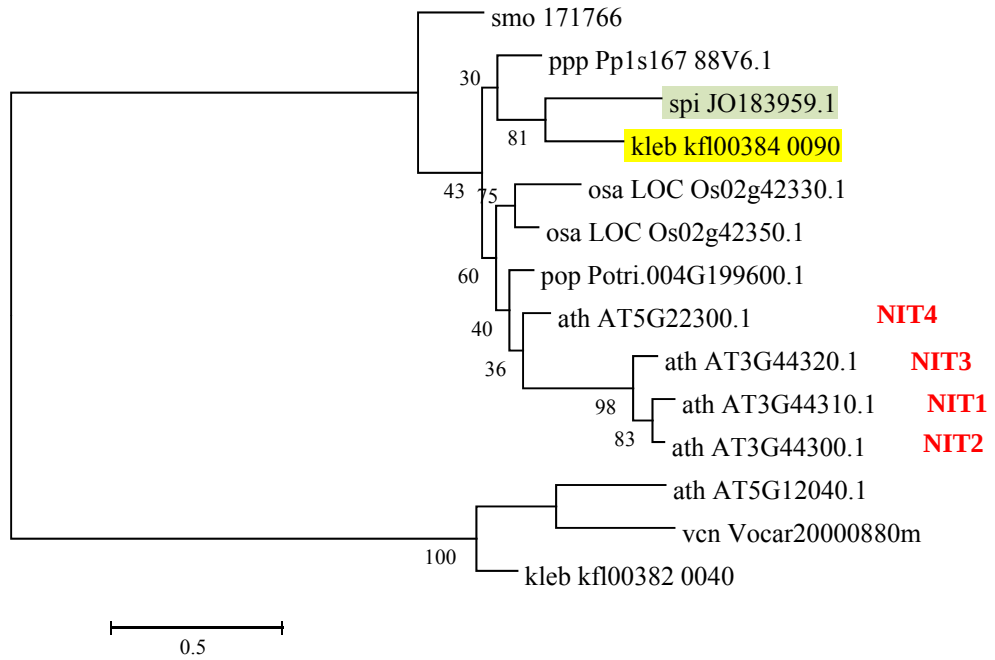
Supplementary Figure 12. In ten species of algae, protein domains which were commonly found in five land plants. Acquisition in algal genomes of conserved domains (black bars) and domain combinations (white bars) commonly found in land plants. For the land plants analyzed (five species), the numbers of conserved domains and domain combinations were 4,676 and 2,708, respectively. Dataset2 (mainly refseq data, Supplementary table 5 and Supplementary table 7) was used, and other approach was identical to Fig. 4c.



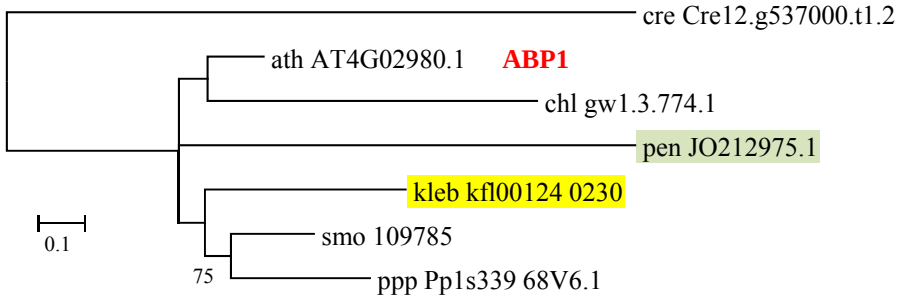
Supplementary Figure 13. Phylogenetic analysis of TAA, TAR and similar proteins of 15 spices and translated sequences of 7 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



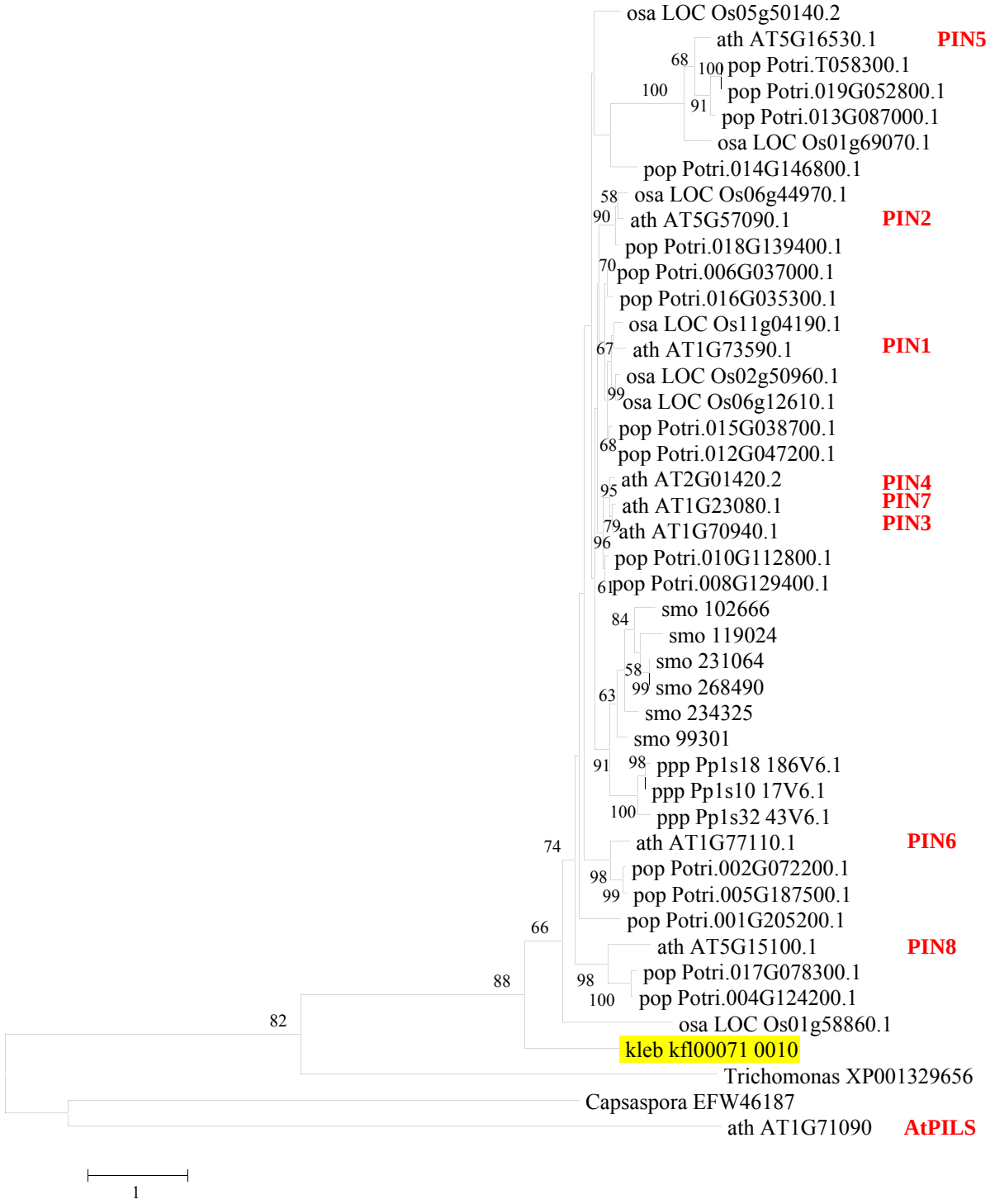
Supplementary Figure 14. Phylogenetic analysis of YUCCA and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



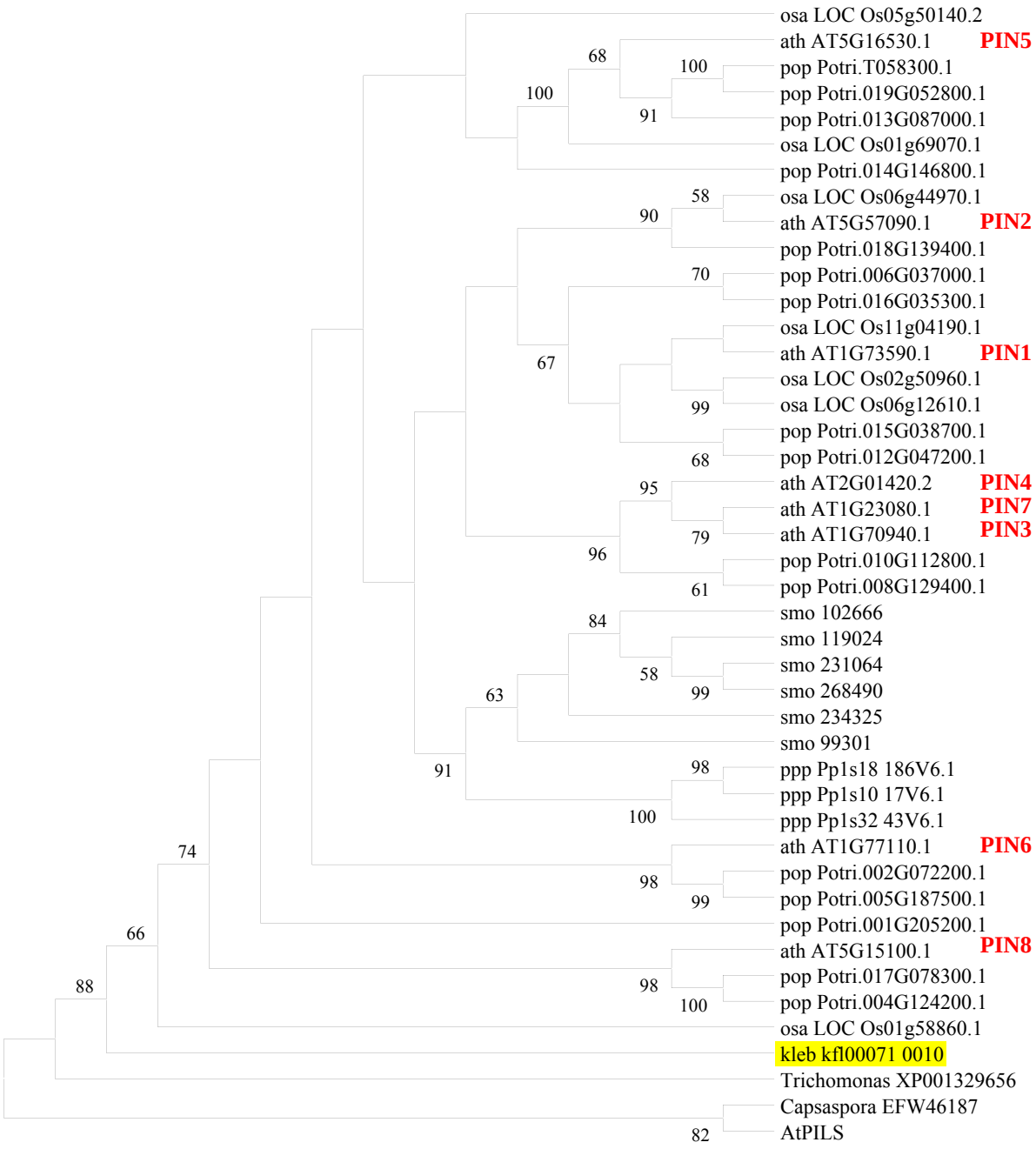
Supplementary Figure 15. Phylogenetic analysis of NIT and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



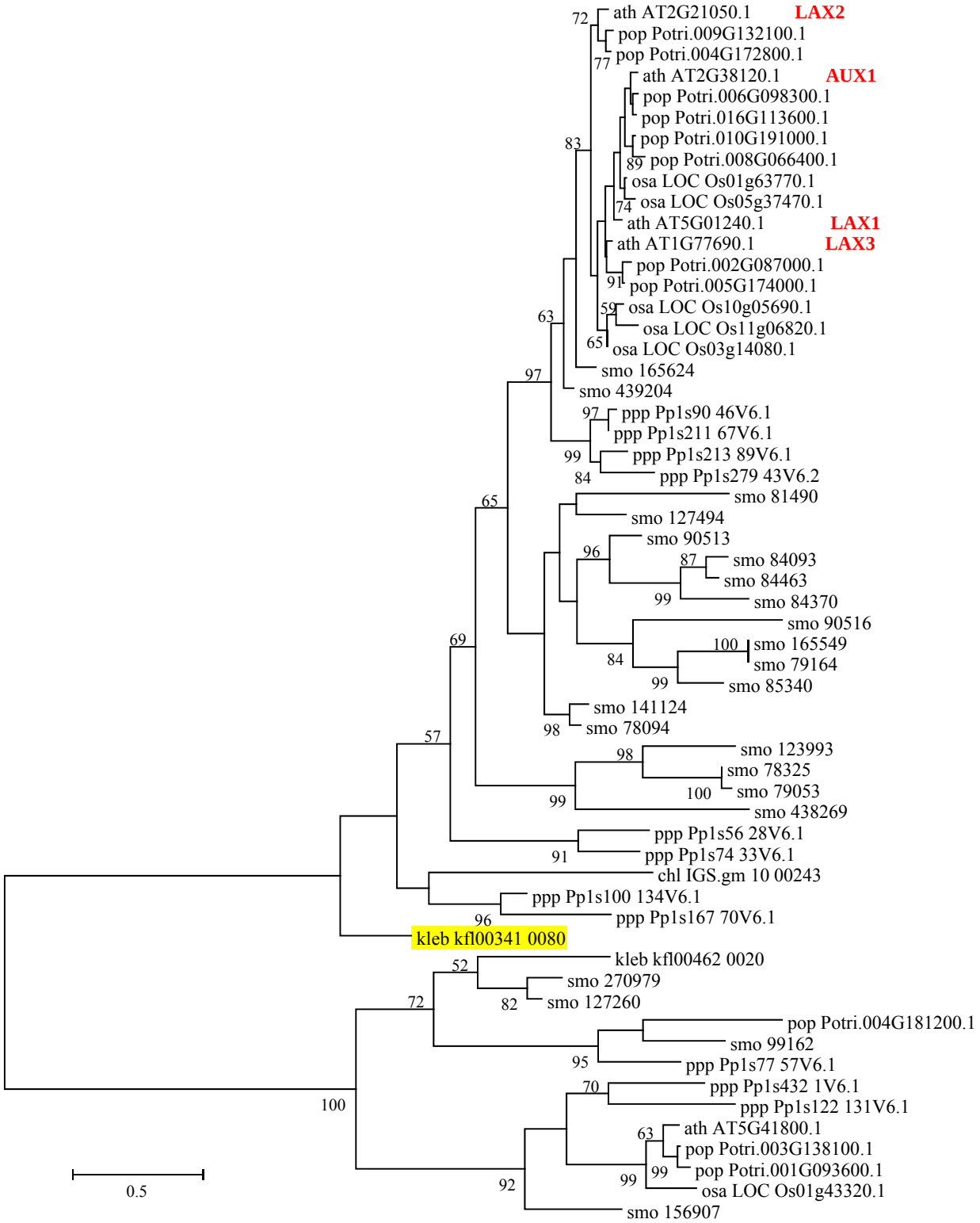
Supplementary Figure 16. Phylogenetic analysis of ABP1 and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "WAGF+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



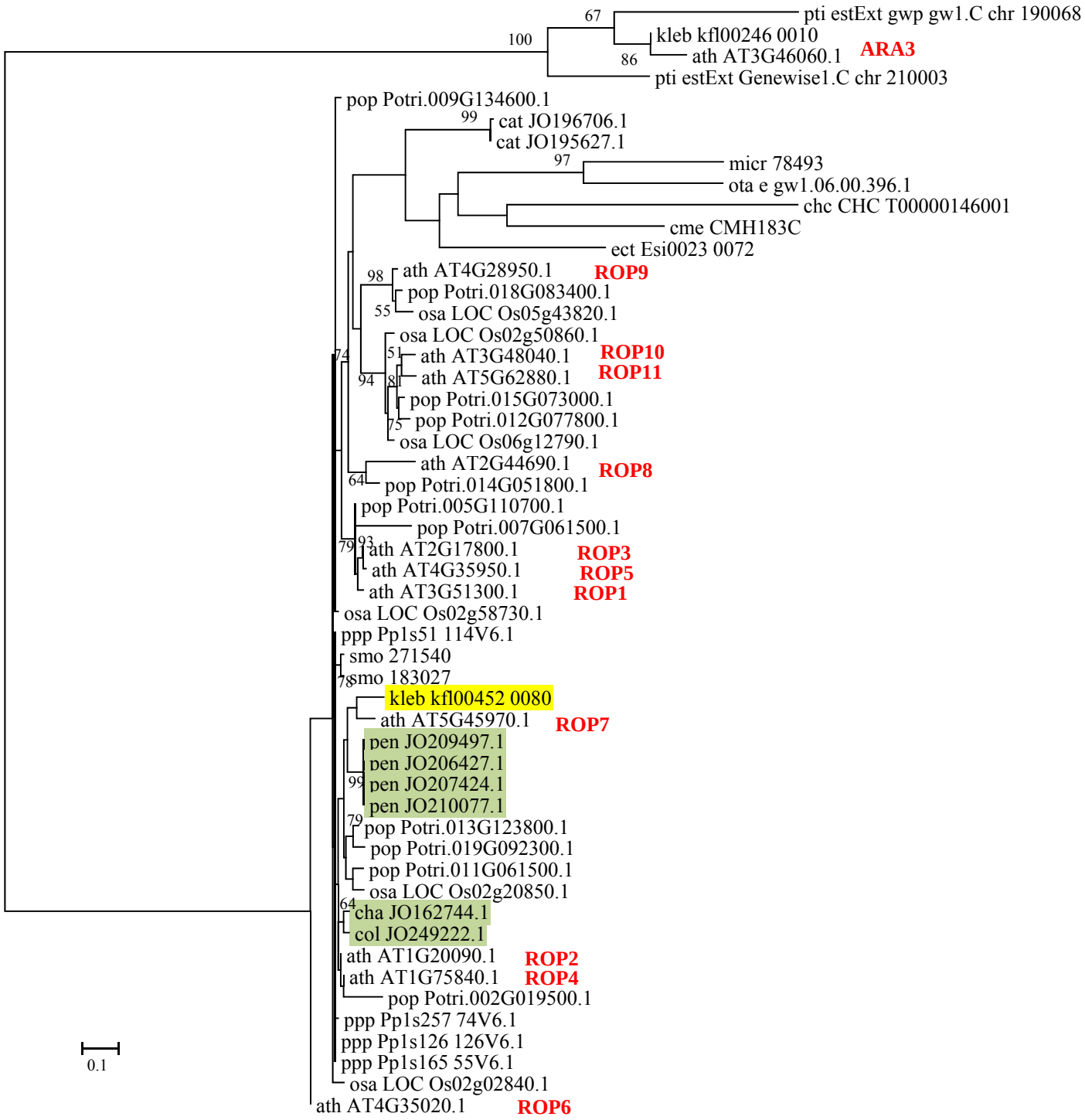
Supplementary Figure 17. Phylogenetic analysis of PIN and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LGF+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



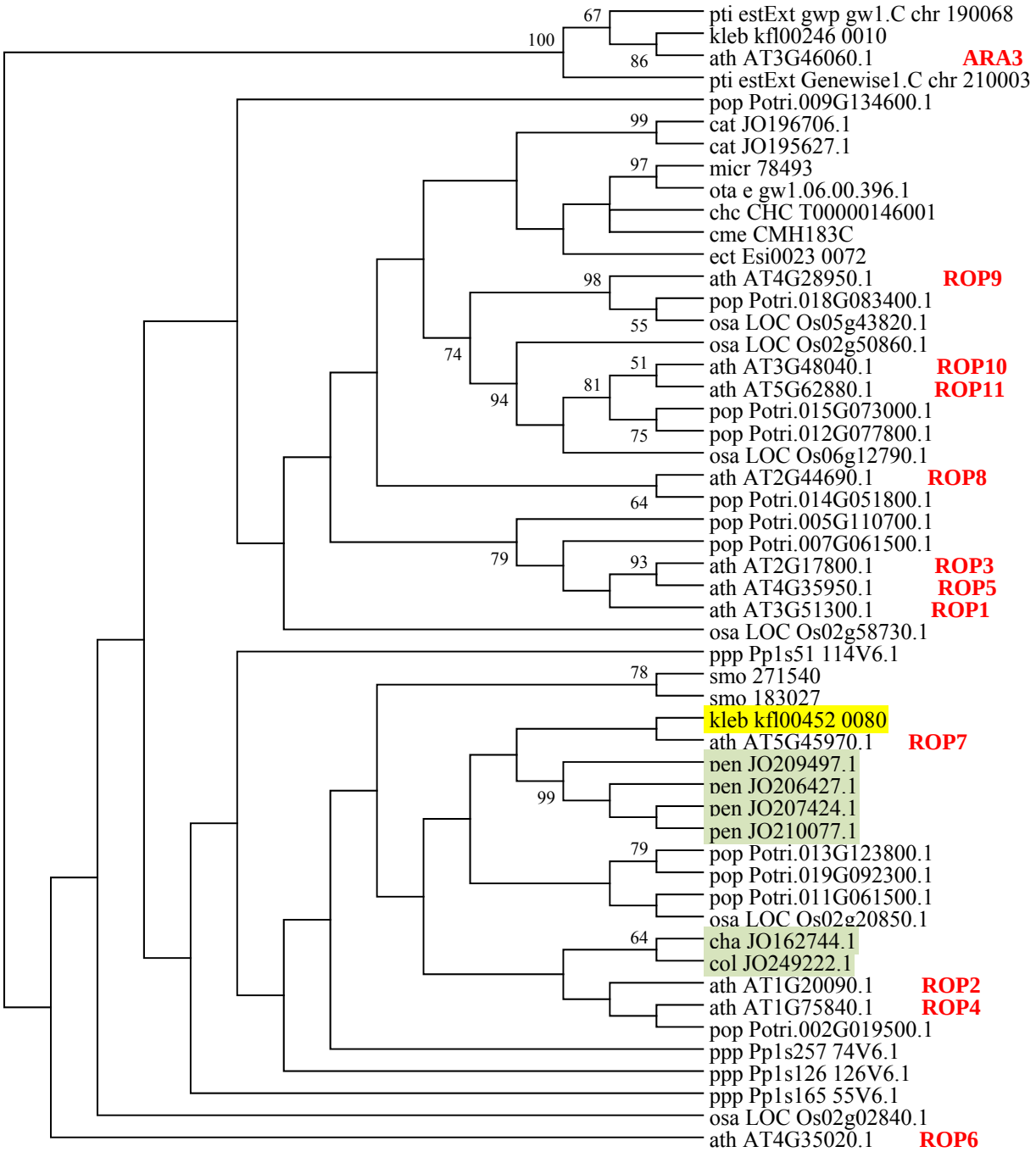
Supplementary Figure 18. Topology only style of Supplementary Figure 17.



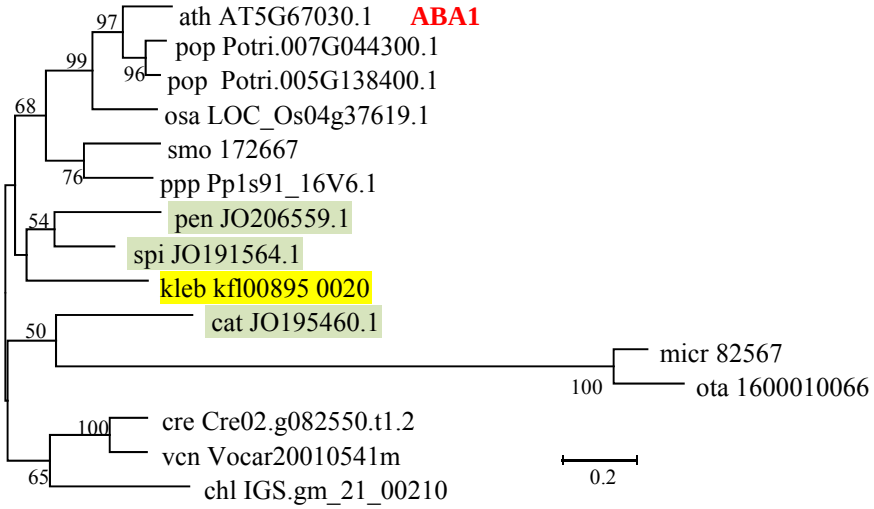
Supplementary Figure 19. Phylogenetic analysis of AUX/LUX and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LGF+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



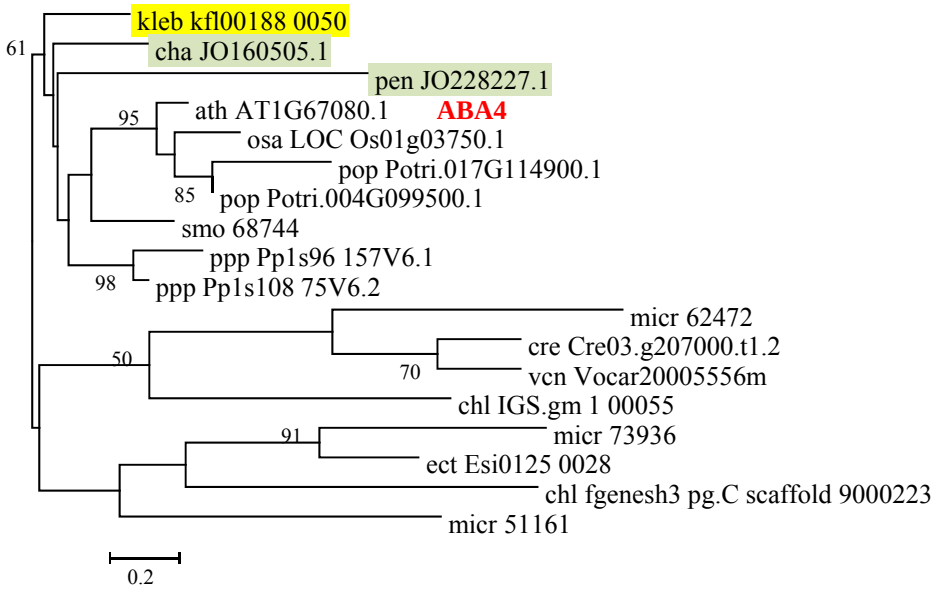
Supplementary Figure 20. Phylogenetic analysis of ROP and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



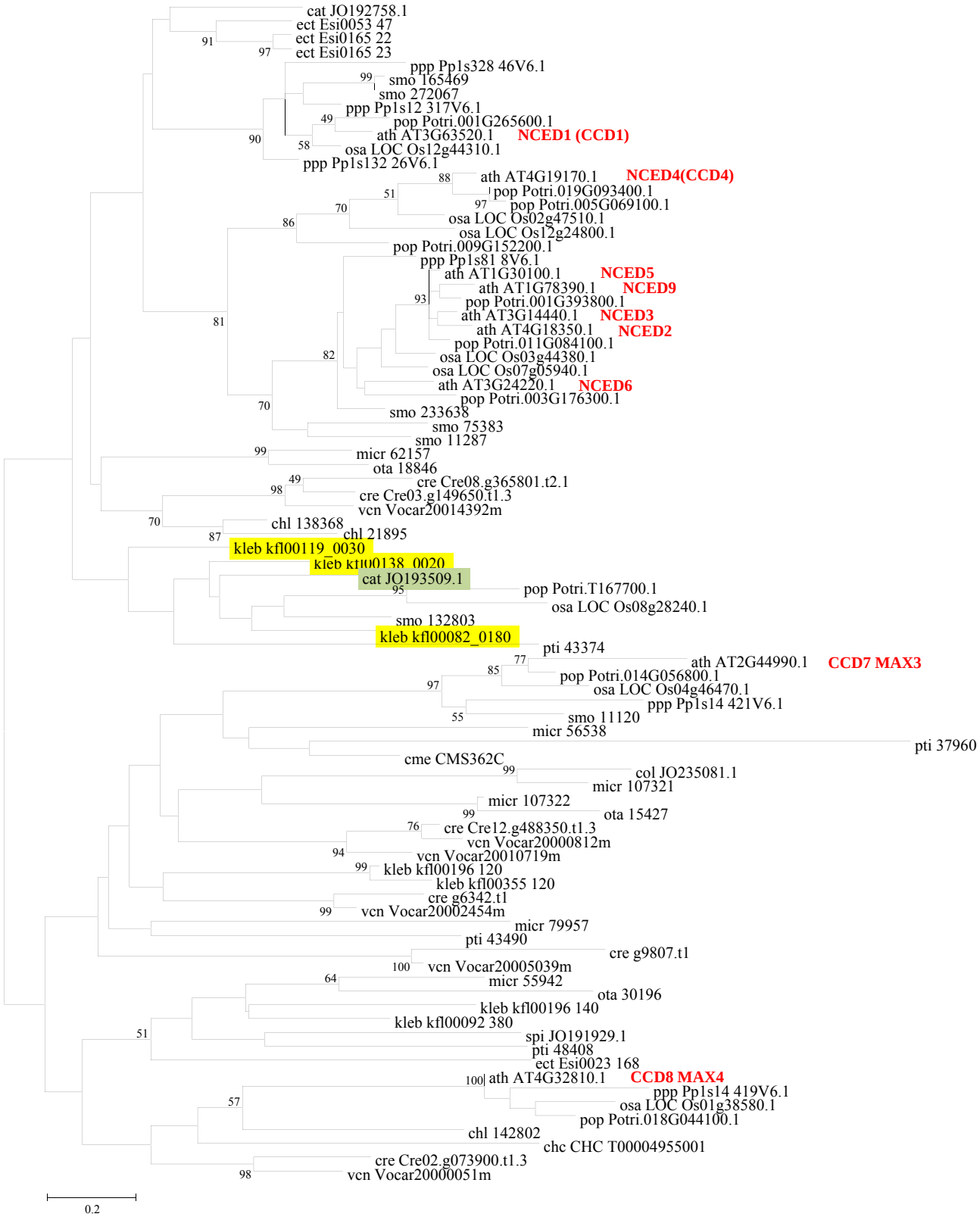
Supplementary Figure 21. Topology only style of Supplementary Figure 20.



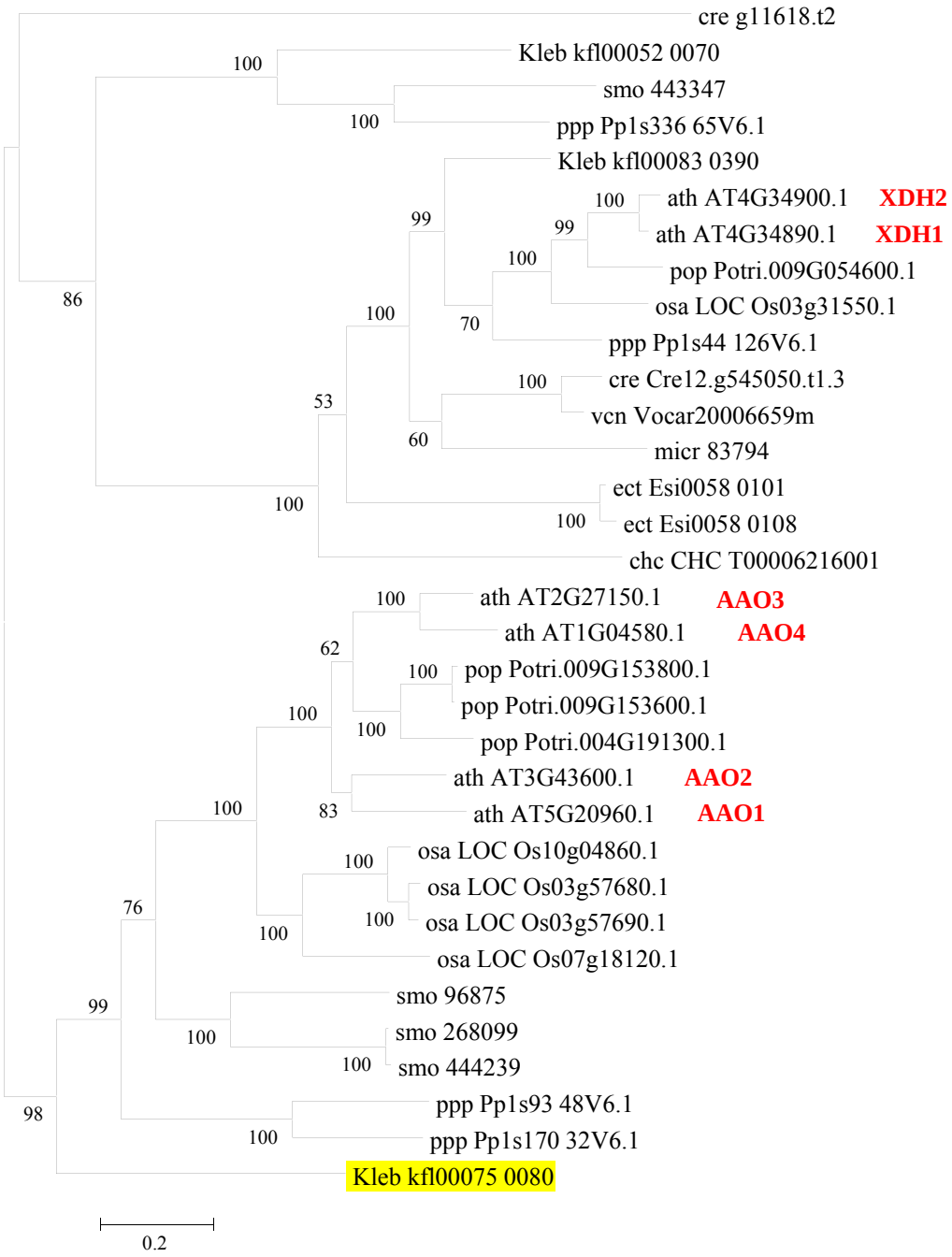
Supplementary Figure 22. Phylogenetic analysis of ZEP and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



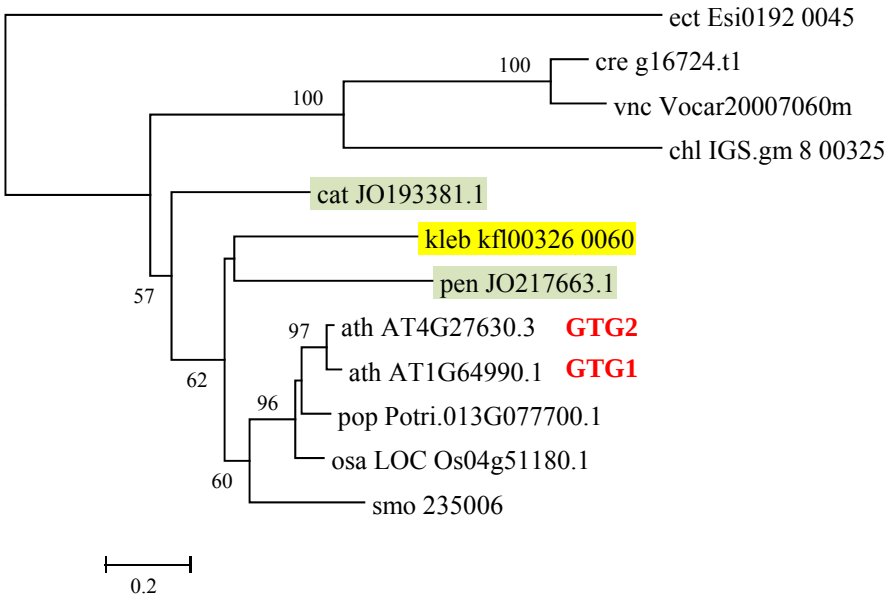
Supplementary Figure 23. Phylogenetic analysis of NSY and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LGF+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



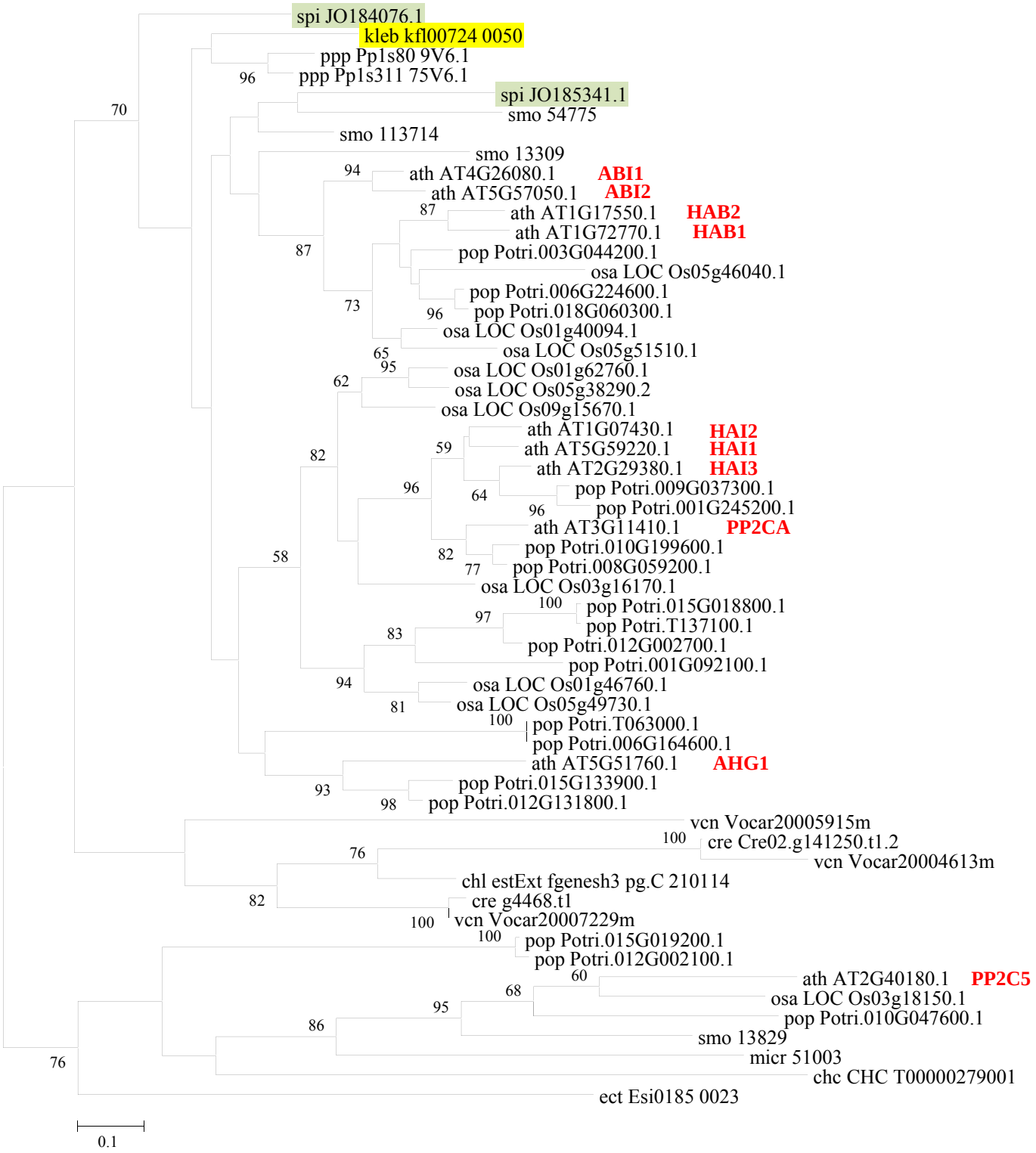
Supplementary Figure 24. Phylogenetic analysis of NCED and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



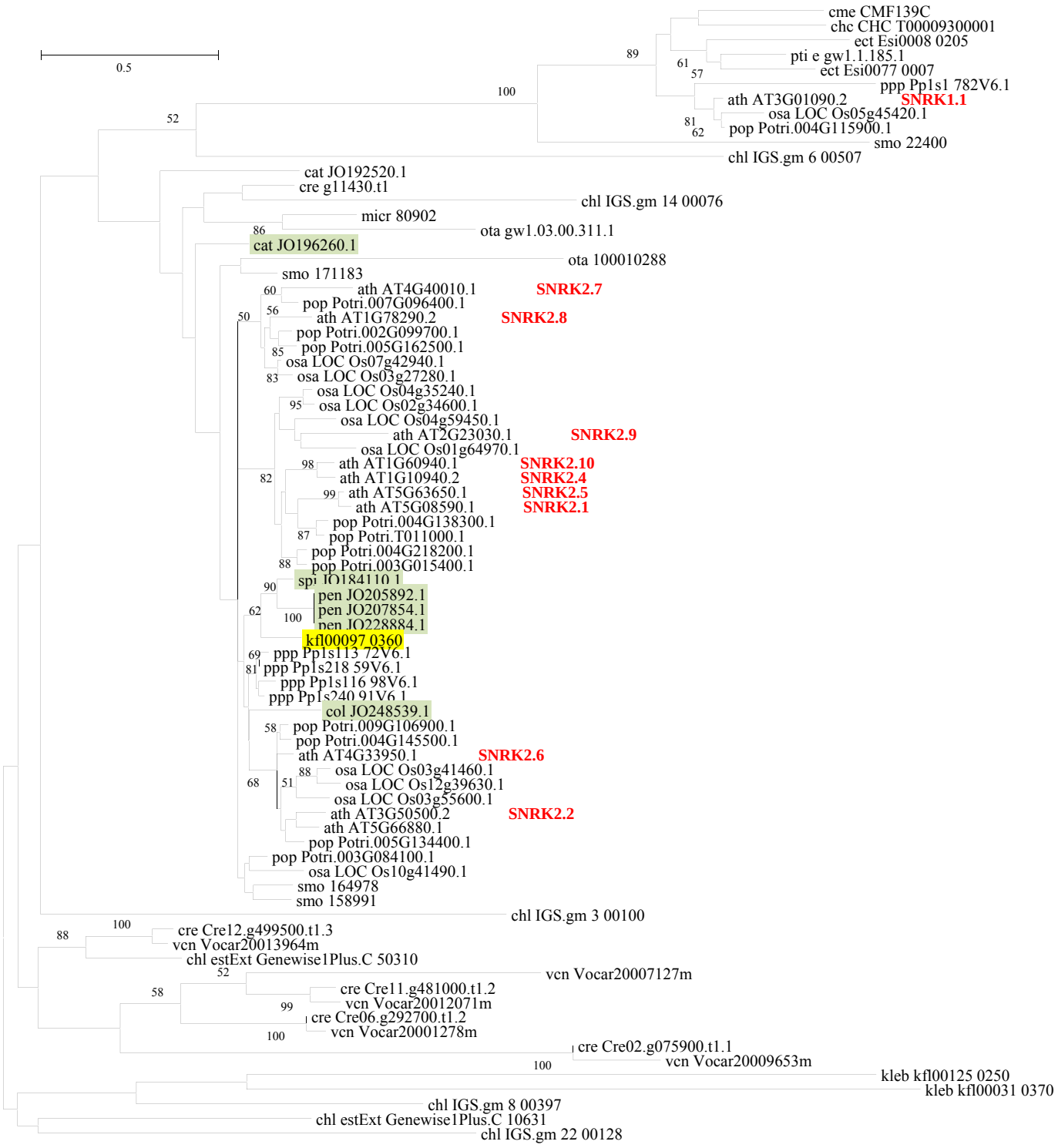
Supplementary Figure 25. Phylogenetic analysis of ABAO and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LGF+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



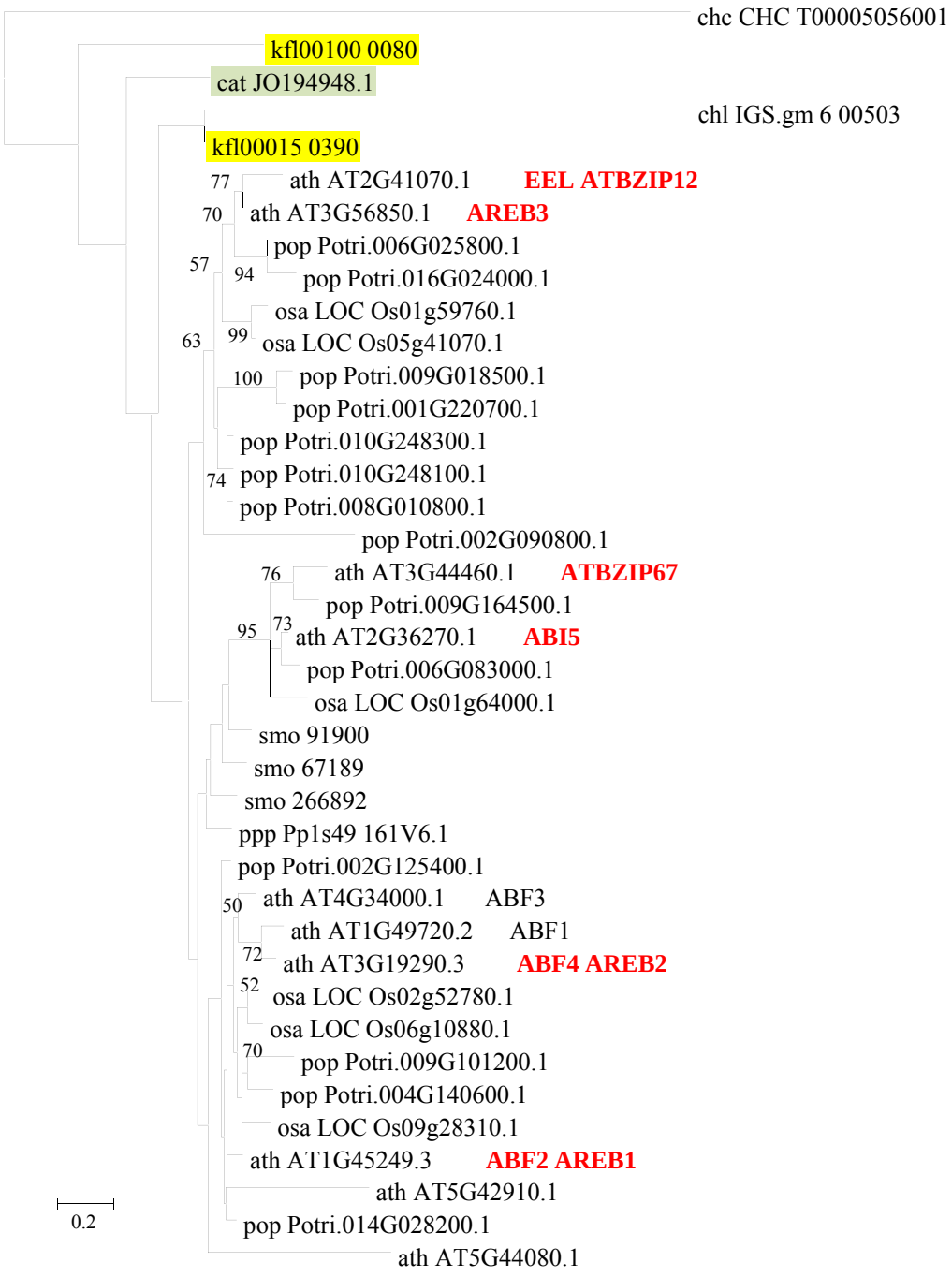
Supplementary Figure 26. Phylogenetic analysis of GTG and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LGF+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



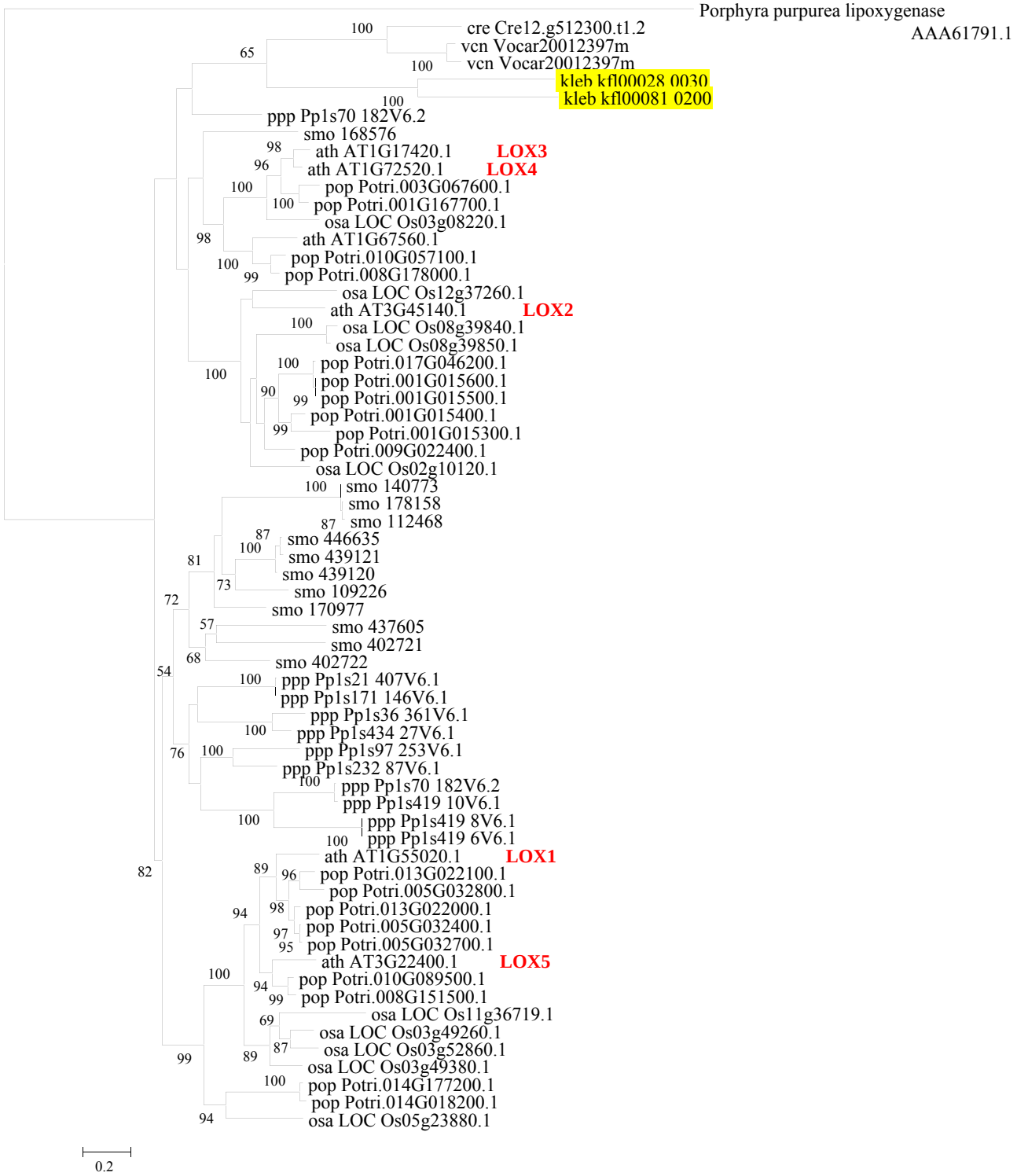
Supplementary Figure 27. Phylogenetic analysis of PP2Cs and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



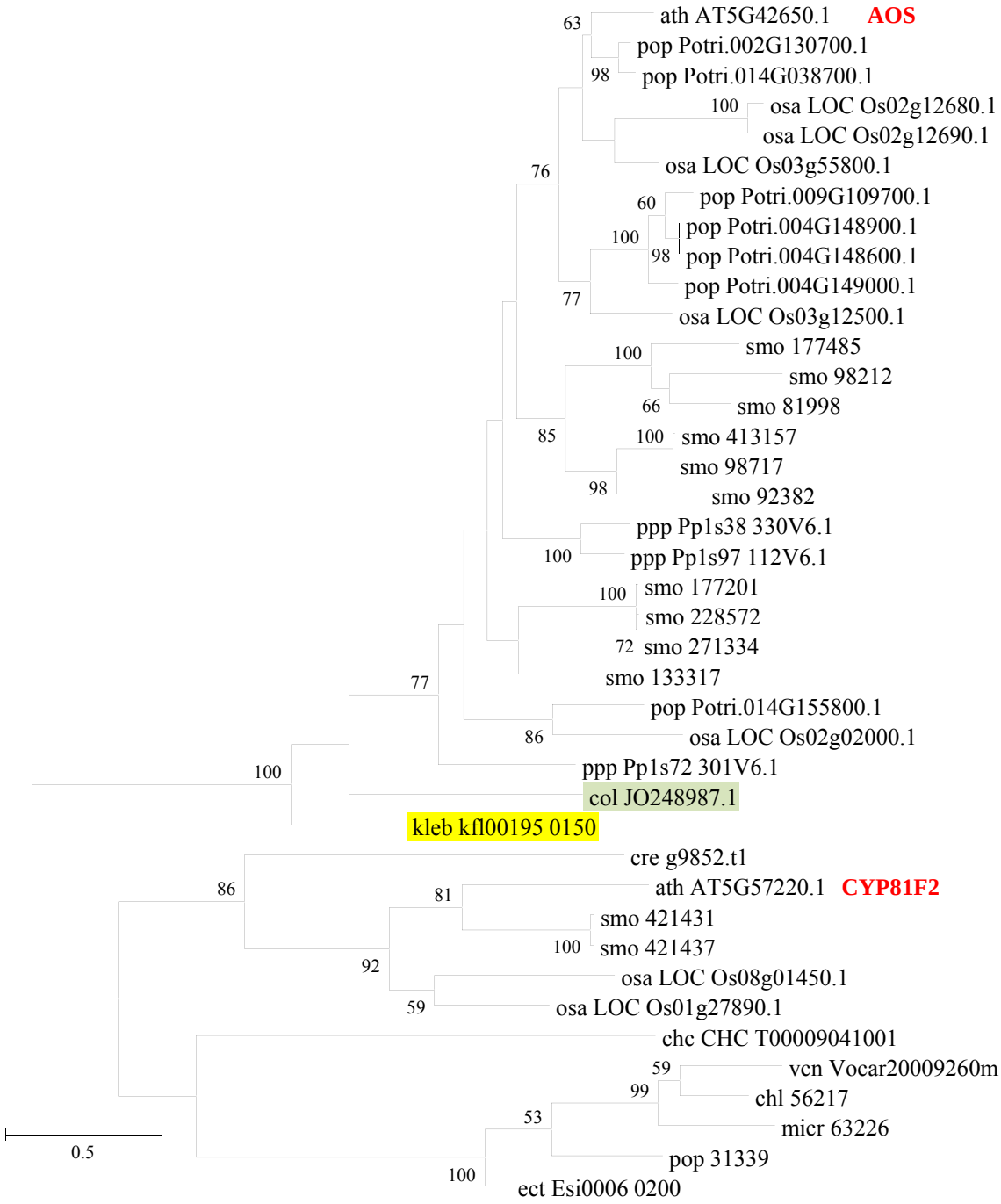
Supplementary Figure 28. Phylogenetic analysis of SNRKs and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



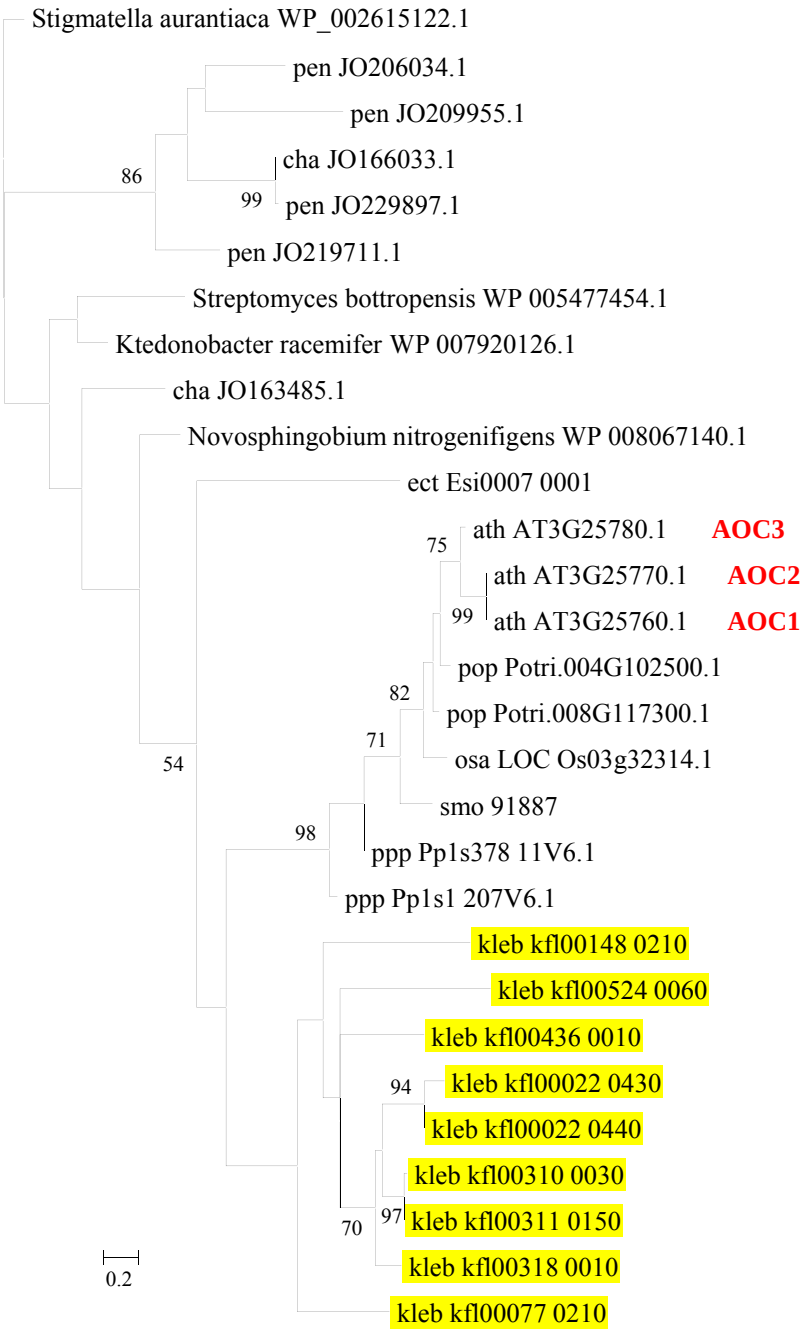
Supplementary Figure 29. Phylogenetic analysis of AREB and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "JTT+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



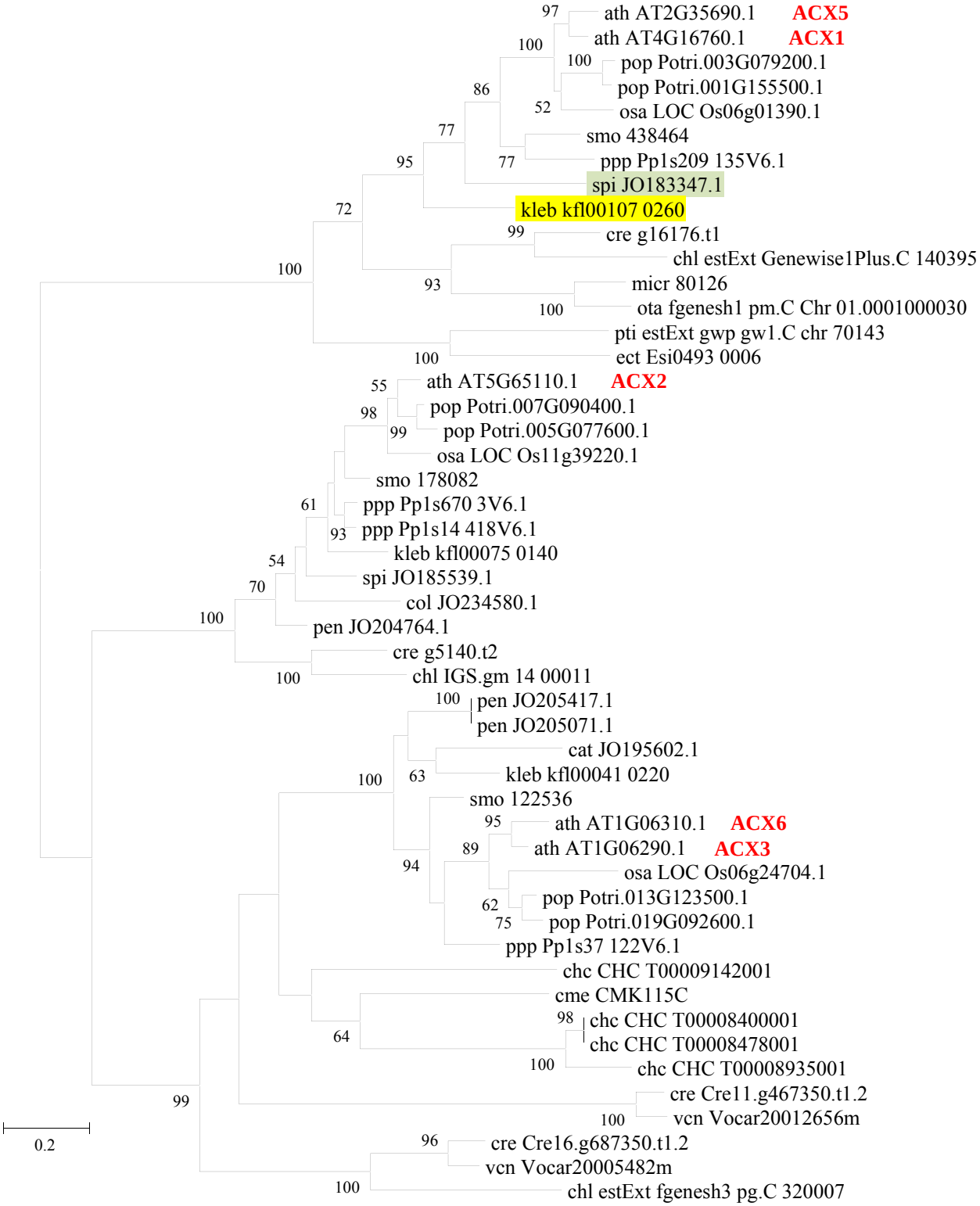
Supplementary Figure 30. Phylogenetic analysis of LOX and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



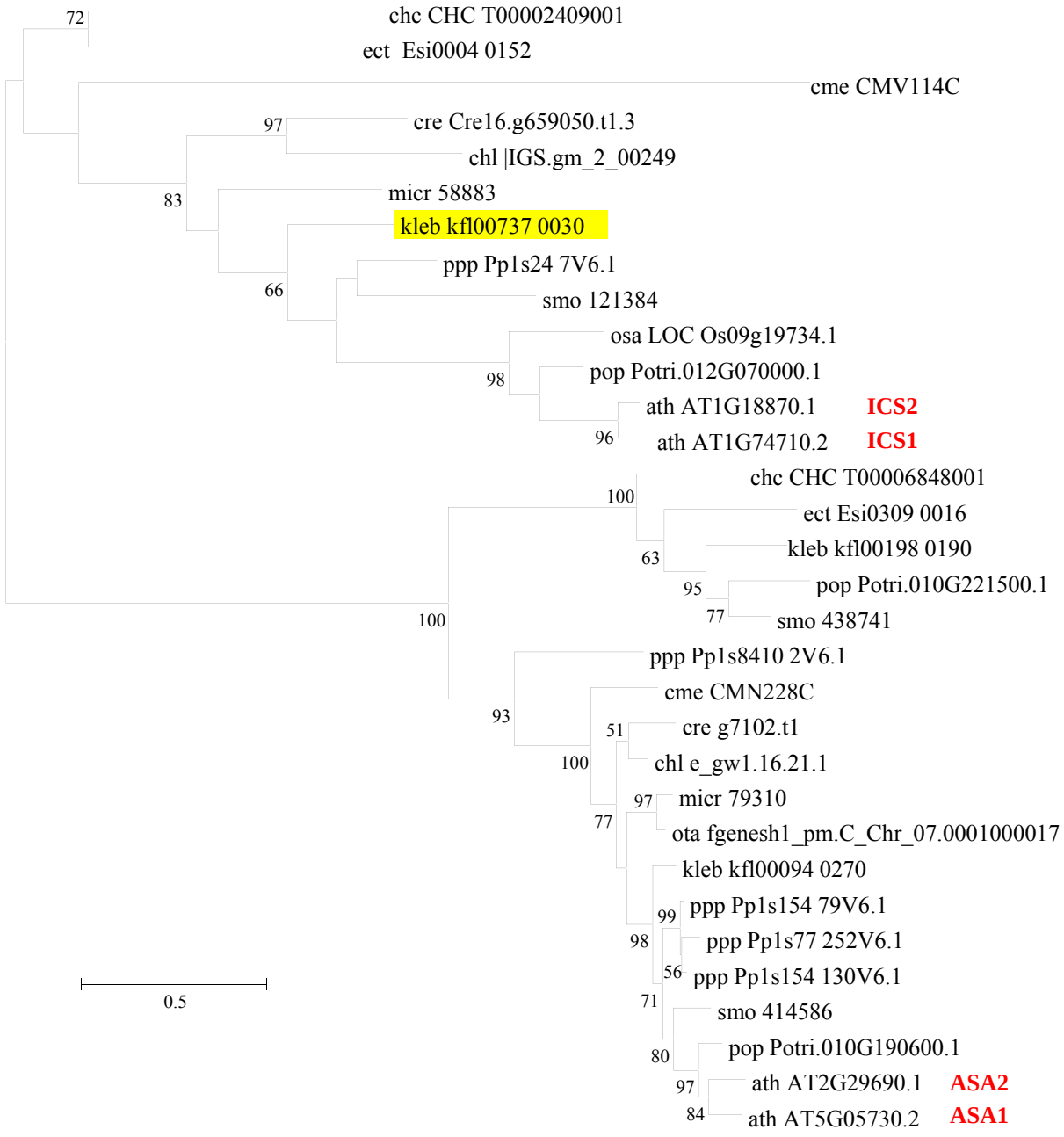
Supplementary Figure 31. Phylogenetic analysis of AOS and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



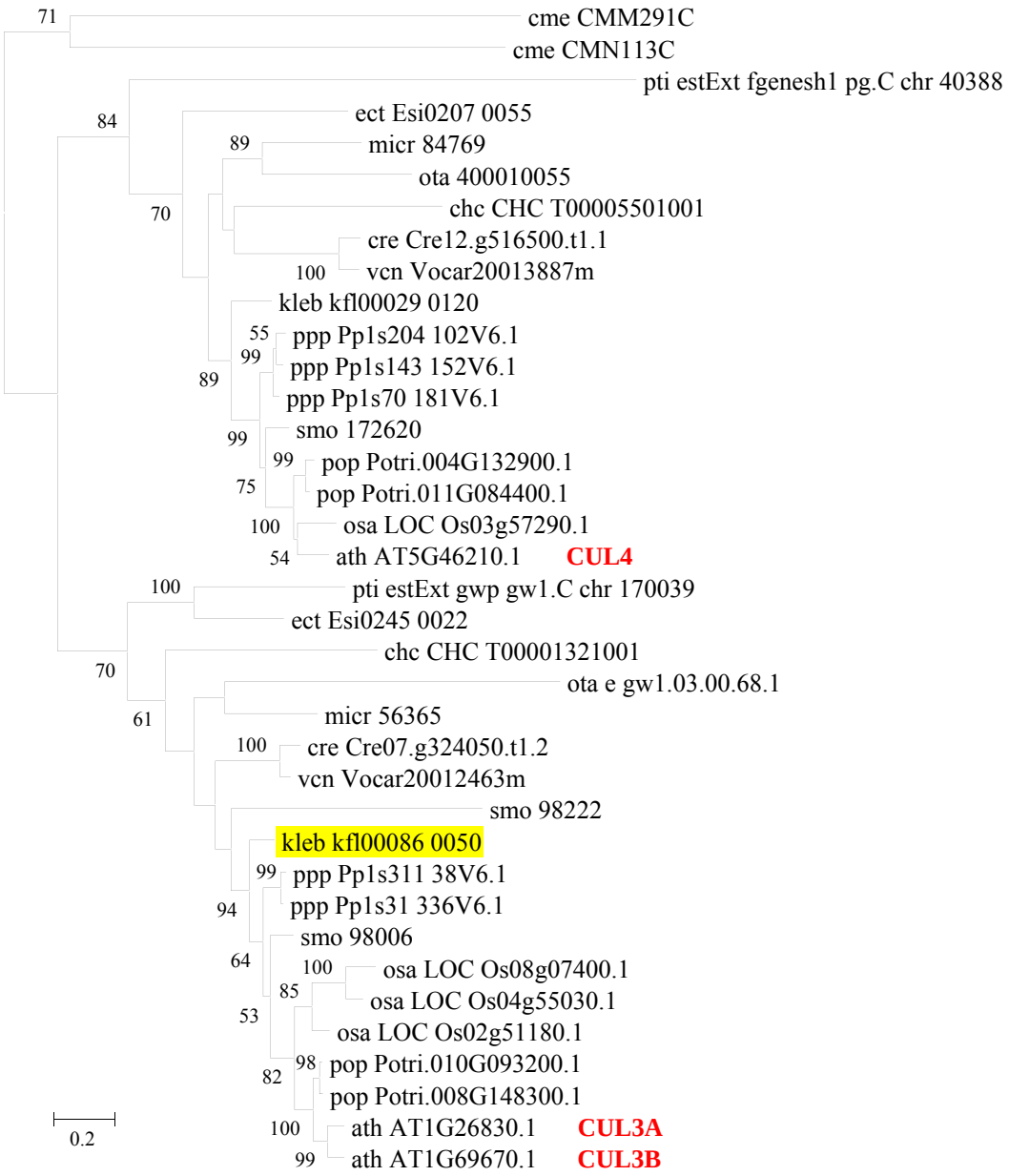
Supplementary Figure 32. Phylogenetic analysis of AOC and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "WAG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



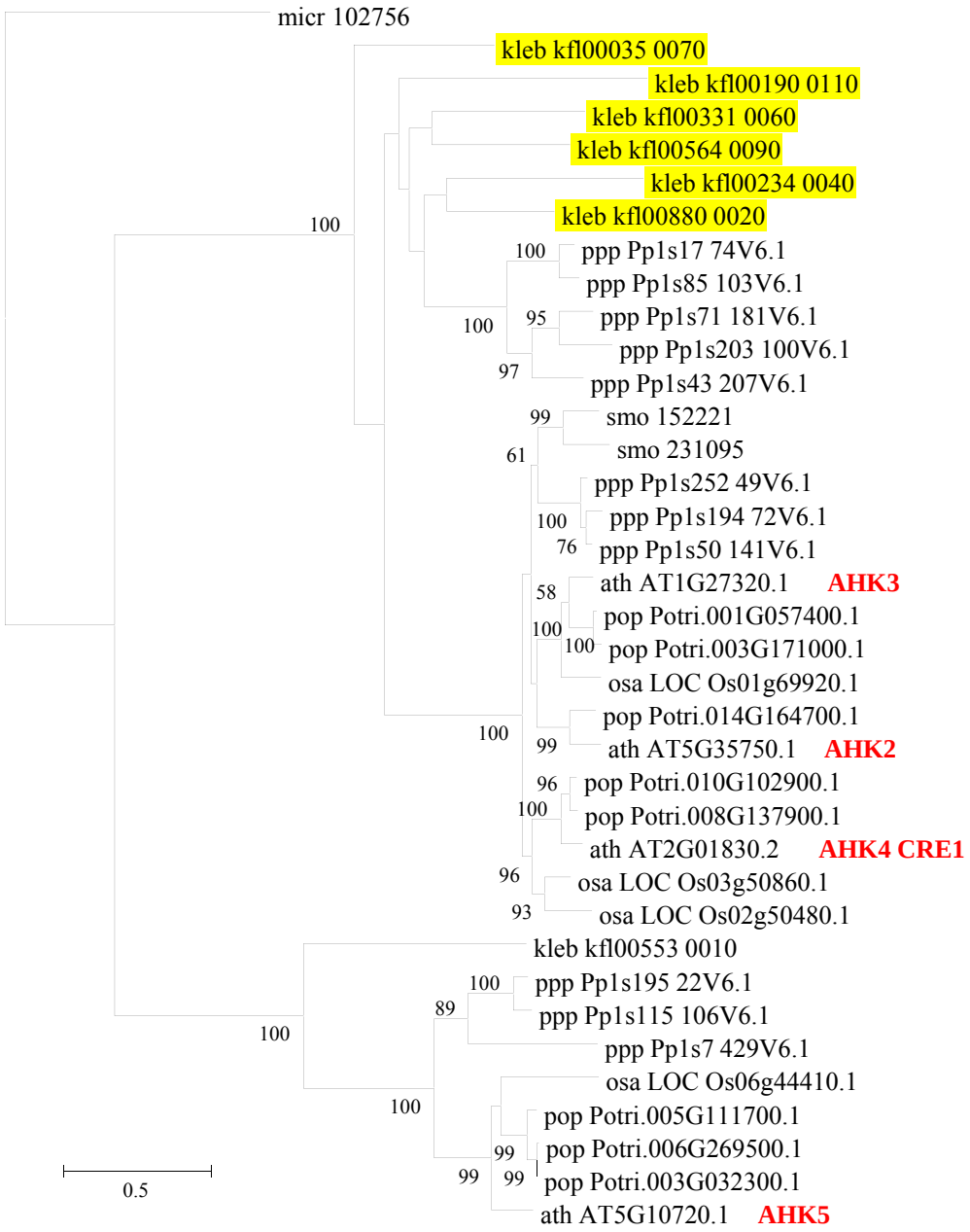
Supplementary Figure 33. Phylogenetic analysis of ACX and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LGF+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



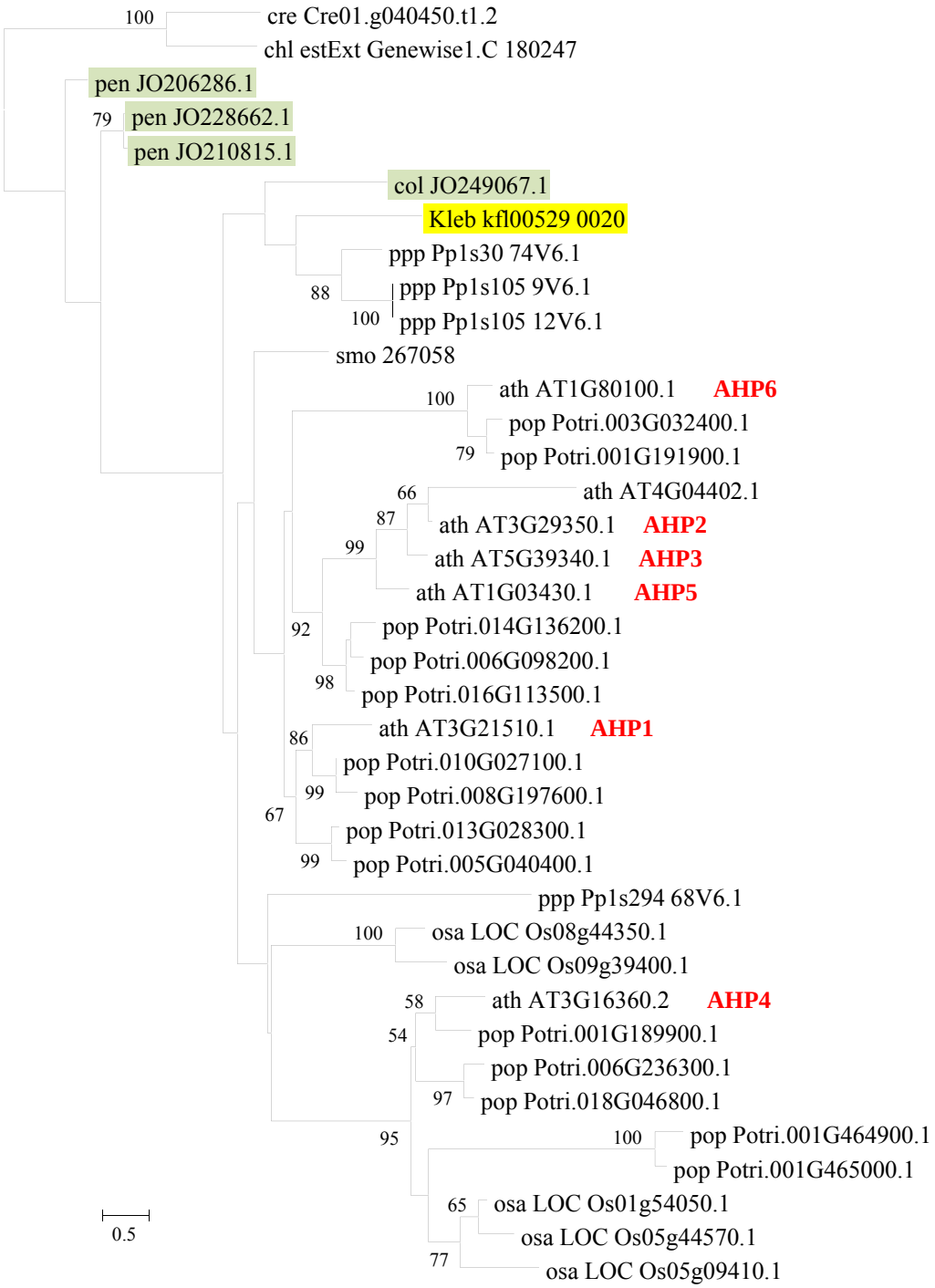
Supplementary Figure 34. Phylogenetic analysis of ICS and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



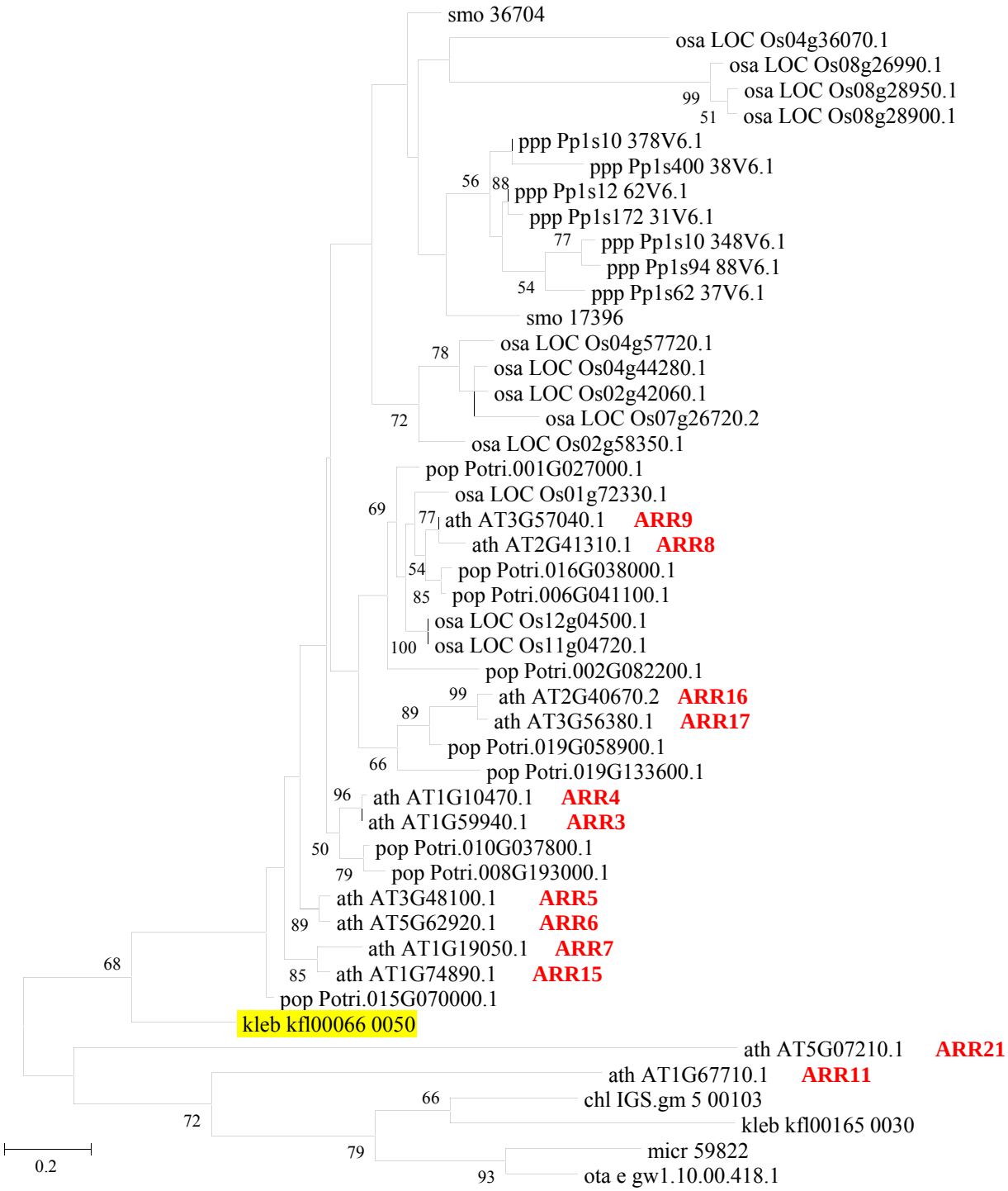
Supplementary Figure 35. Phylogenetic analysis of CUL3 and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



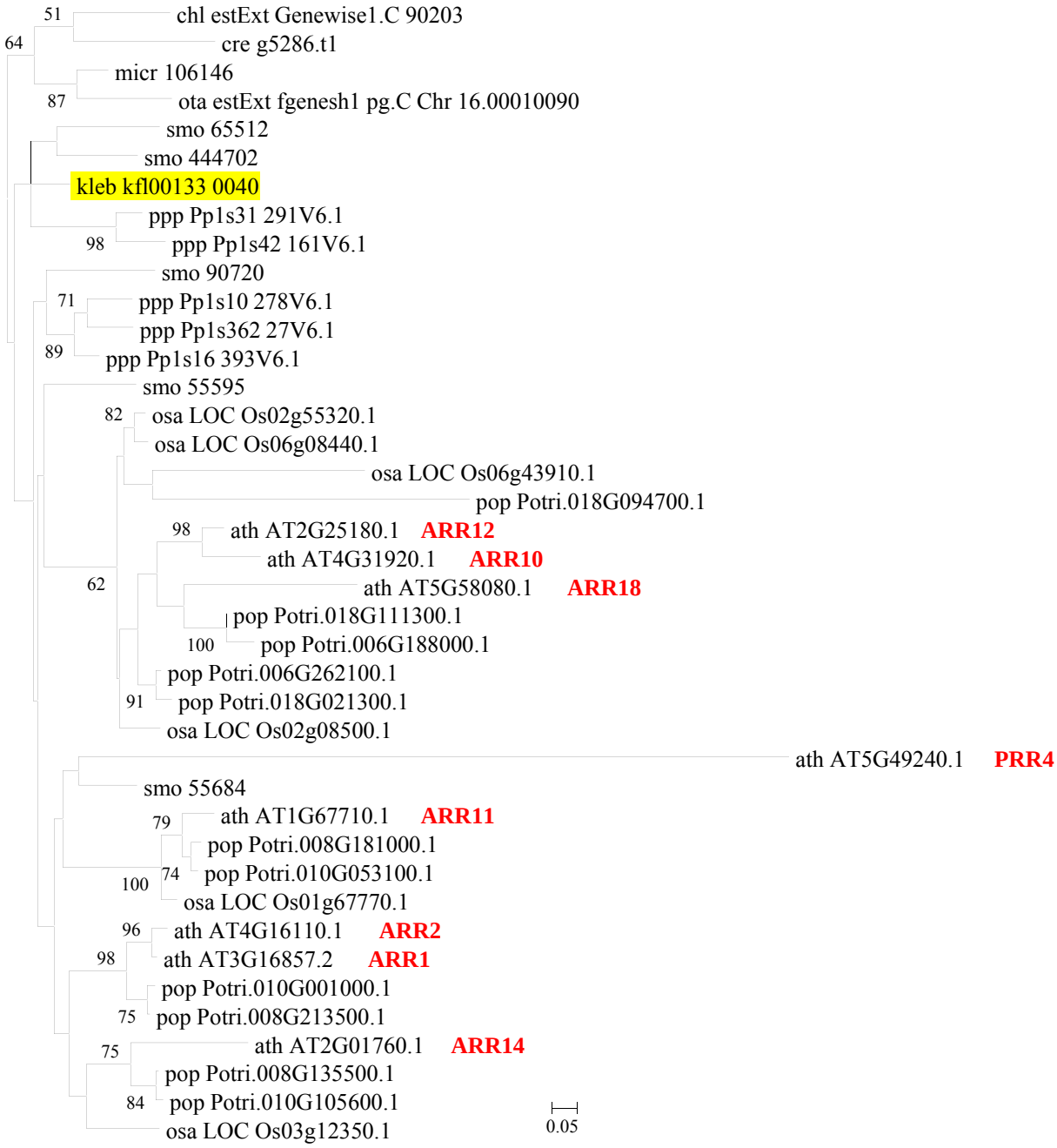
Supplementary Figure 36. Phylogenetic analysis of AHK and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



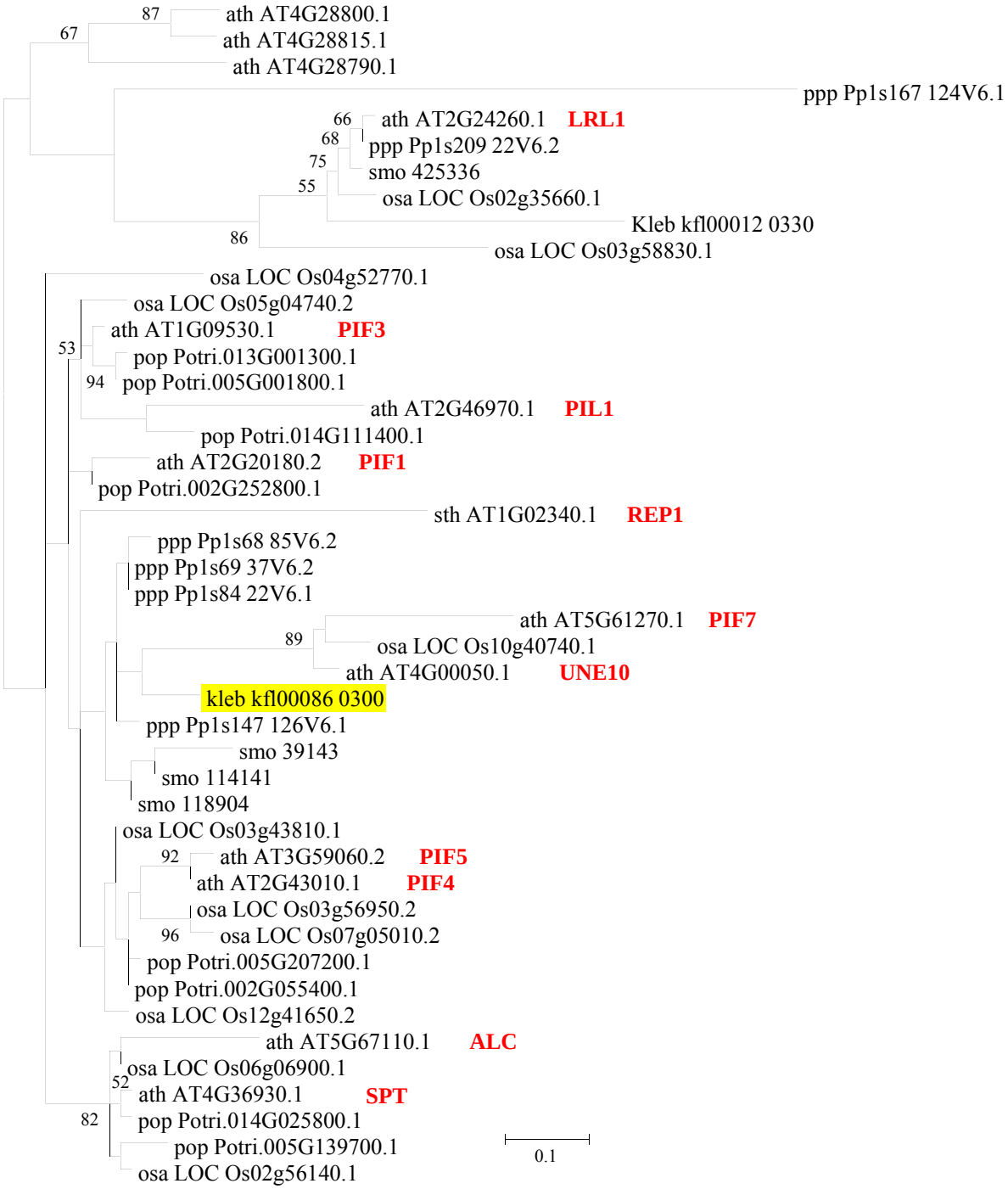
Supplementary Figure 37. Phylogenetic analysis of AHP and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "JTT+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



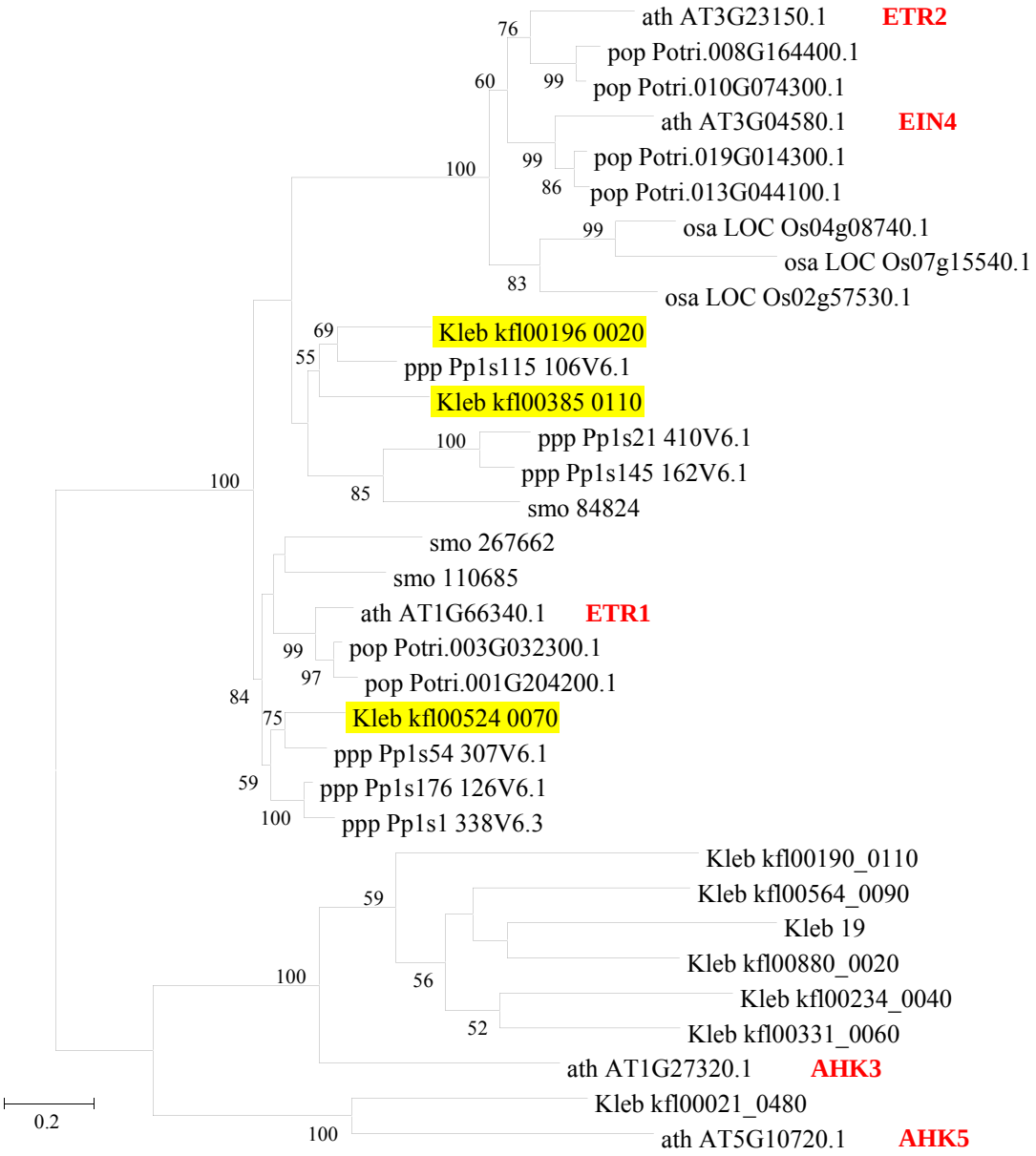
Supplementary Figure 38. Phylogenetic analysis of ARRA and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



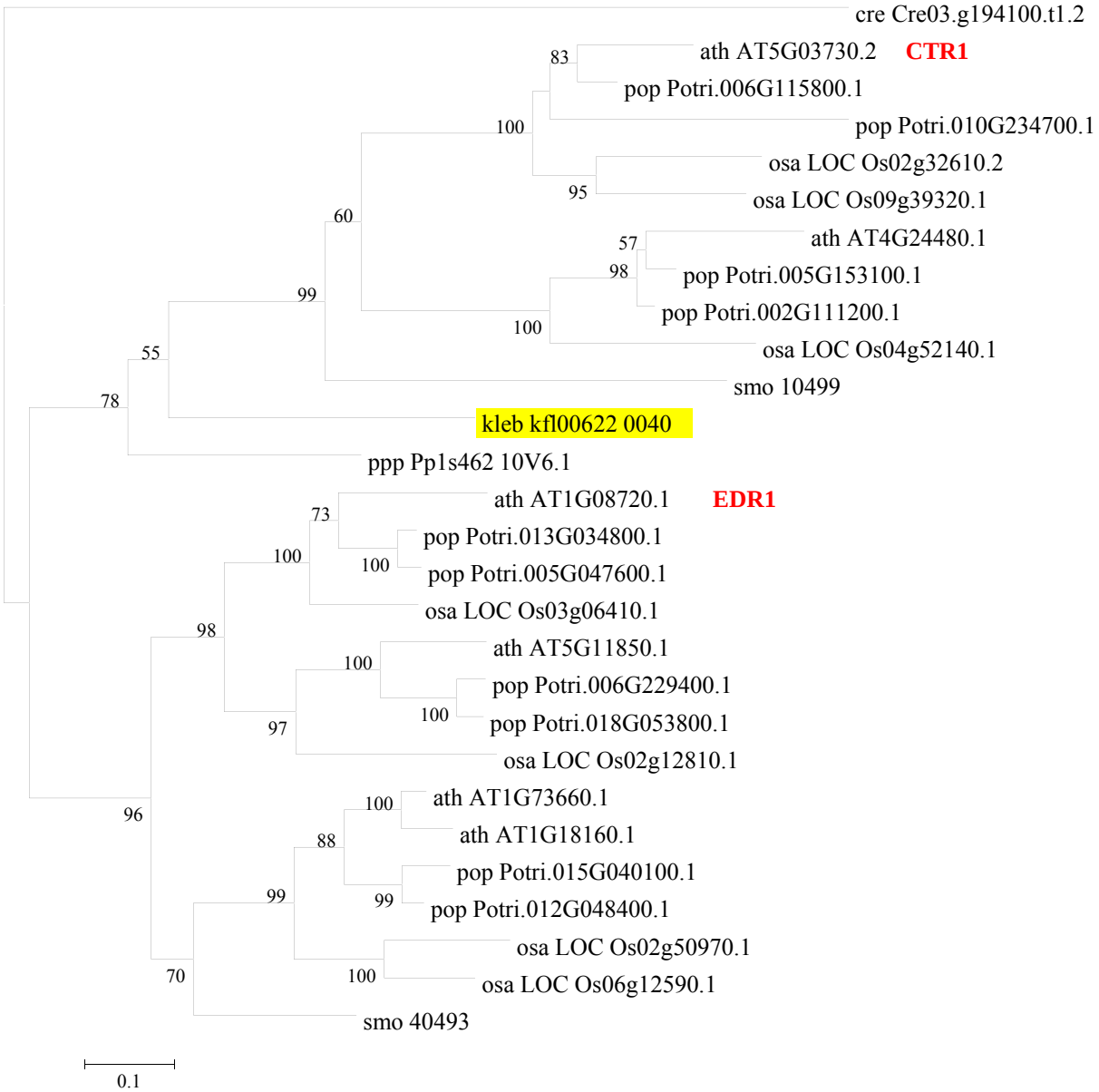
Supplementary Figure 39. Phylogenetic analysis of ARRB and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



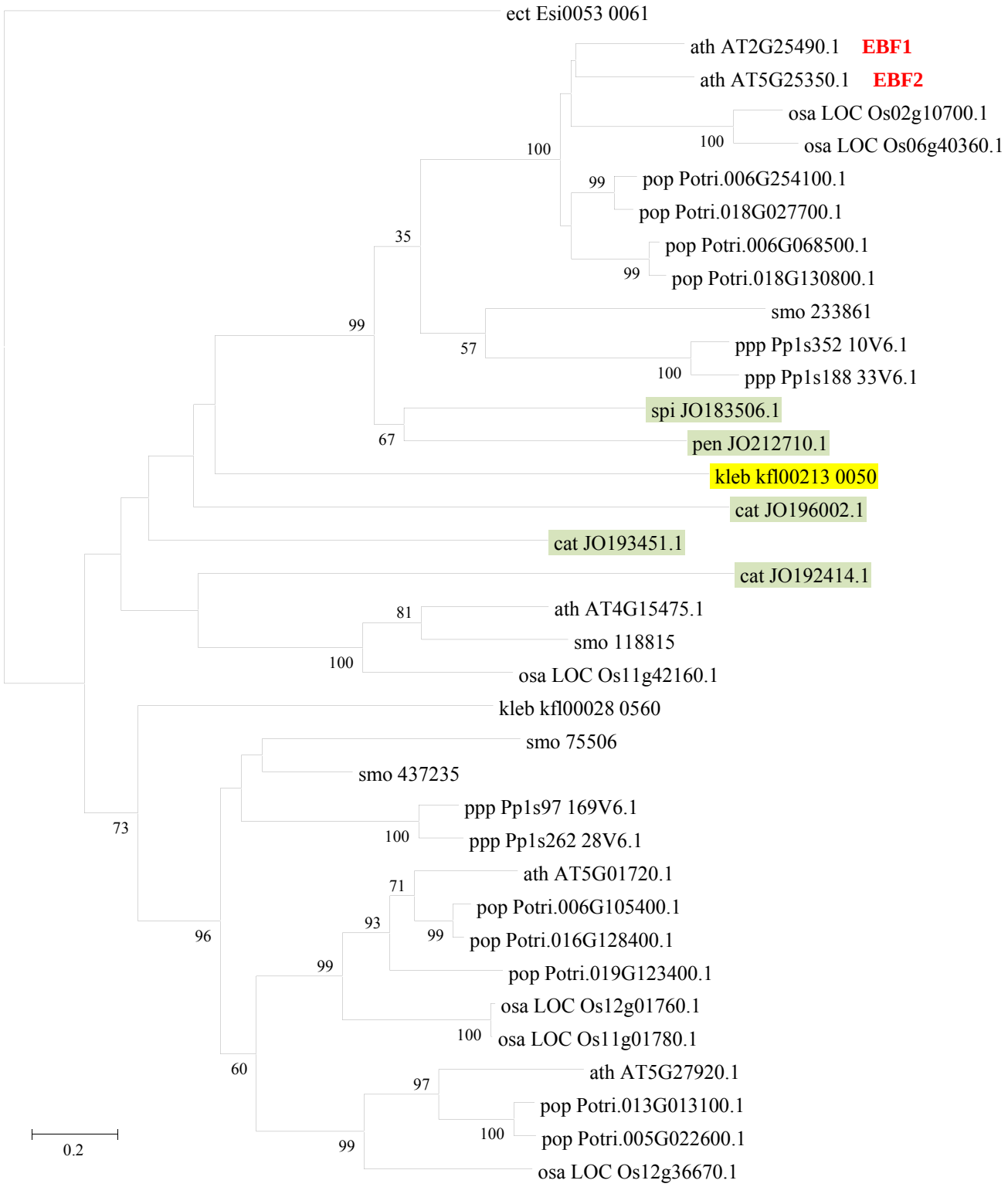
Supplementary Figure 40. Phylogenetic analysis of PIF and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



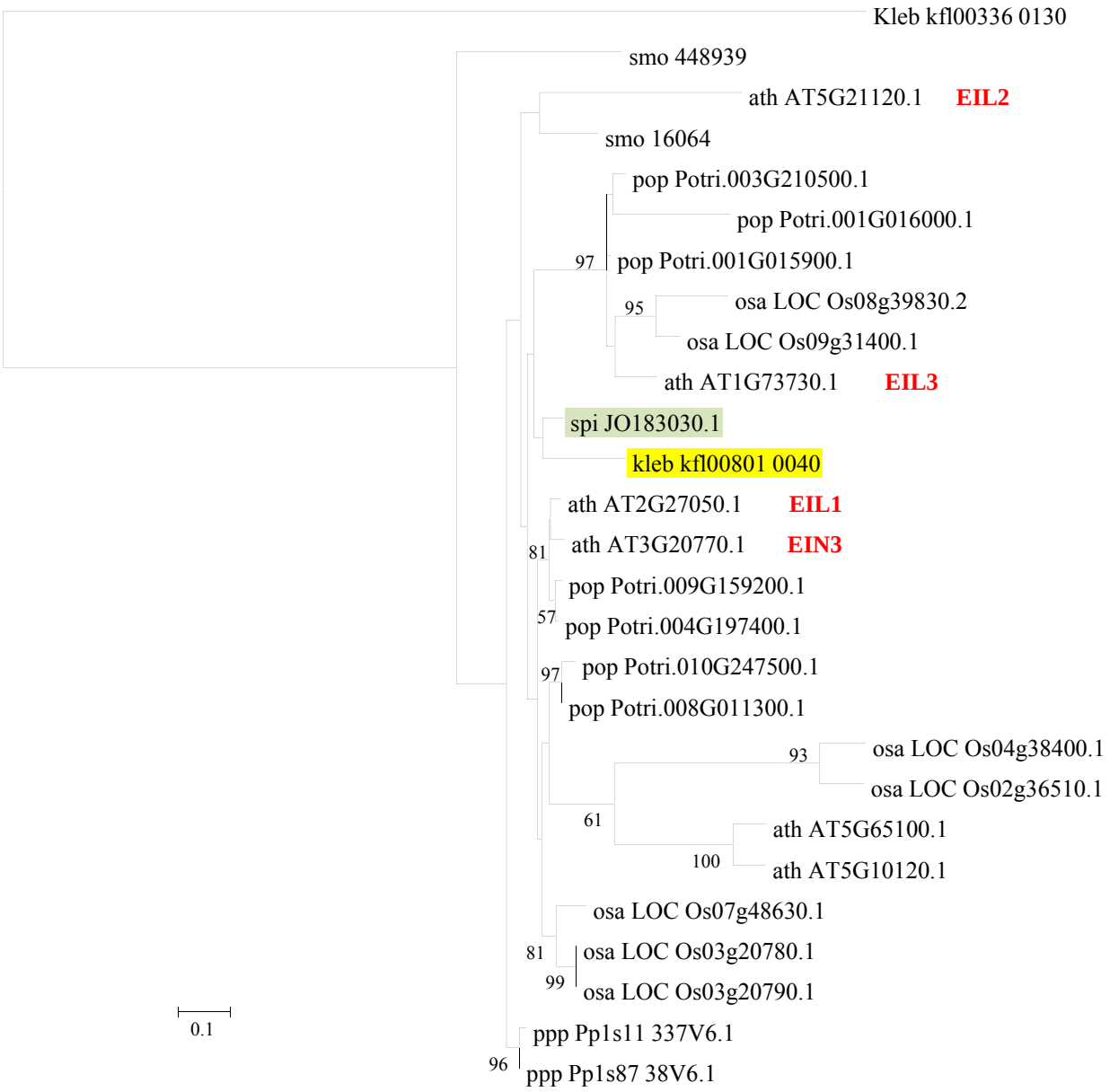
Supplementary Figure 41. Phylogenetic analysis of ETR and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "JTT+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



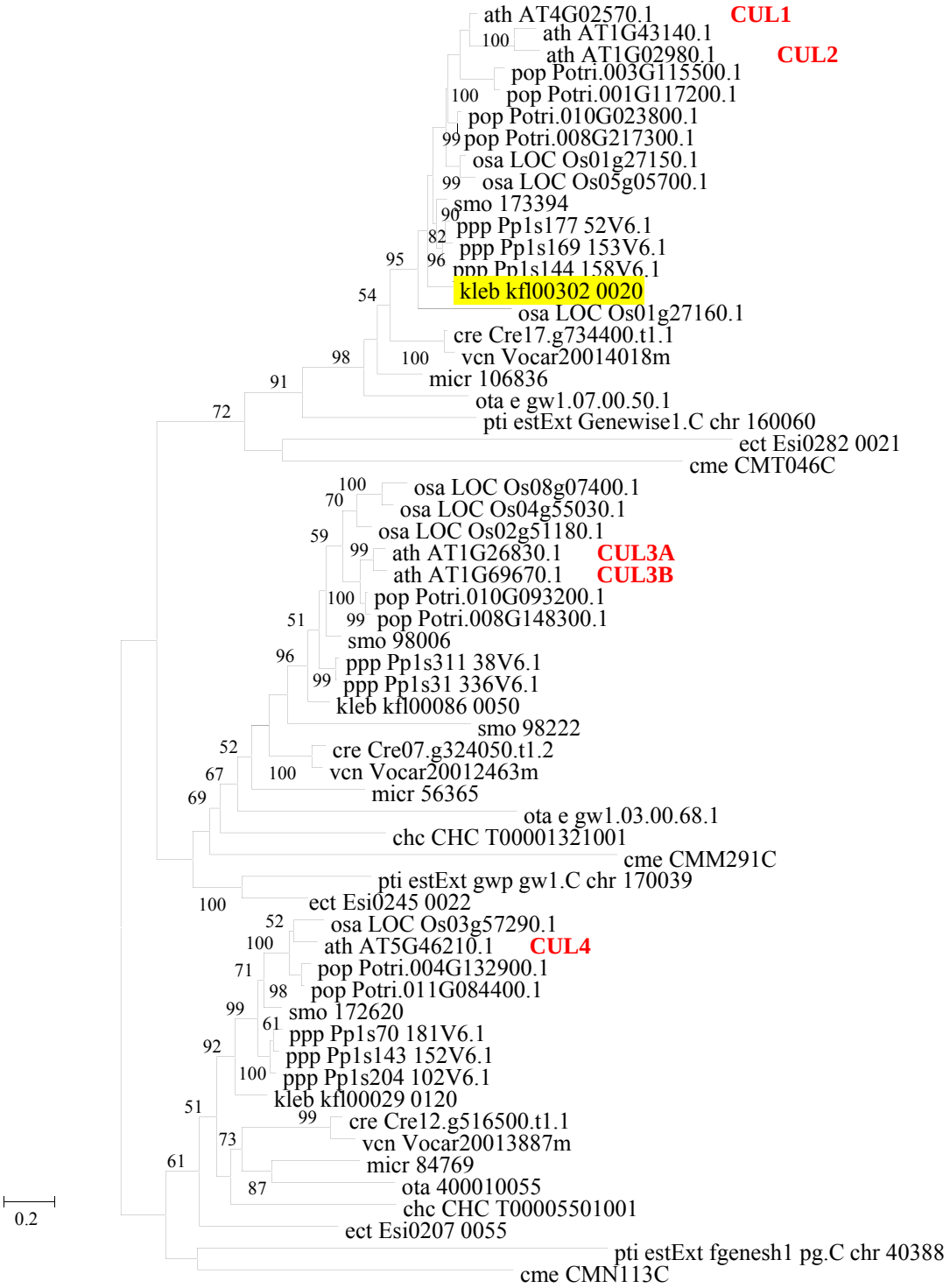
Supplementary Figure 42. Phylogenetic analysis of CTR and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



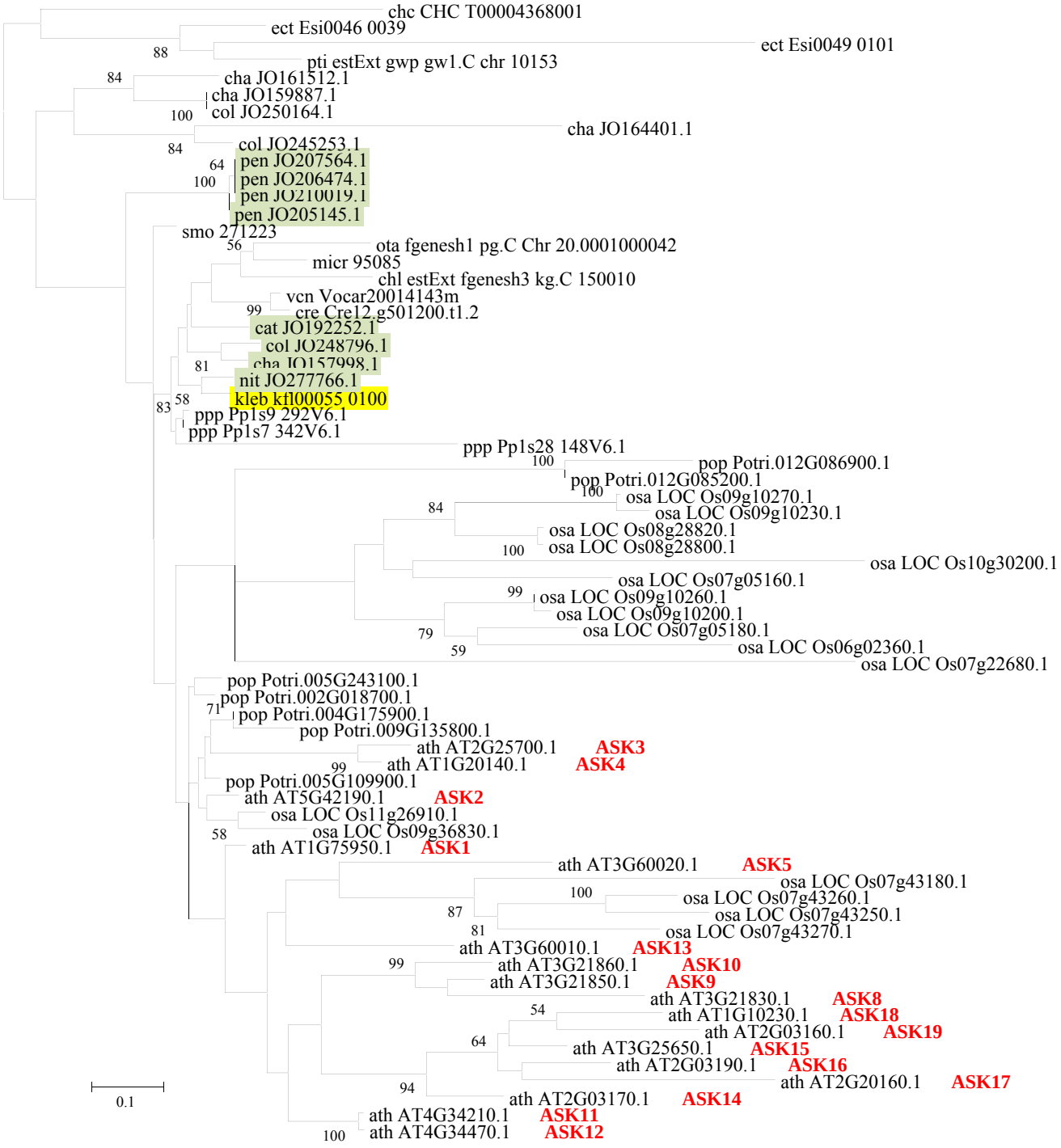
Supplementary Figure 43. Phylogenetic analysis of EBF and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



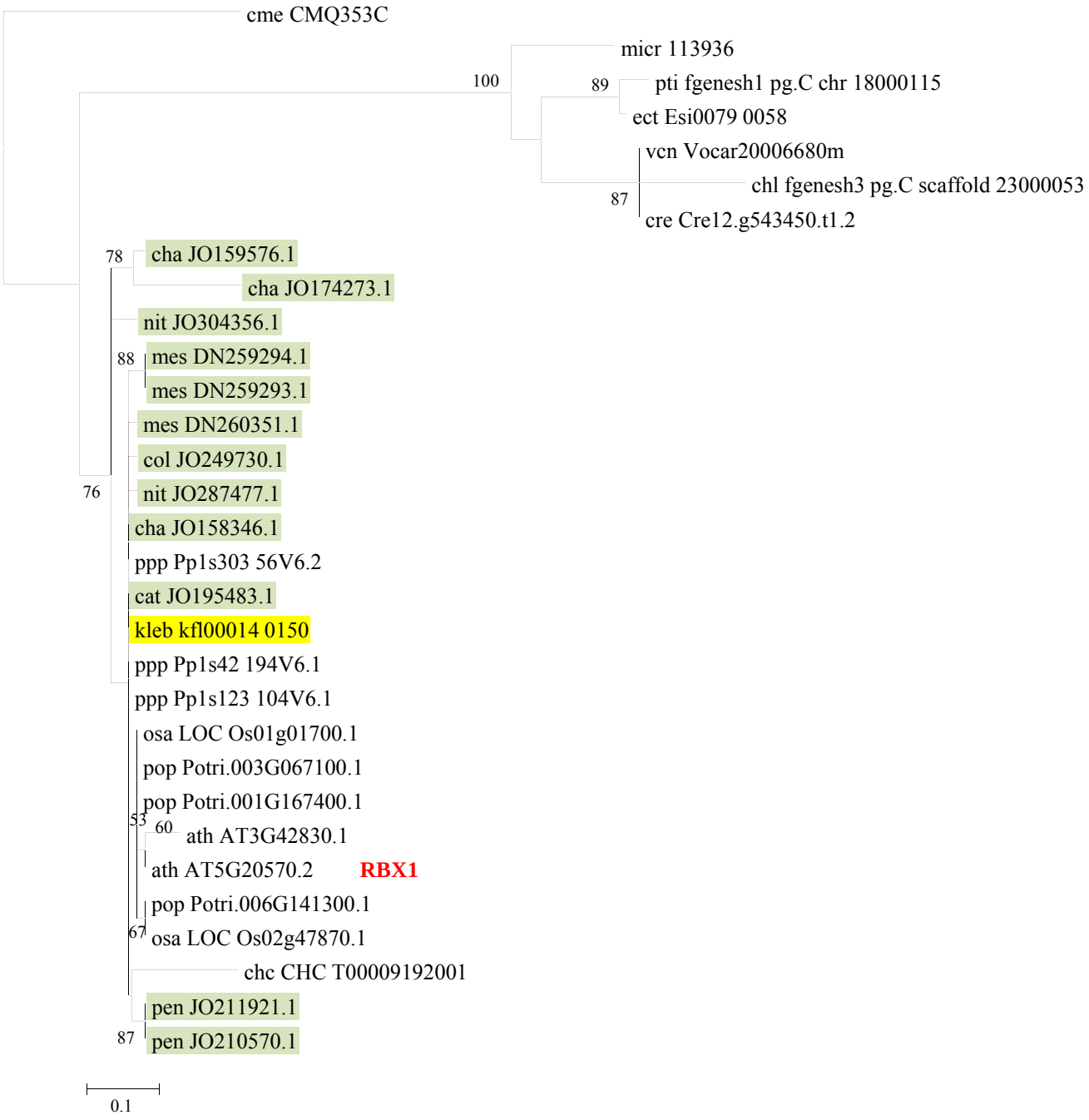
Supplementary Figure 44. Phylogenetic analysis of EIN3 and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "JTT+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



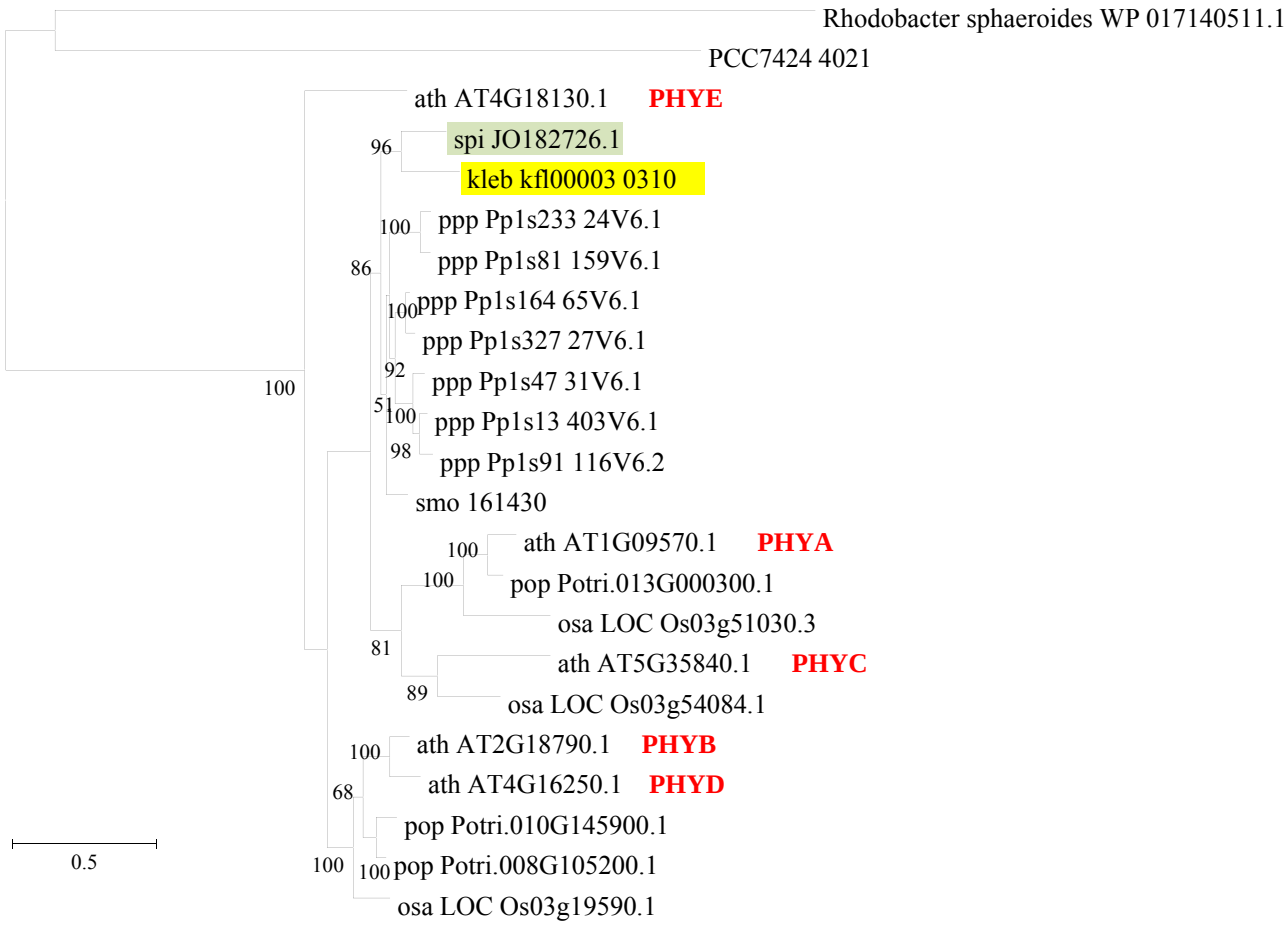
Supplementary Figure 45. Phylogenetic analysis of CUL1 and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



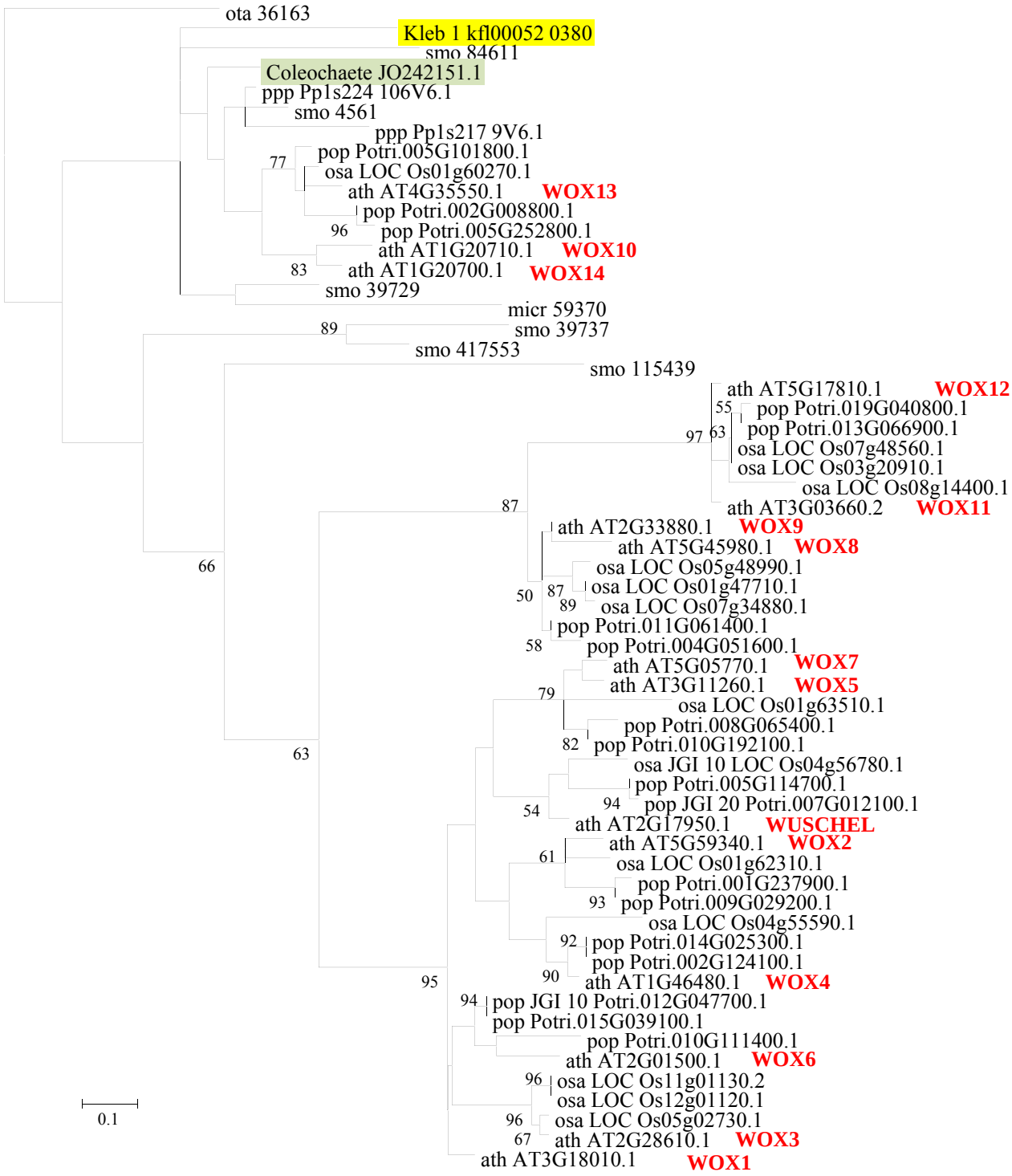
Supplementary Figure 46. Phylogenetic analysis of ASK and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



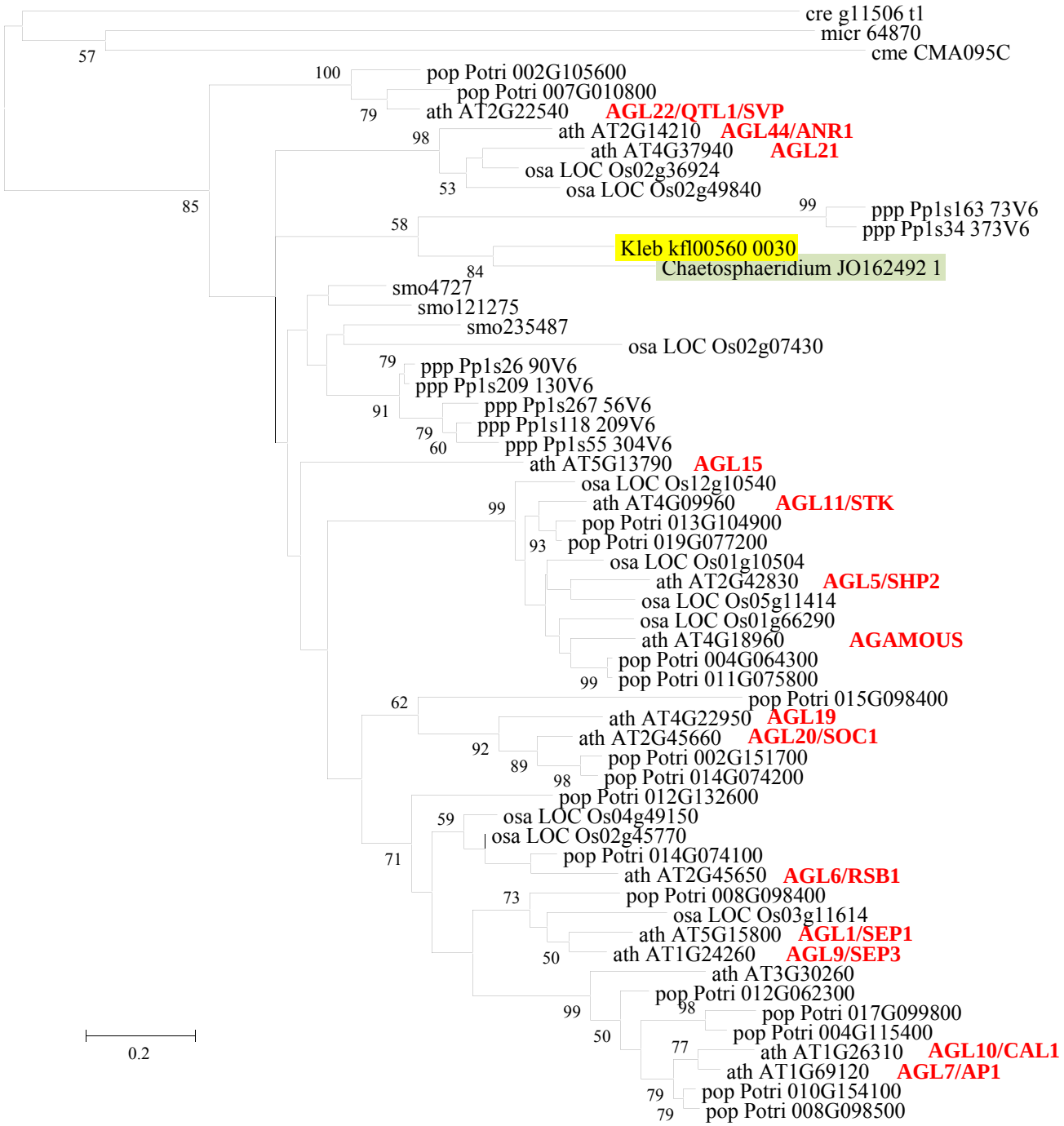
Supplementary Figure 47. Phylogenetic analysis of RBX and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "JTT+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



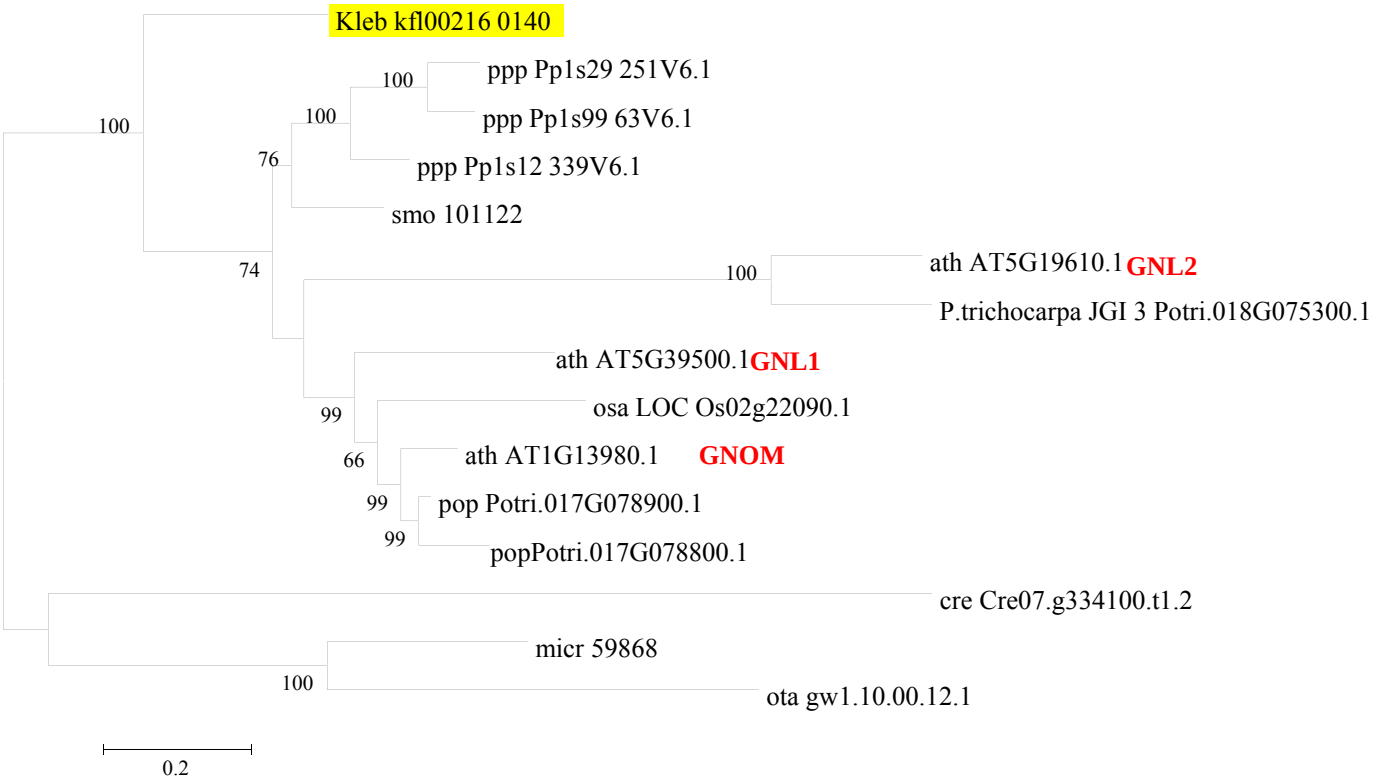
Supplementary Figure 48. Phylogenetic analysis of phytochrome and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "JTT+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



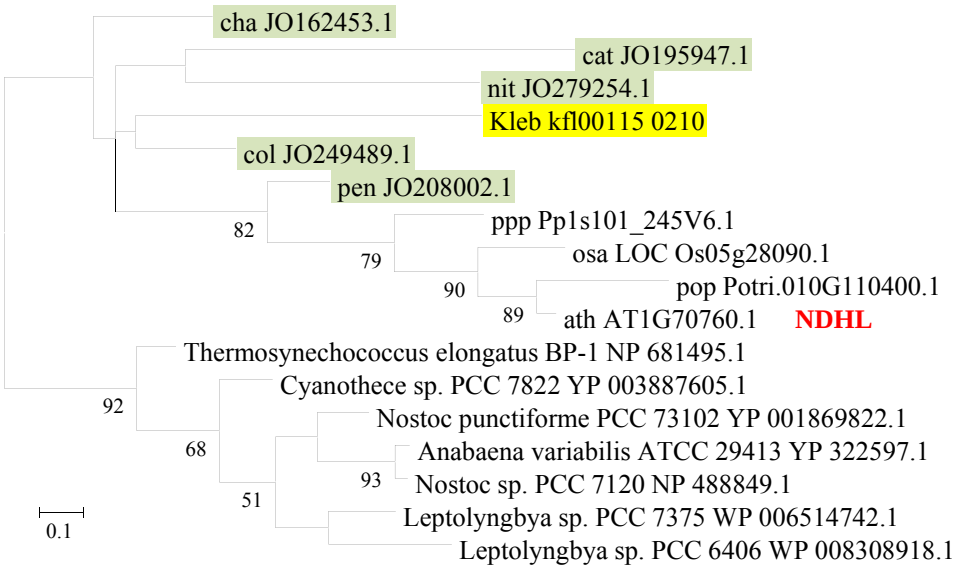
Supplementary Figure 49. Phylogenetic analysis of WUSCHEL proteins and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



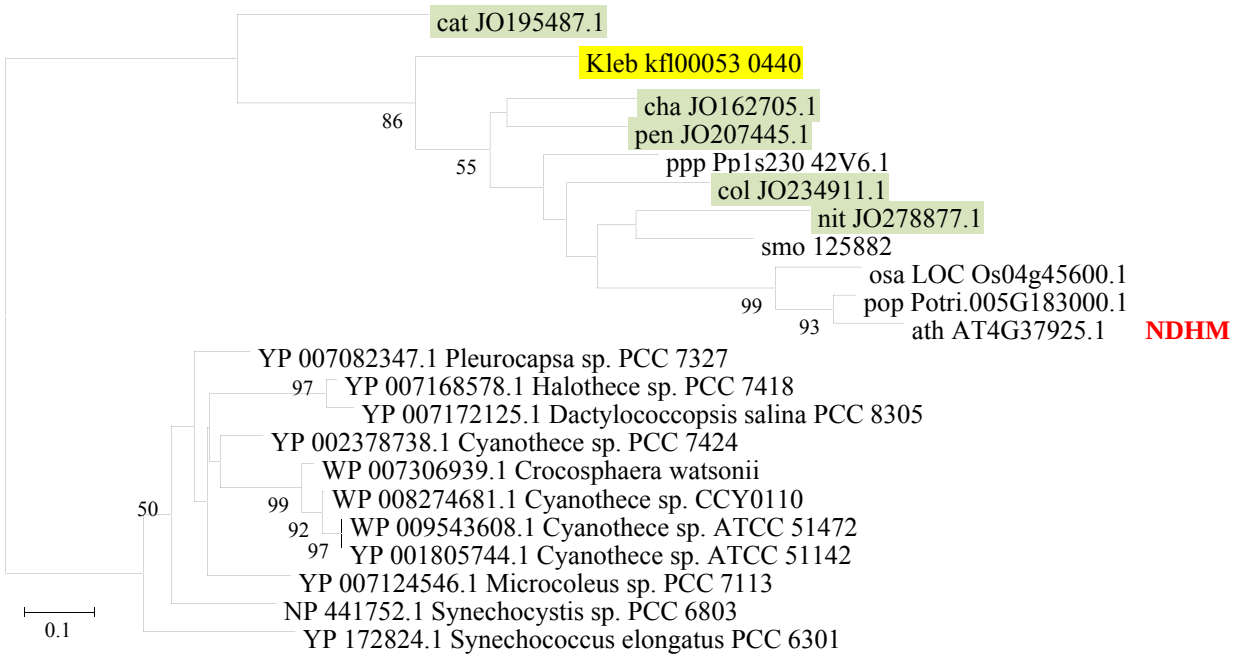
Supplementary Figure 50. Phylogenetic analysis of AGAMOUS like MADS-box proteins and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



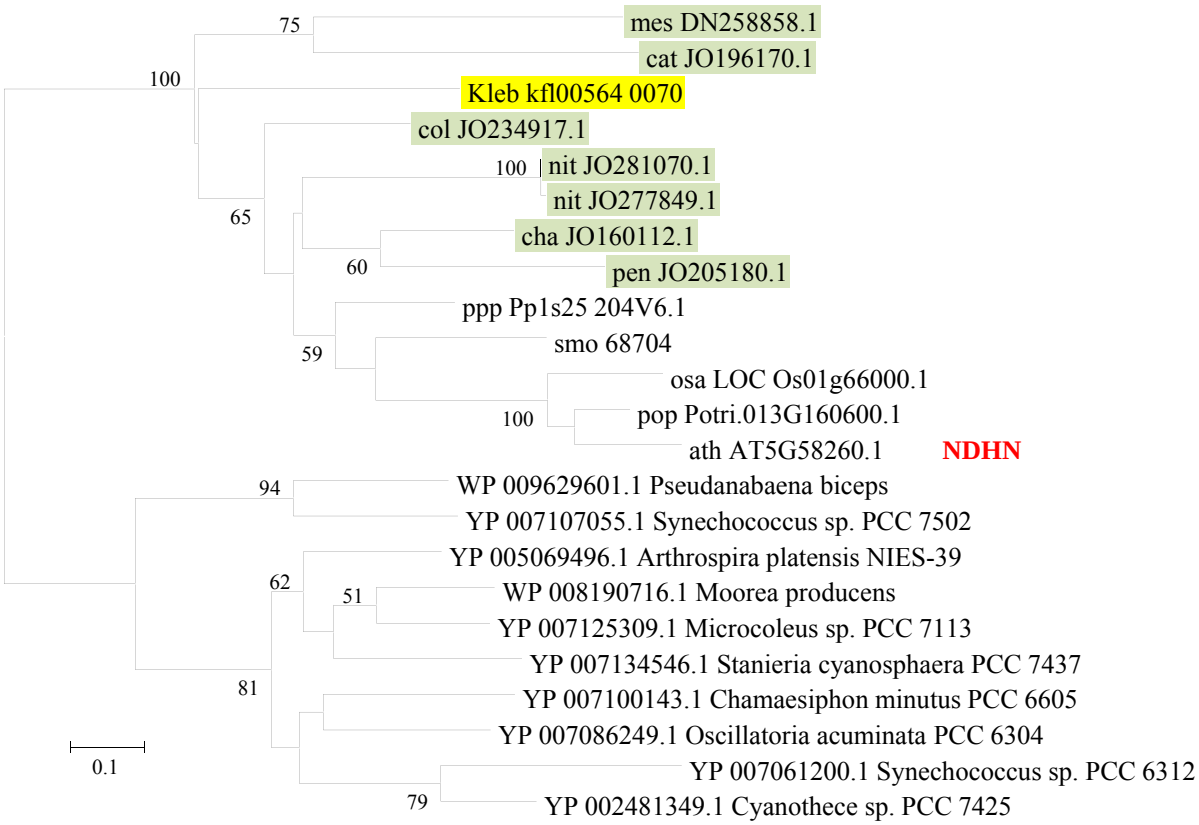
Supplementary Figure 51. Phylogenetic analysis of GNOM proteins and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "JTT+G+F". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



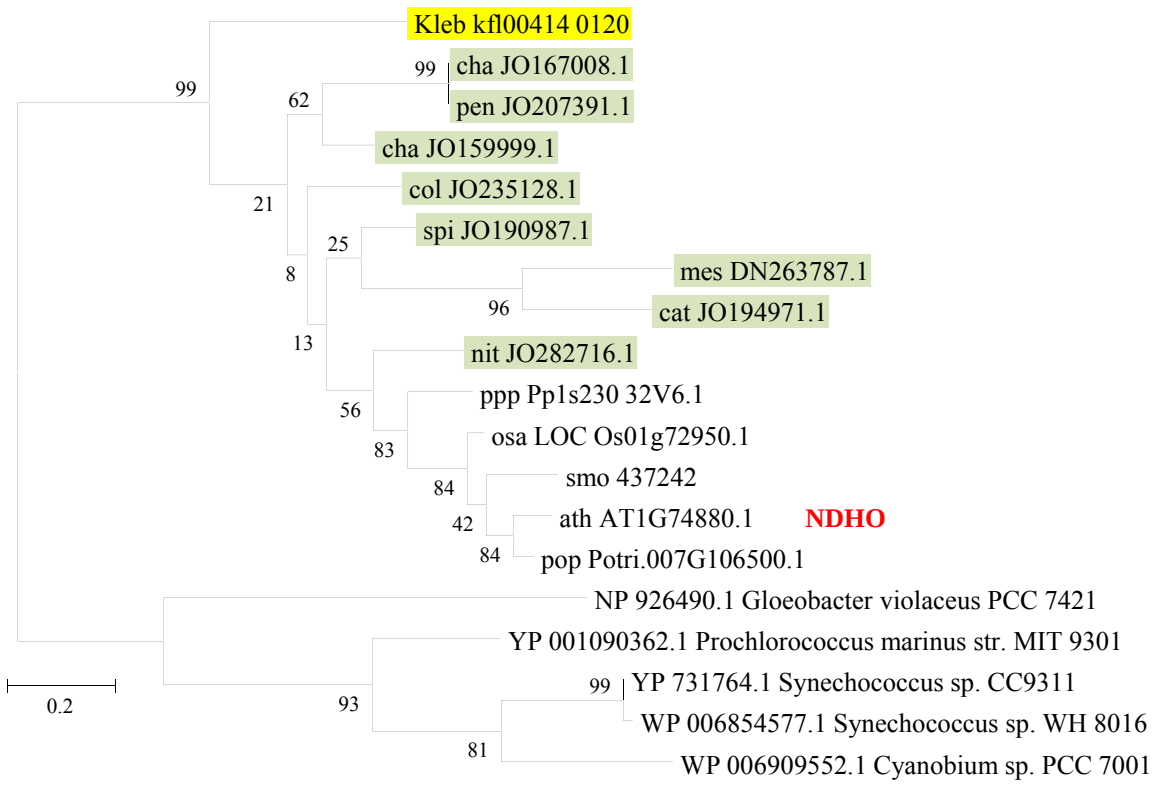
Supplementary Figure 52. Phylogenetic analysis of NDHL and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LGF+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



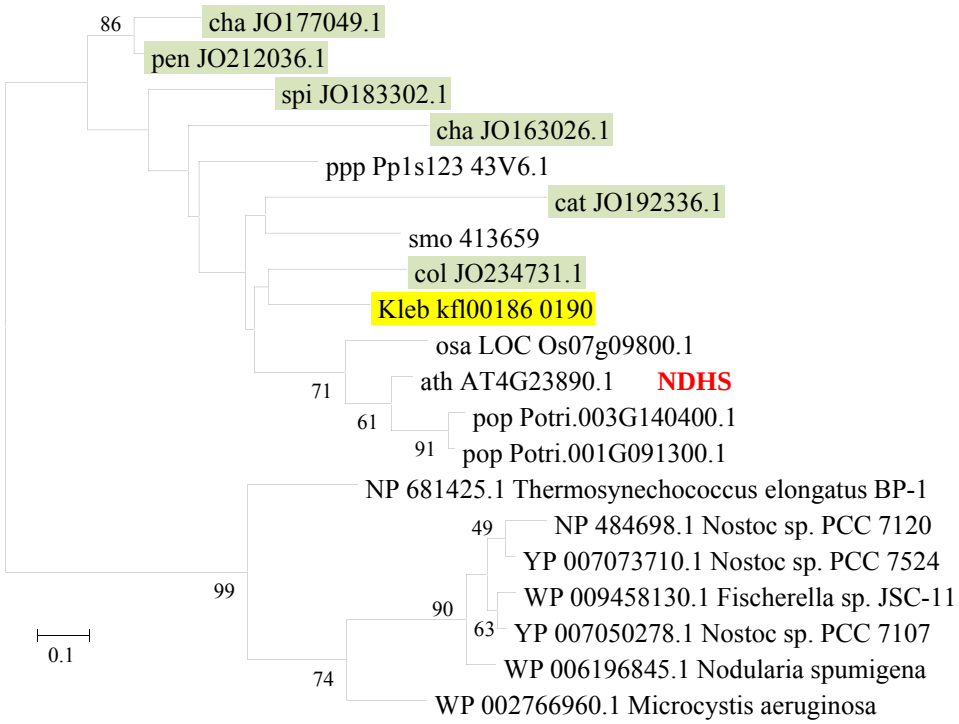
Supplementary Figure 53. Phylogenetic analysis of NDHM and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



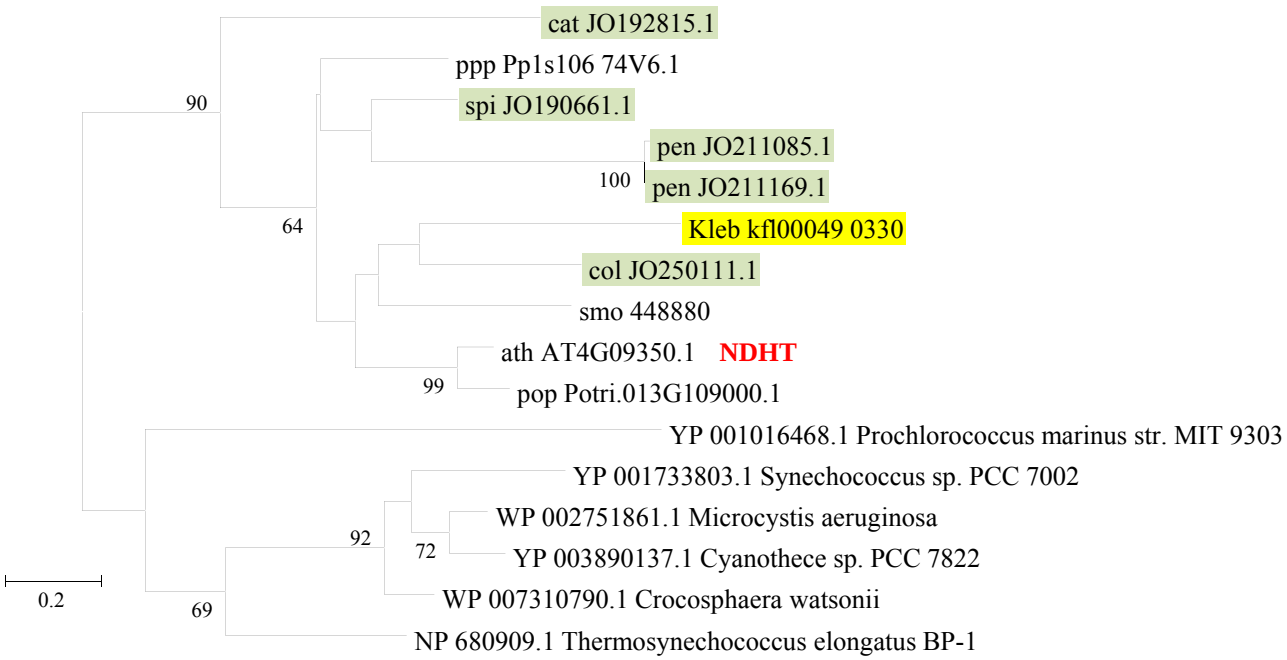
Supplementary Figure 54. Phylogenetic analysis of NDHN and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



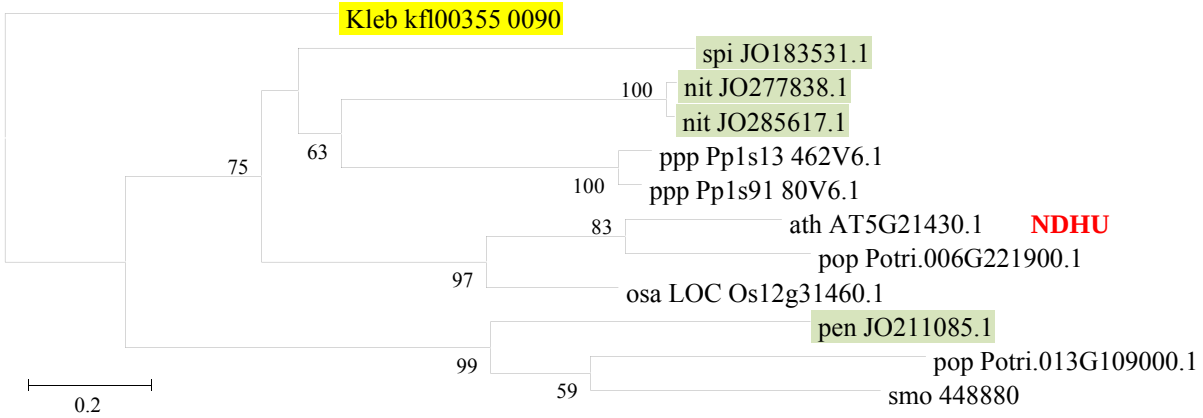
Supplementary Figure 55. Phylogenetic analysis of NDHO and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "WAG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



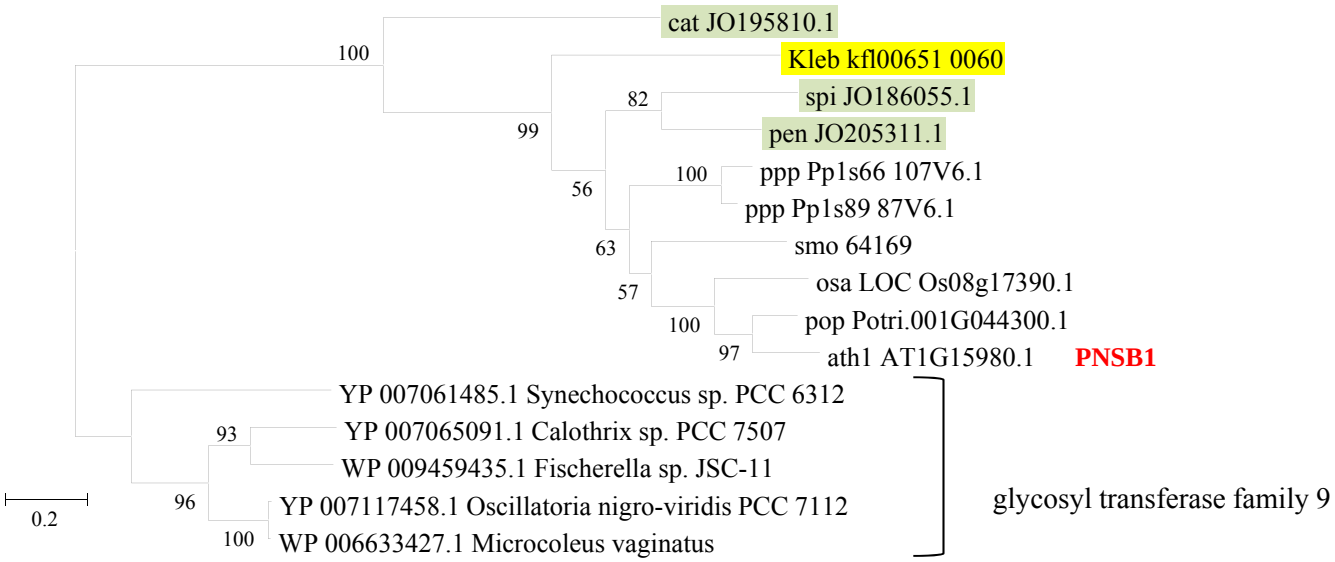
Supplementary Figure 56. Phylogenetic analysis of NDHS and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "WAG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



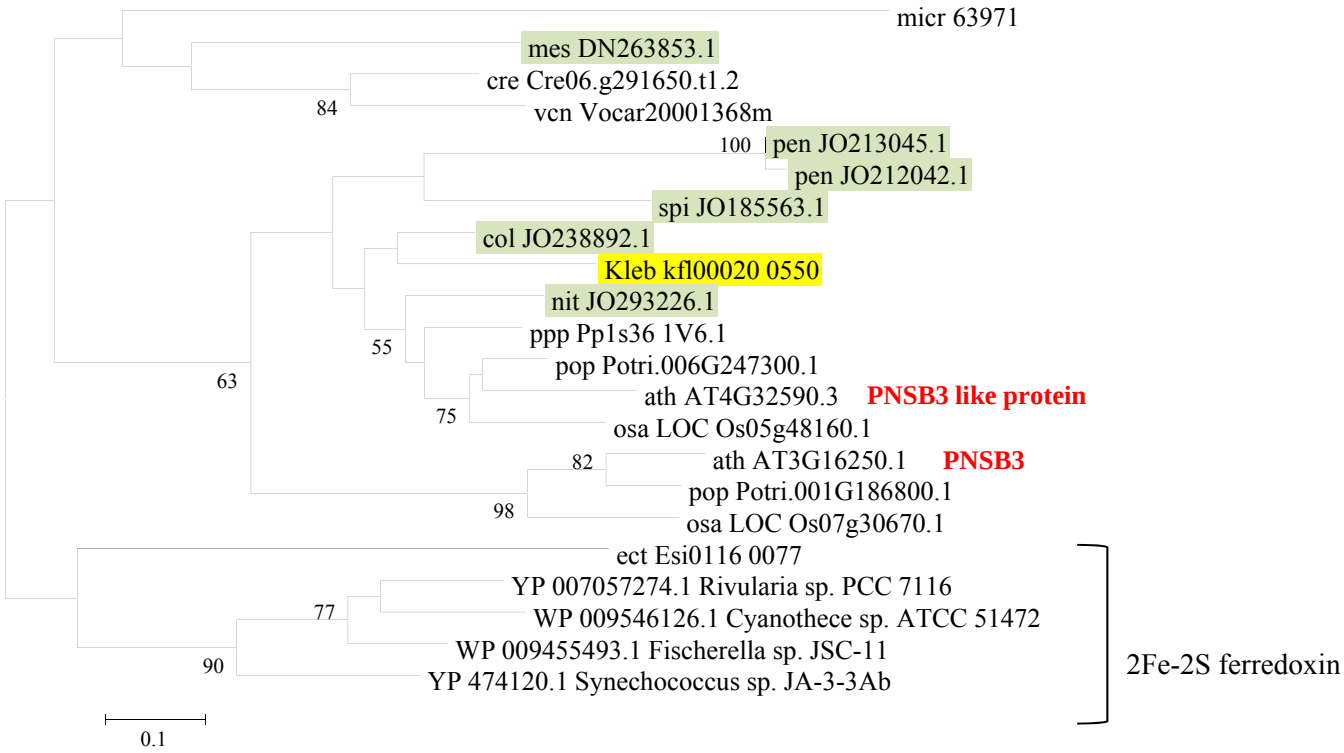
Supplementary Figure 57. Phylogenetic analysis of NDHT and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



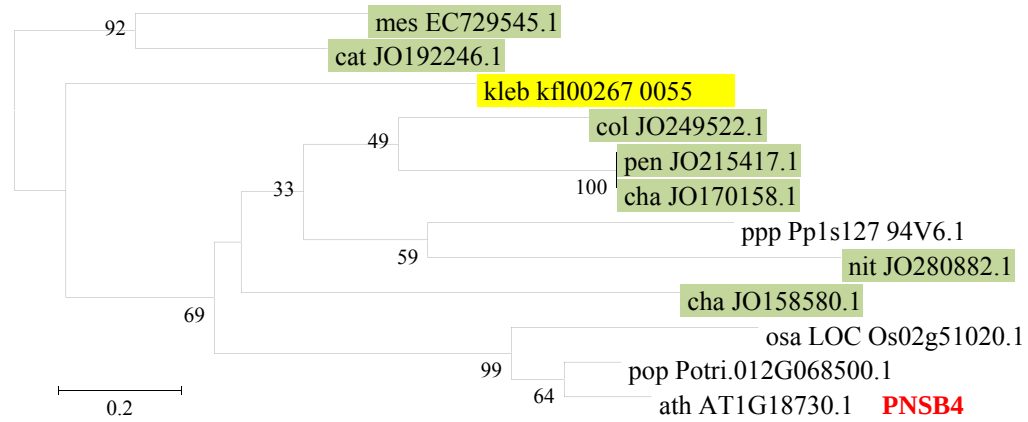
Supplementary Figure 58. Phylogenetic analysis of NDHU and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



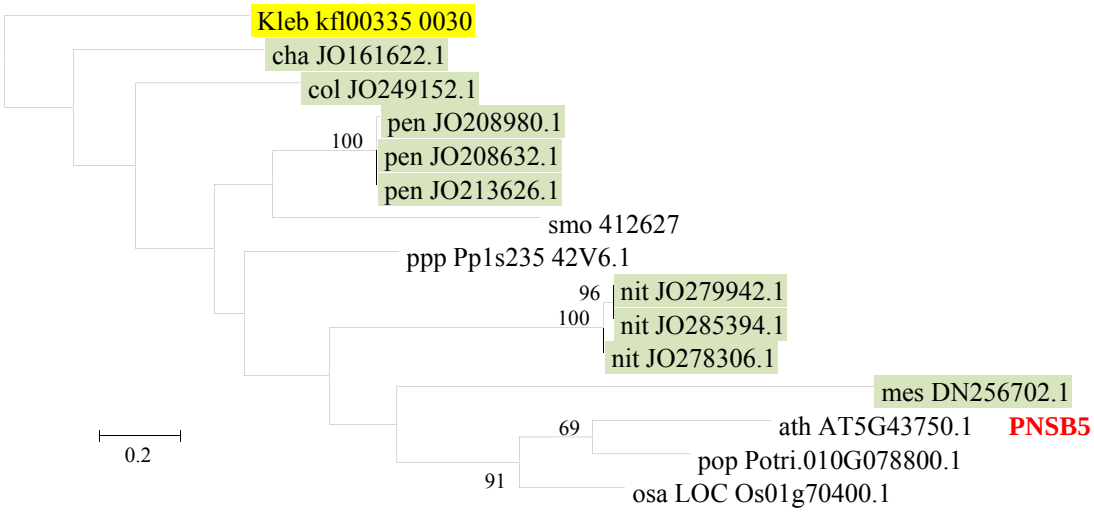
Supplementary Figure 59. Phylogenetic analysis of PNSB1 and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



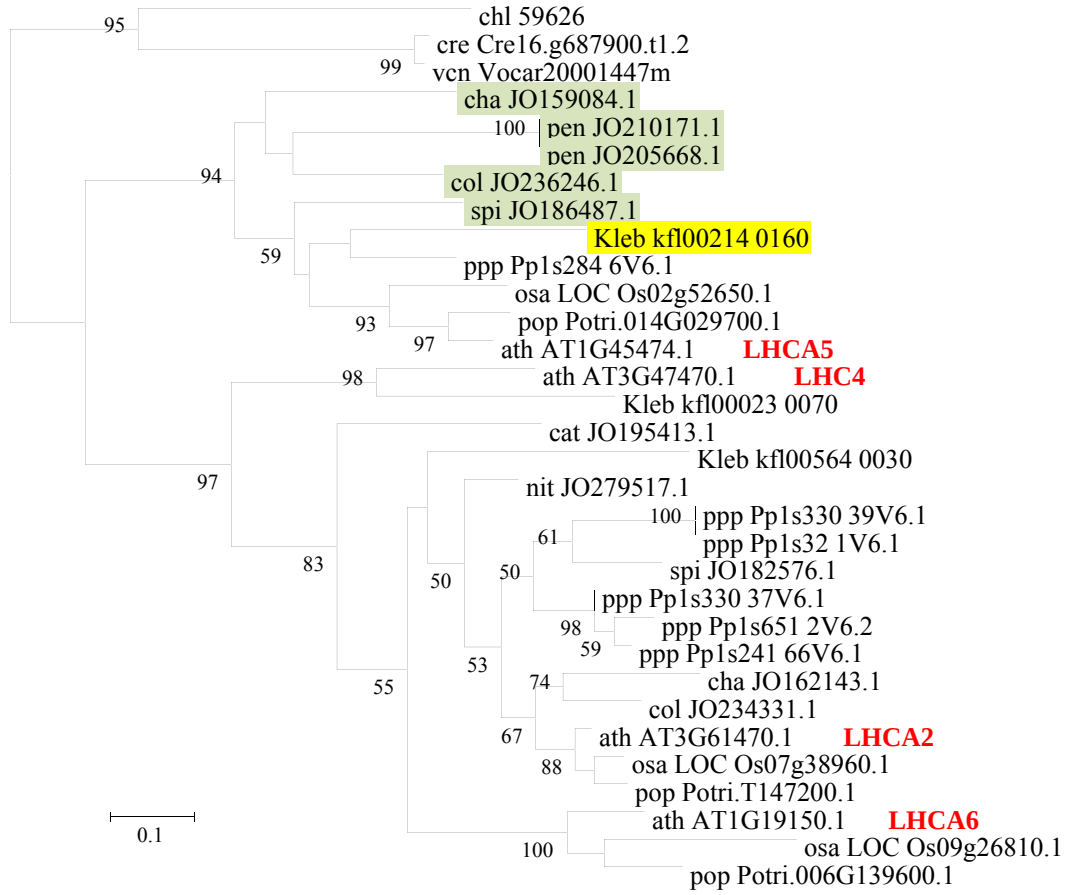
Supplementary Figure 60. Phylogenetic analysis of PNSB3 and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "WAFF+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



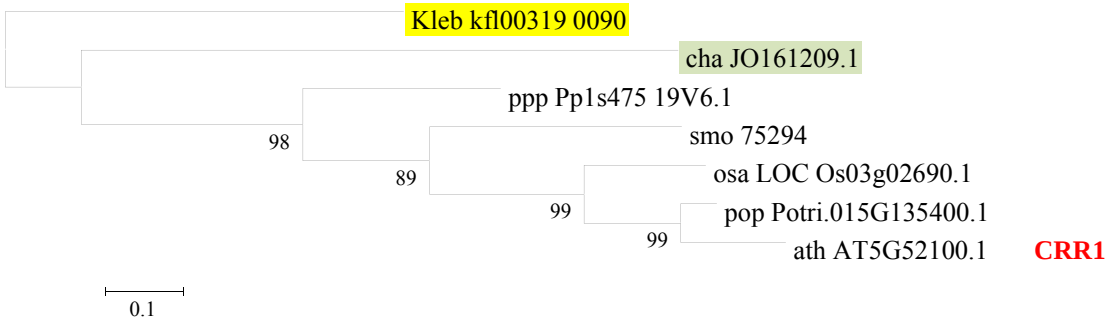
Supplementary Figure 61. Phylogenetic analysis of PNSB4 and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LGF+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



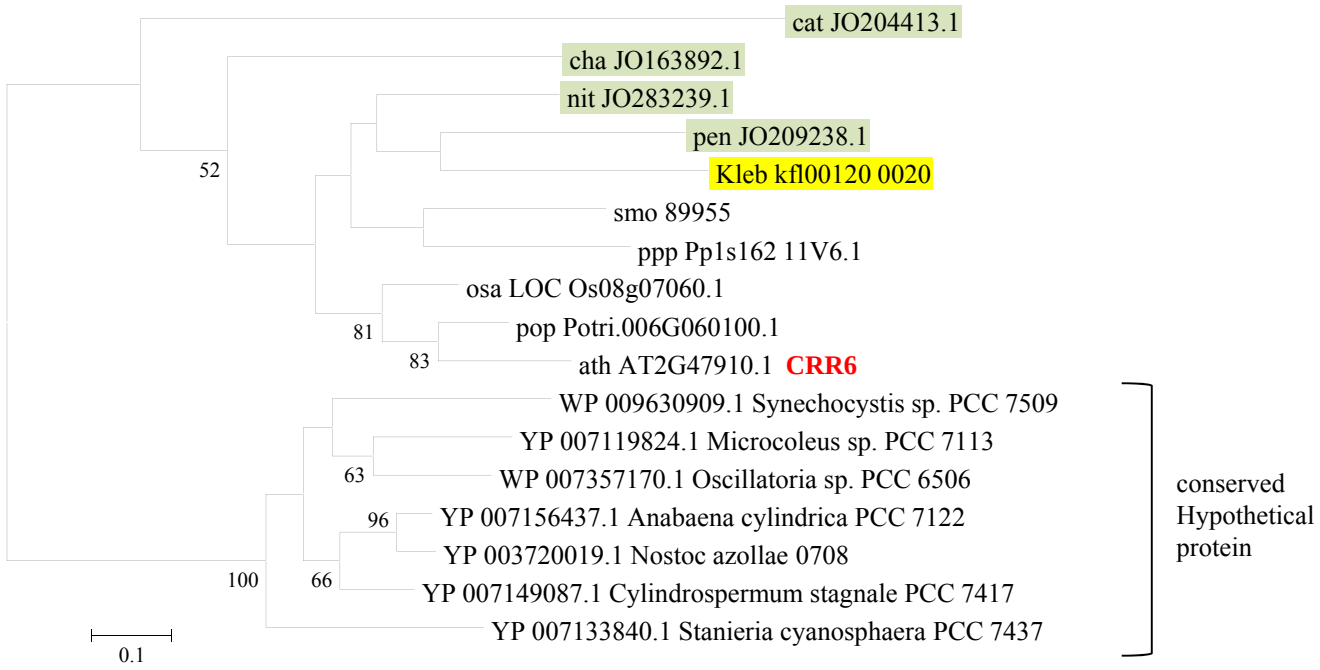
Supplementary Figure 62. Phylogenetic analysis of PNSB5 and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LGF+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



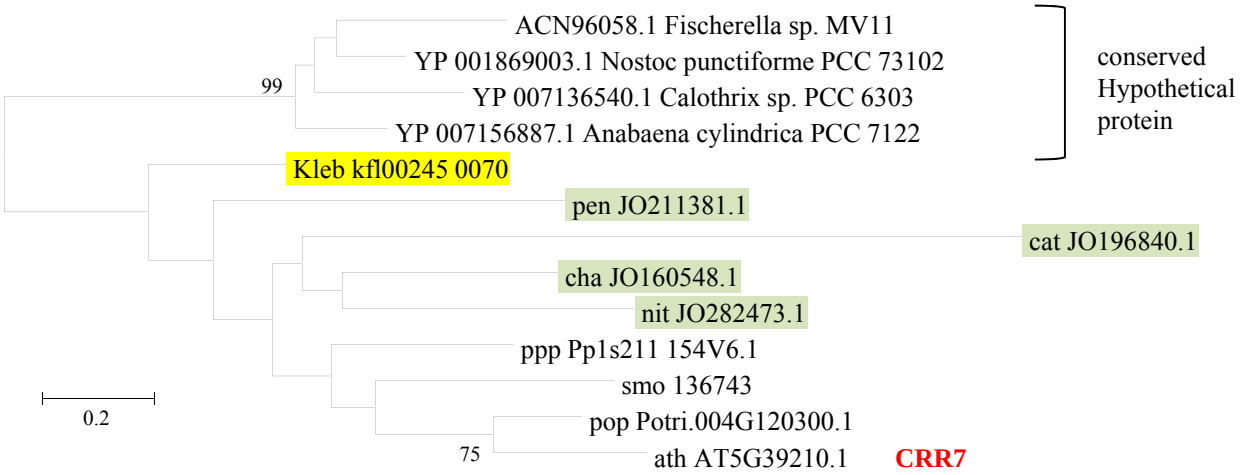
Supplementary Figure 63. Phylogenetic analysis of LHCA5 and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LGF+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



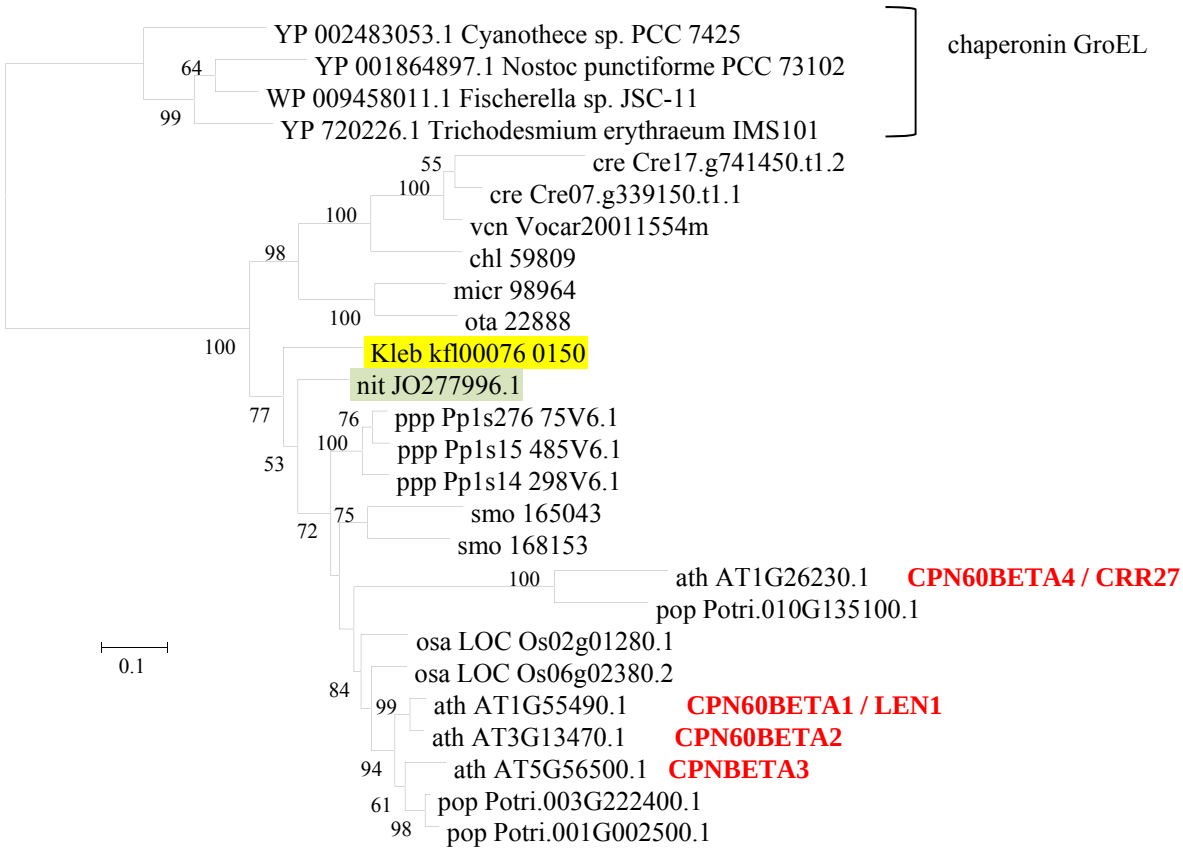
Supplementary Figure 64. Phylogenetic analysis of CRR1 and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



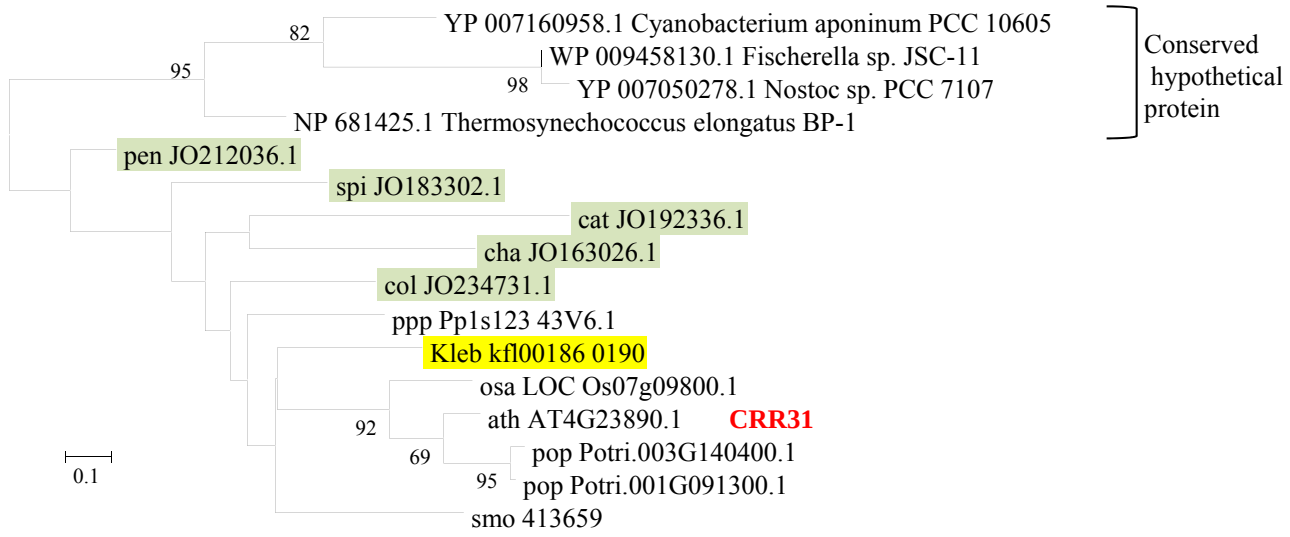
Supplementary Figure 65. Phylogenetic analysis of CRR6 and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "WAG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



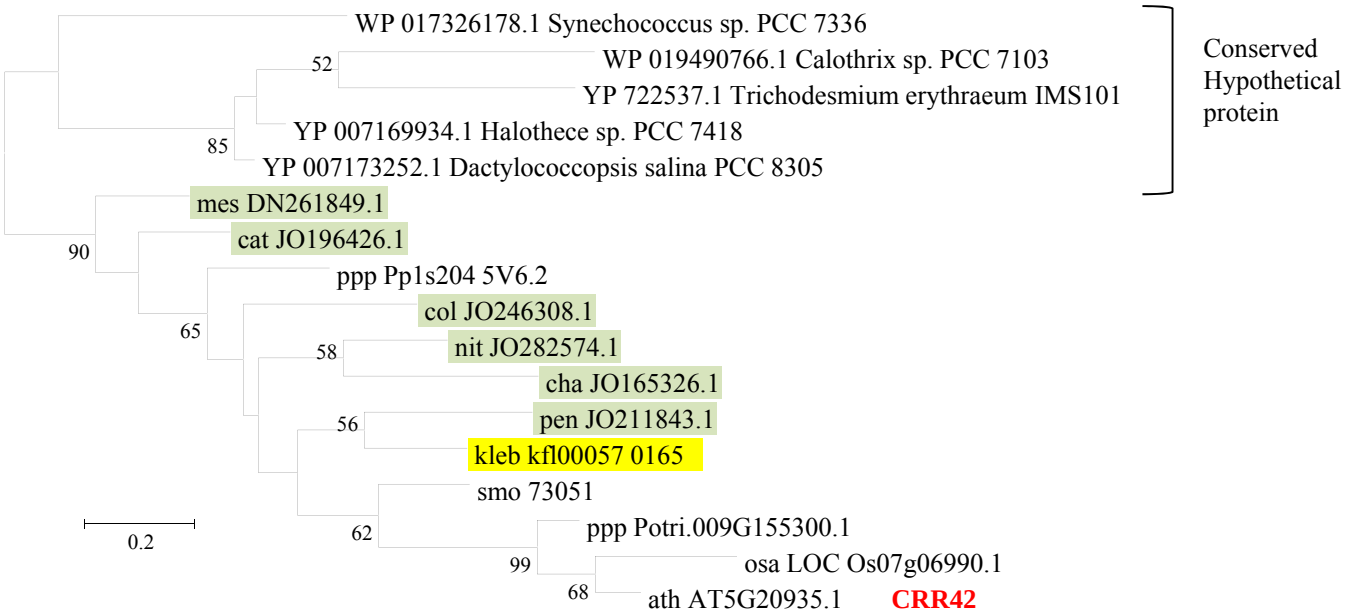
Supplementary Figure 66. Phylogenetic analysis of CRR7 and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



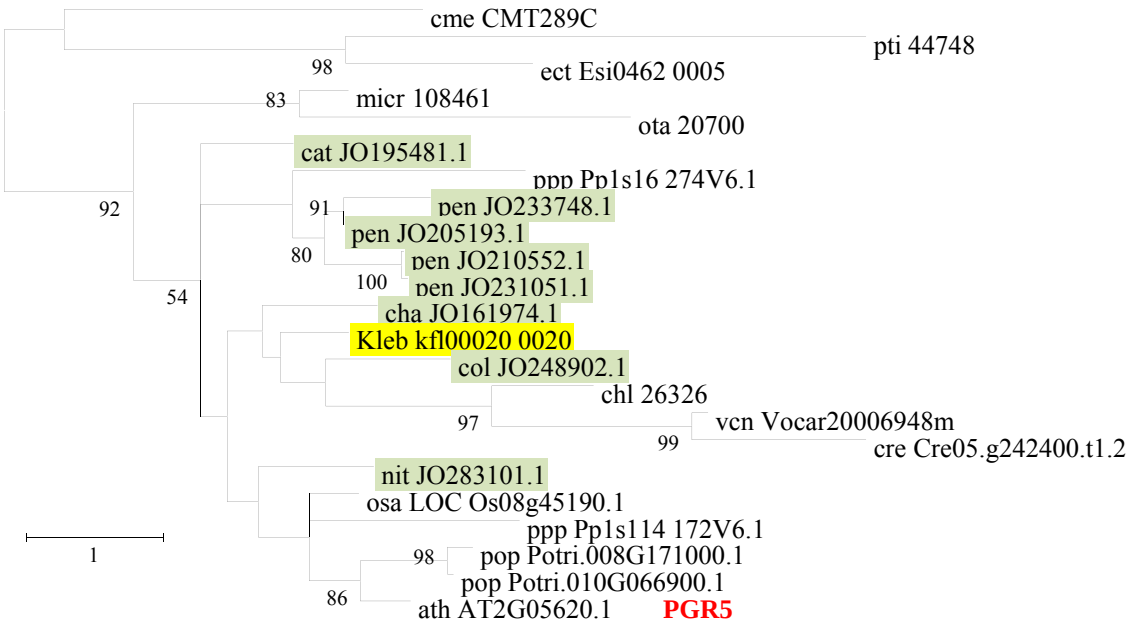
Supplementary Figure 67. Phylogenetic analysis of CRR27 and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



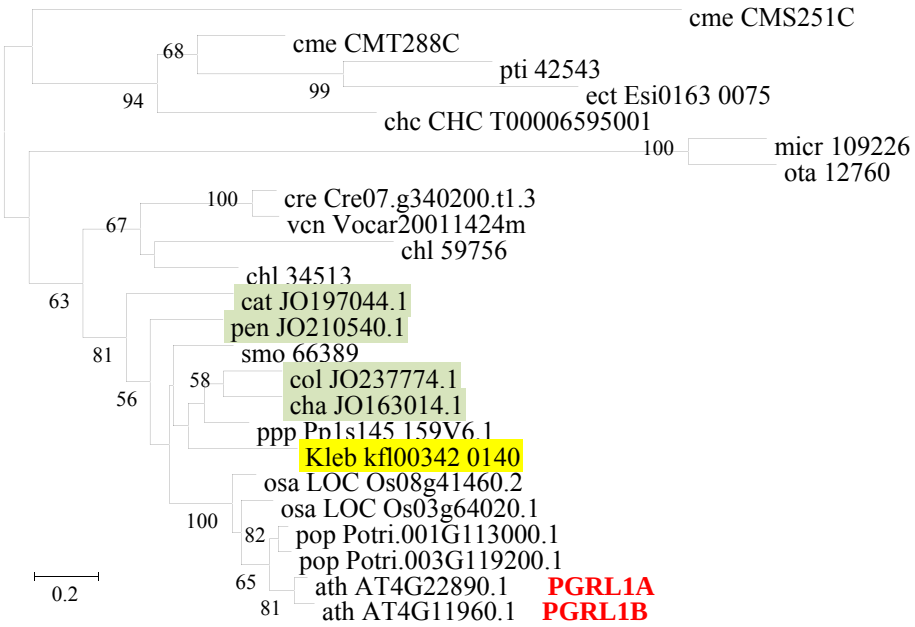
Supplementary Figure 68. Phylogenetic analysis of CRR31 and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



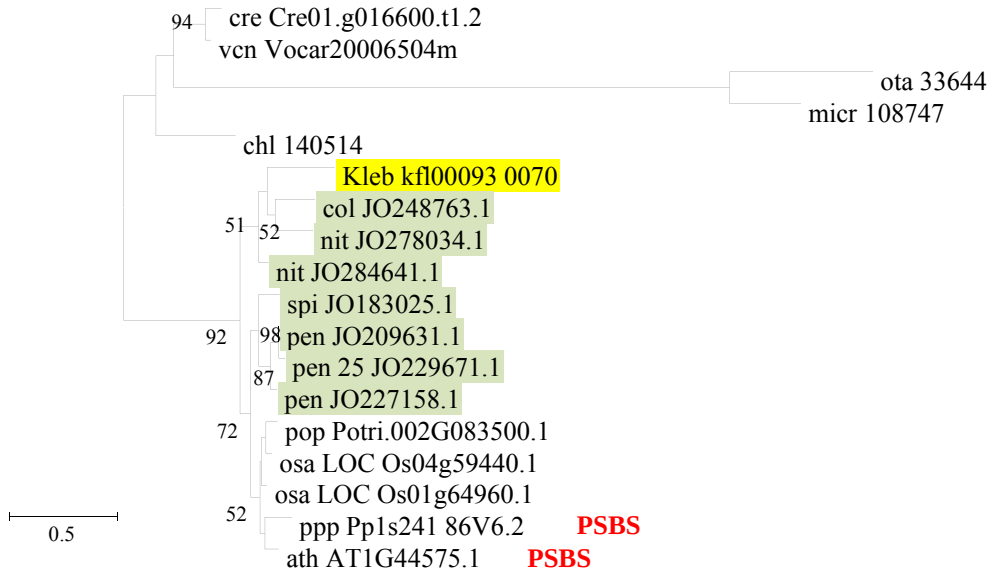
Supplementary Figure 69. Phylogenetic analysis of CRR42 and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



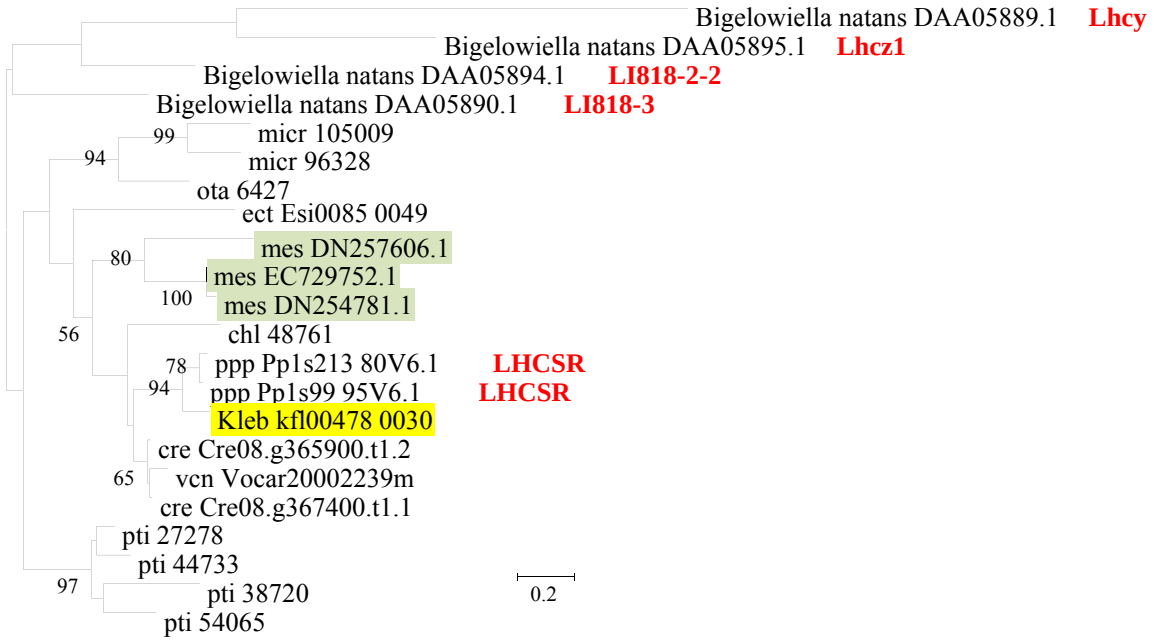
Supplementary Figure 70. Phylogenetic analysis of PGR5 and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



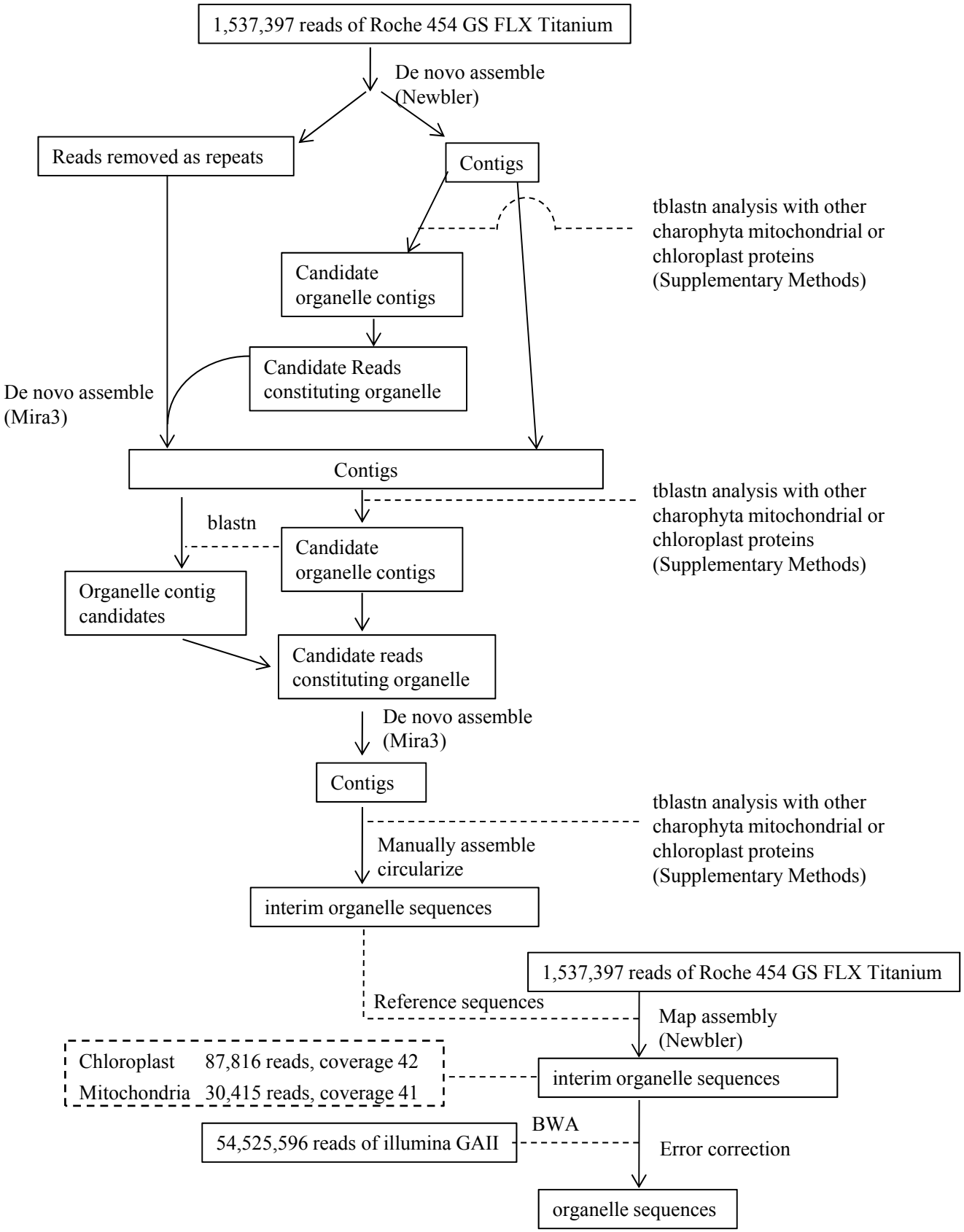
Supplementary Figure 71. Phylogenetic analysis of PGRL1 and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



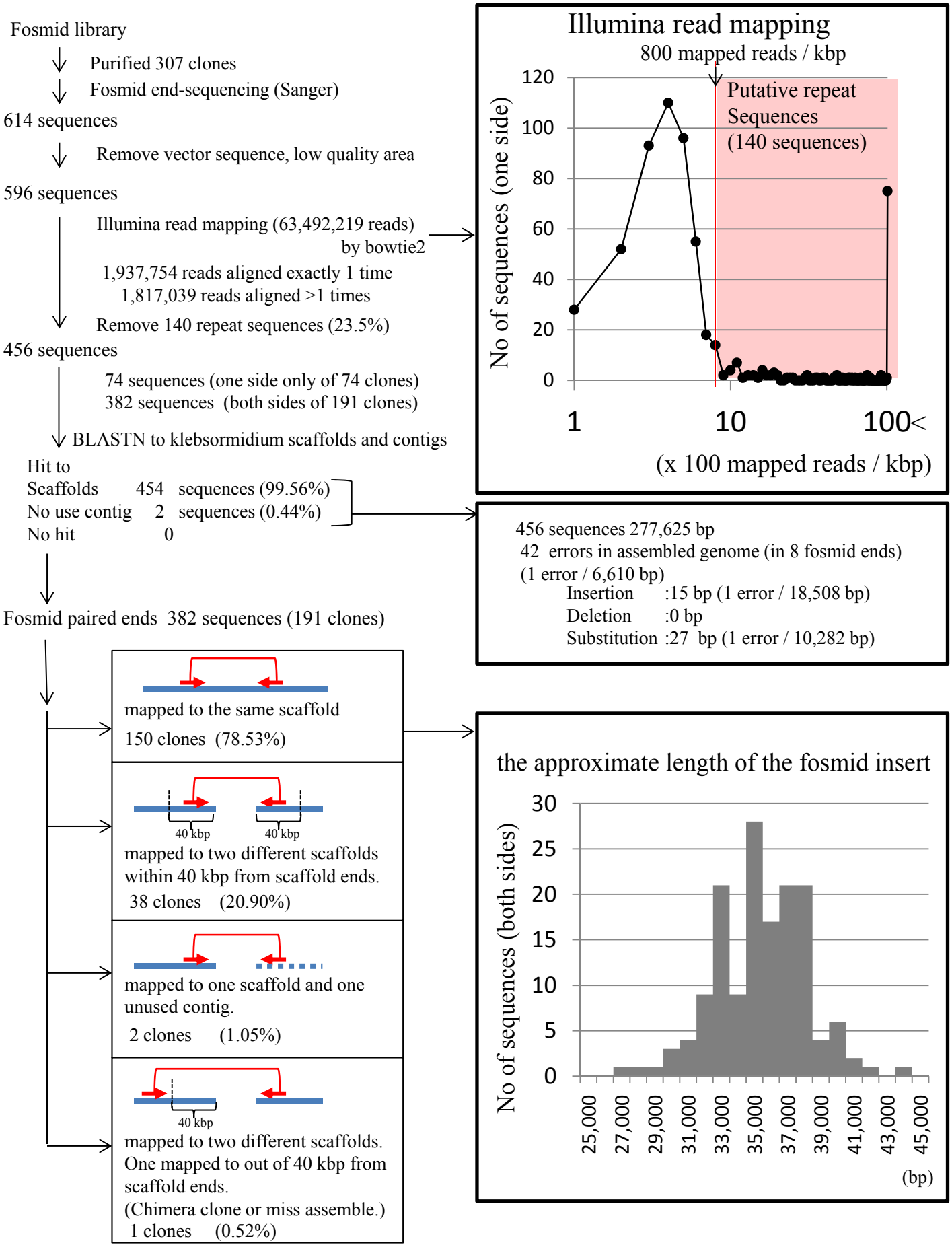
Supplementary Figure 72. Phylogenetic analysis of PSBS and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LGF+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



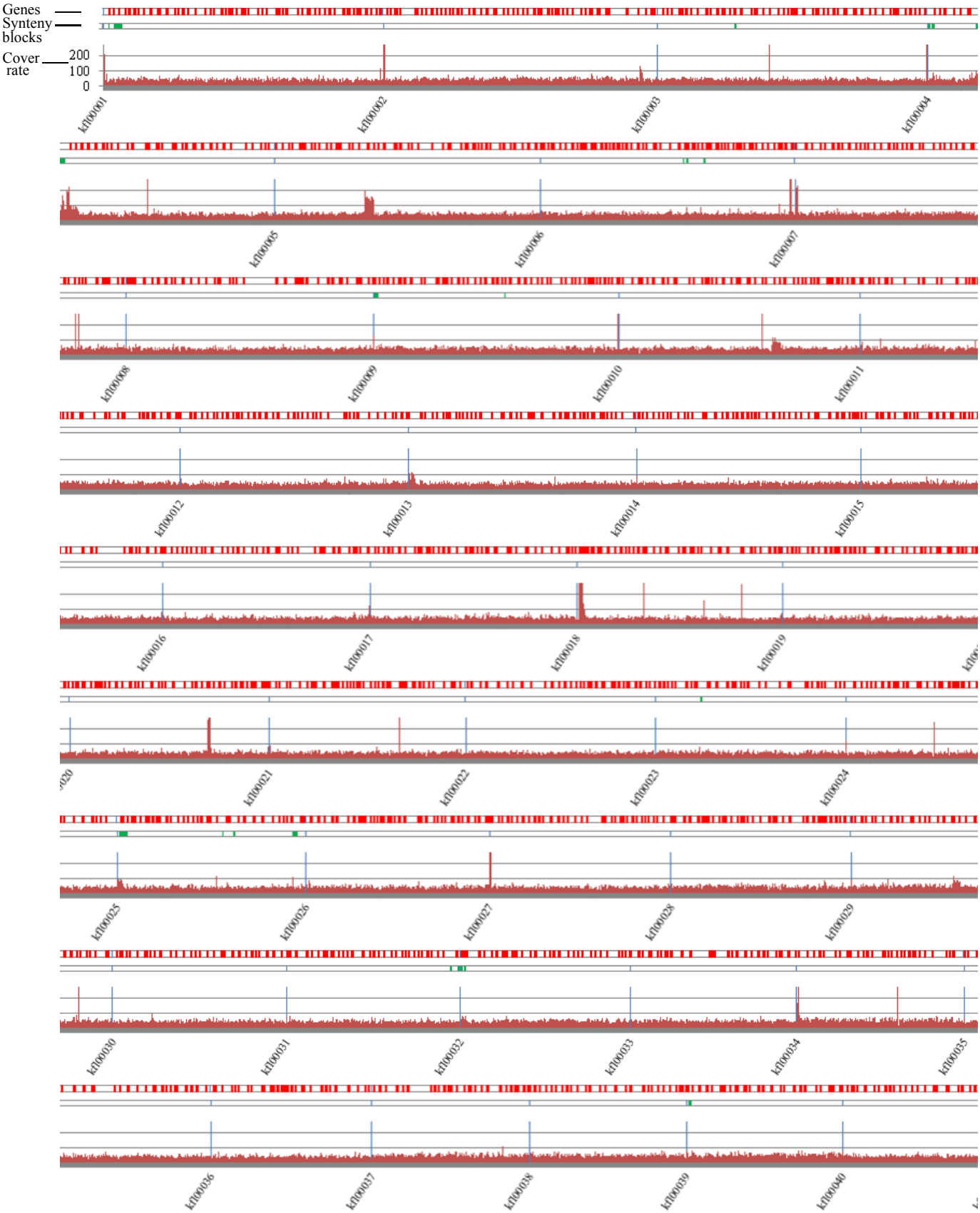
Supplementary Figure 73. Phylogenetic analysis of LHCSR and similar proteins of 15 spices and translated sequence of 6 charophyte ESTs. The amino acid substitution model was selected the optimal model of "LG+G". Phylogenetic analysis using maximum likelihood was performed in MEGA-CC ver 5.2 with 500 bootstraps. Bootstrap values higher than 50 are indicated under each branch. Red symbols indicate gene name. Candidate counterparts in *K. flaccidum* and other charophyta were indicated by yellow and green, respectively.



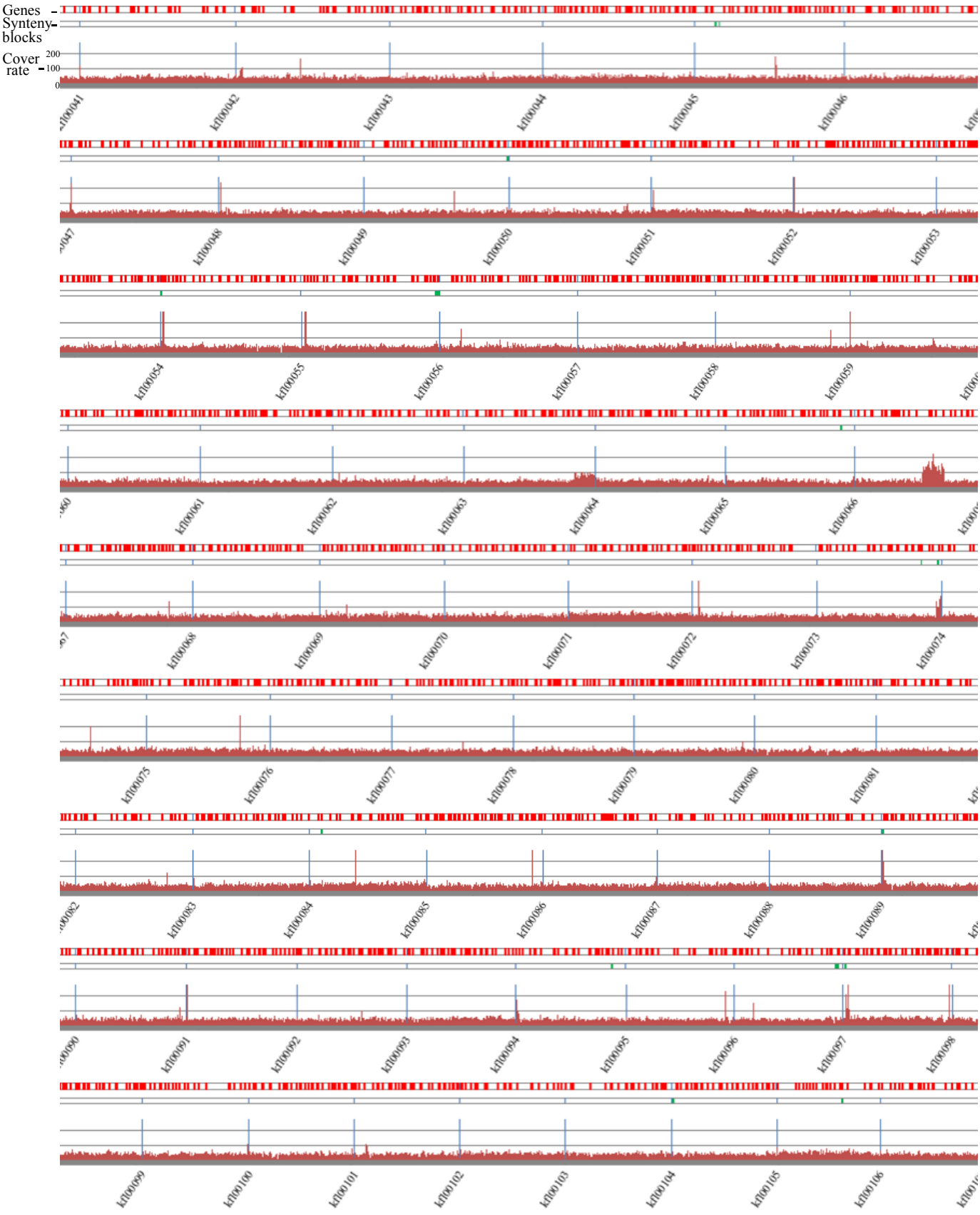
Supplementary Figure 74. Scheme used to assemble organellar genomes. Assembly statistics are shown in Supplementary table 1. The chloroplast genome and mitochondrial genome have been registered as scaffold1813 and scaffold1814, respectively.



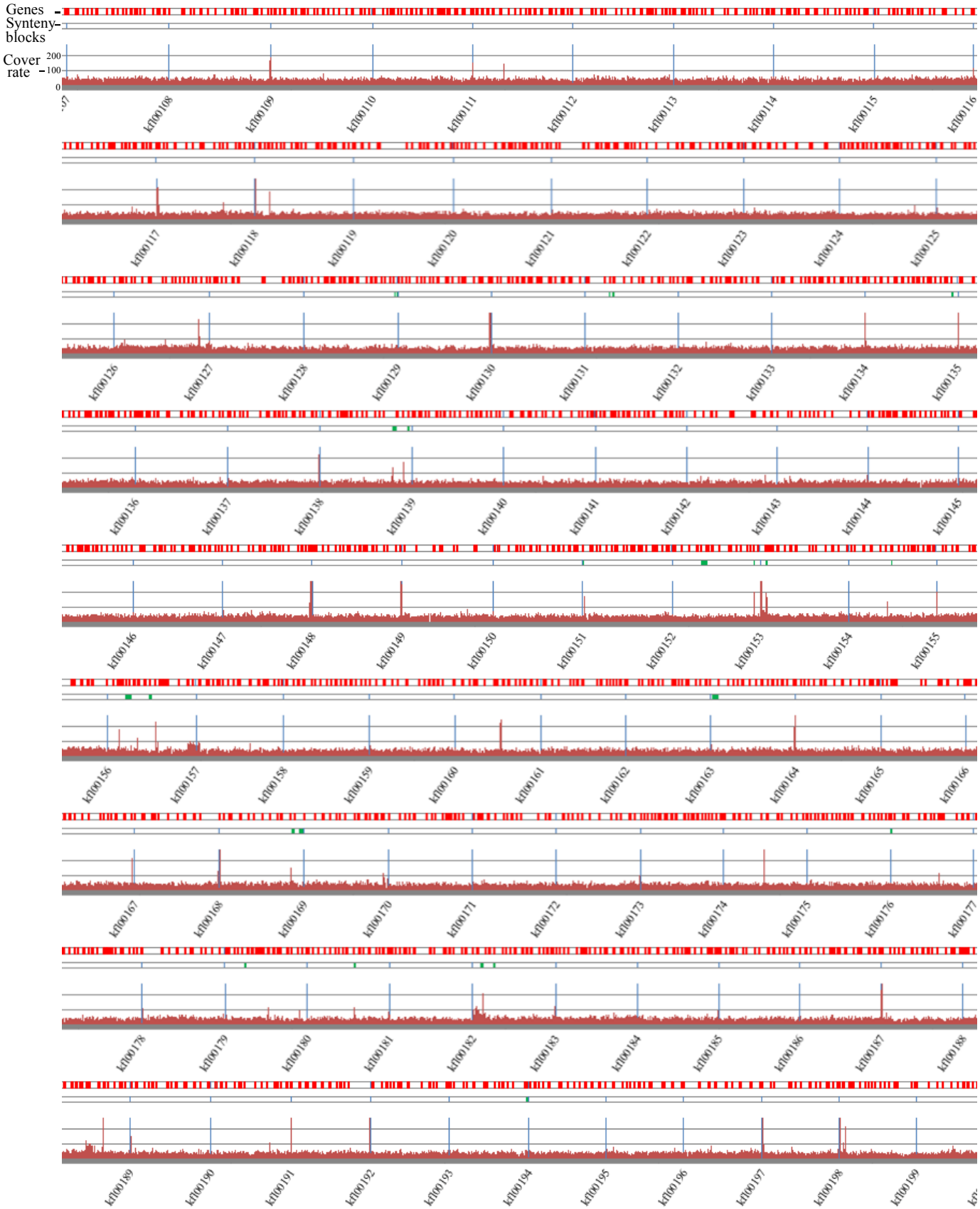
Supplementary Figure 75. Scheme of validation of sequencing and assembly of nuclear genome with fosmid library



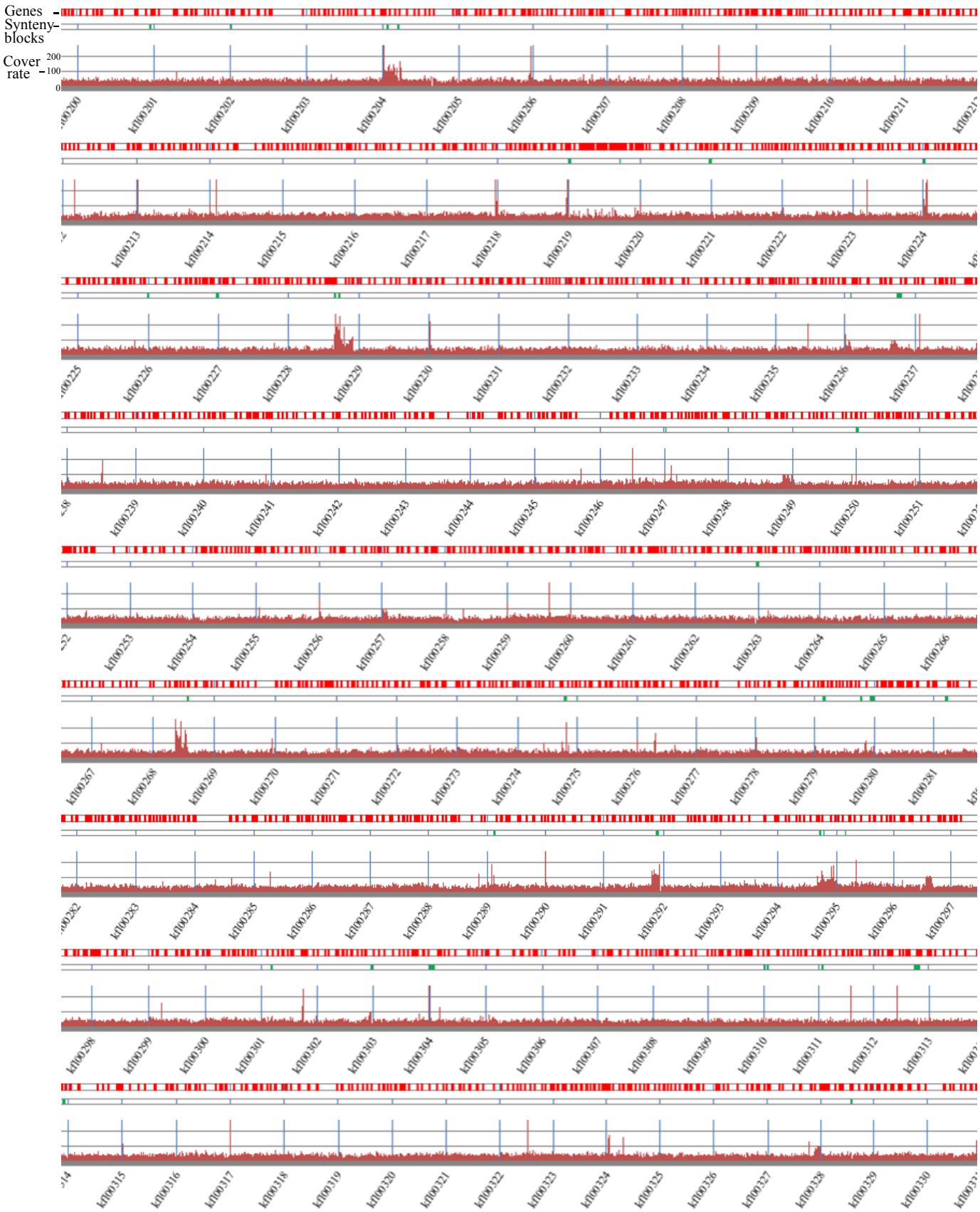
Supplementary Figure 76-1. Scaffold map (scaffold0001 - scaffold0040).
Genes line : Red dots indicate positions of start codon of predicted gene. Synteny blocks line : Green dots indicate synteny blocks area (≥ 1k bp).
Cover rate line : Coverage plots of remapping with 454 reads were shown (brown). Each scaffold was separated by blue bar.



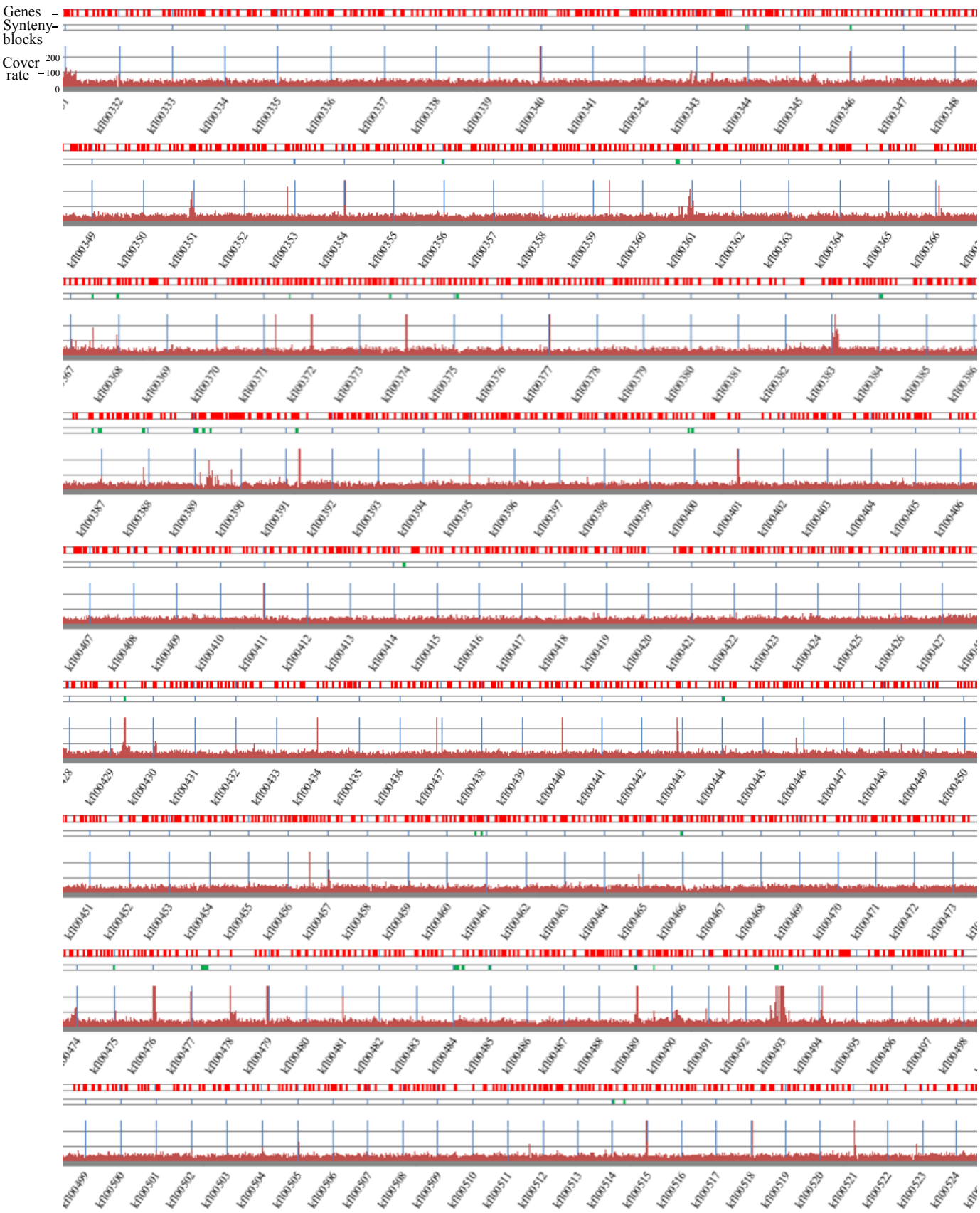
Supplementary Figure 76-2. Scaffold map (scaffold0040 - scaffold0106).
Genes line : Red dots indicate positions of start codon of predicted gene. Synteny blocks line : Green dots indicate synteny blocks area (≥ 1k bp).
Cover rate line : Coverage plots of remapping with 454 reads were shown (brown). Each scaffold was separated by blue bar.



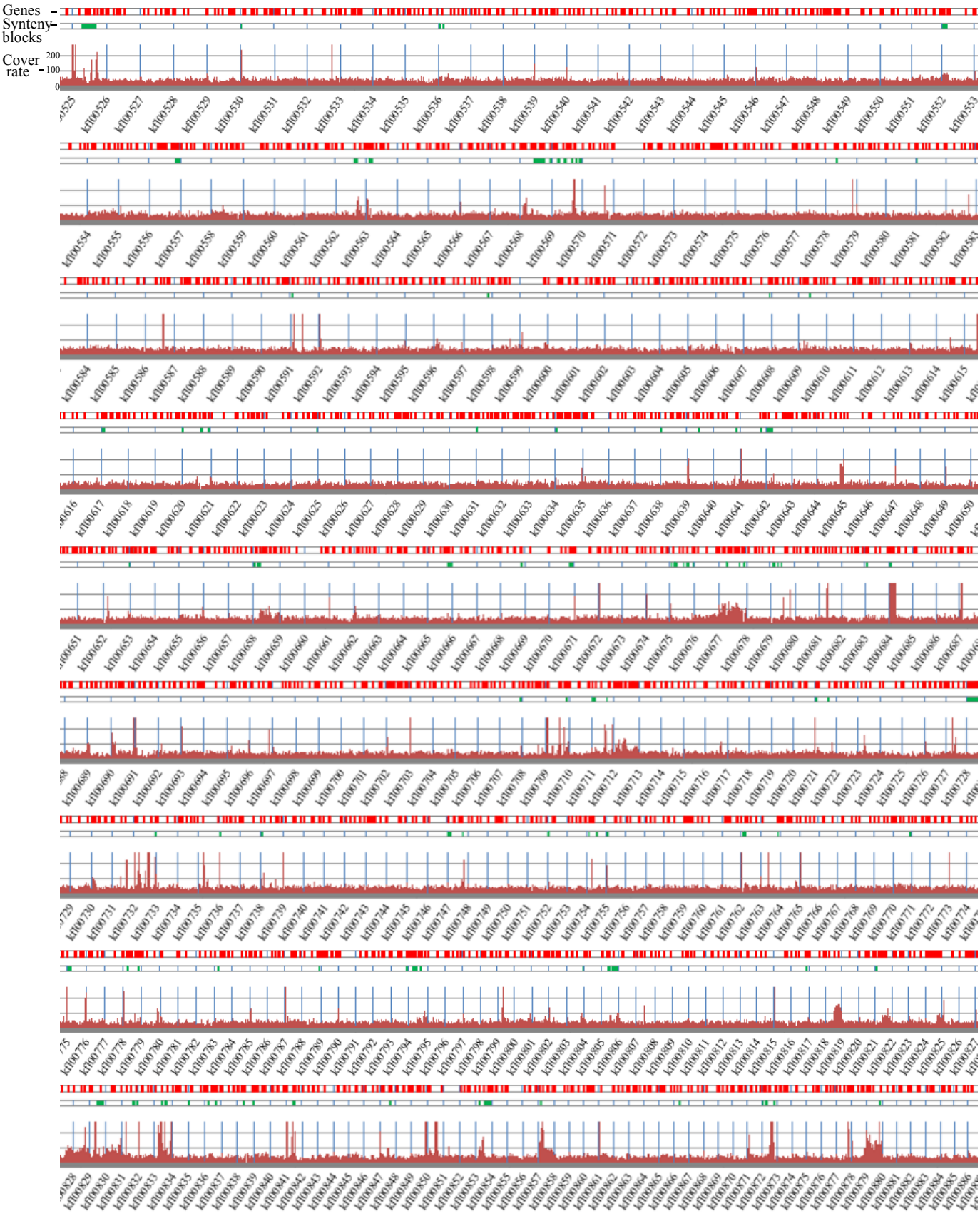
Supplementary Figure 76-3. Scaffold map (scaffold0106 - scaffold0199).
Genes line : Red dots indicate positions of start codon of predicted gene. Synteny blocks line : Green dots indicate synteny blocks area (≥ 1k bp).
Cover rate line : Coverage plots of remapping with 454 reads were shown (brown). Each scaffold was separated by blue bar.



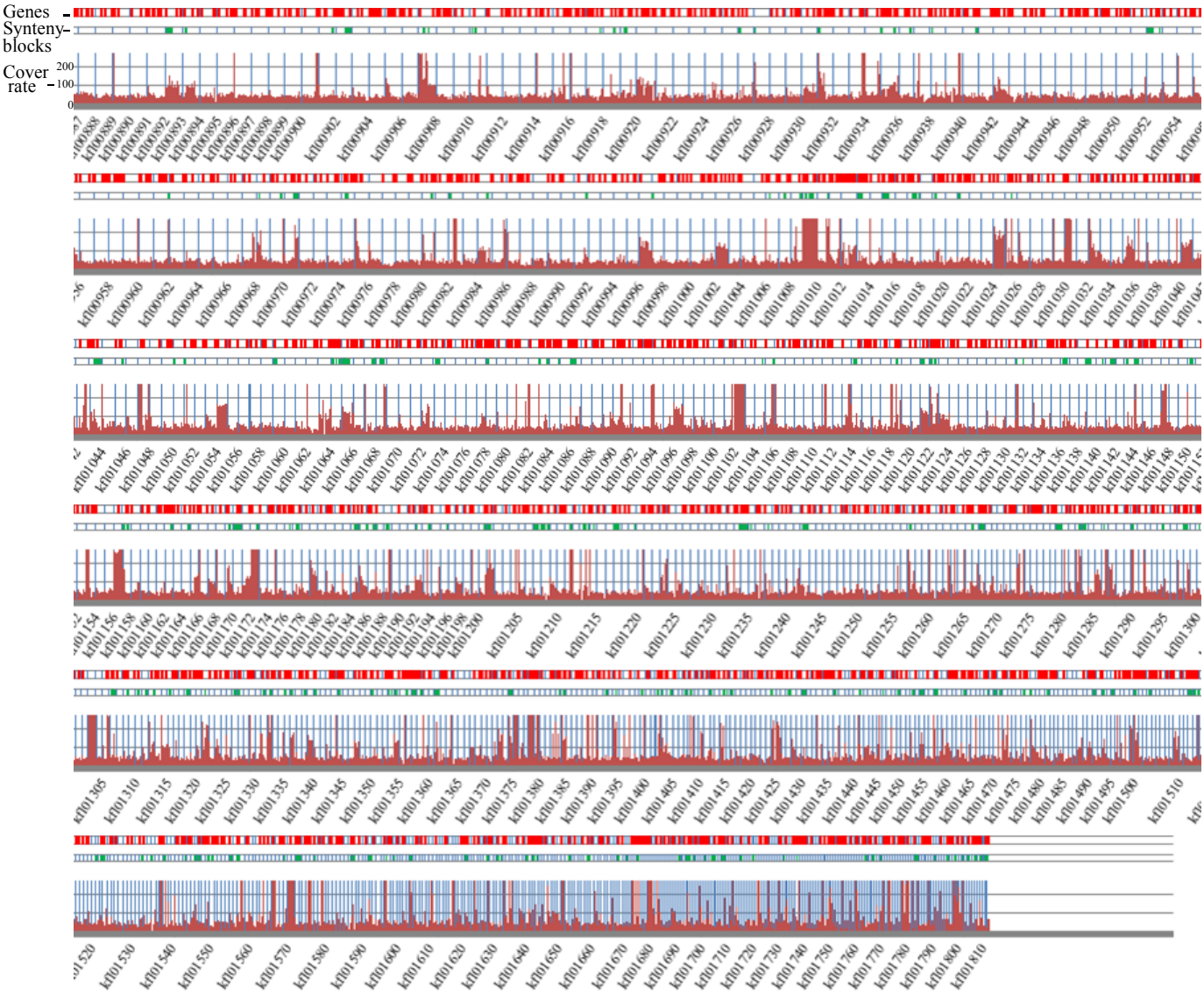
Supplementary Figure 76-4. Scaffold map (scaffold0199 - scaffold0330).
Genes line : Red dots indicate positions of start codon of predicted gene. Synteny blocks line : Green dots indicate synteny blocks area (≥ 1k bp).
Cover rate line : Coverage plots of remapping with 454 reads were shown (brown). Each scaffold was separated by blue bar.



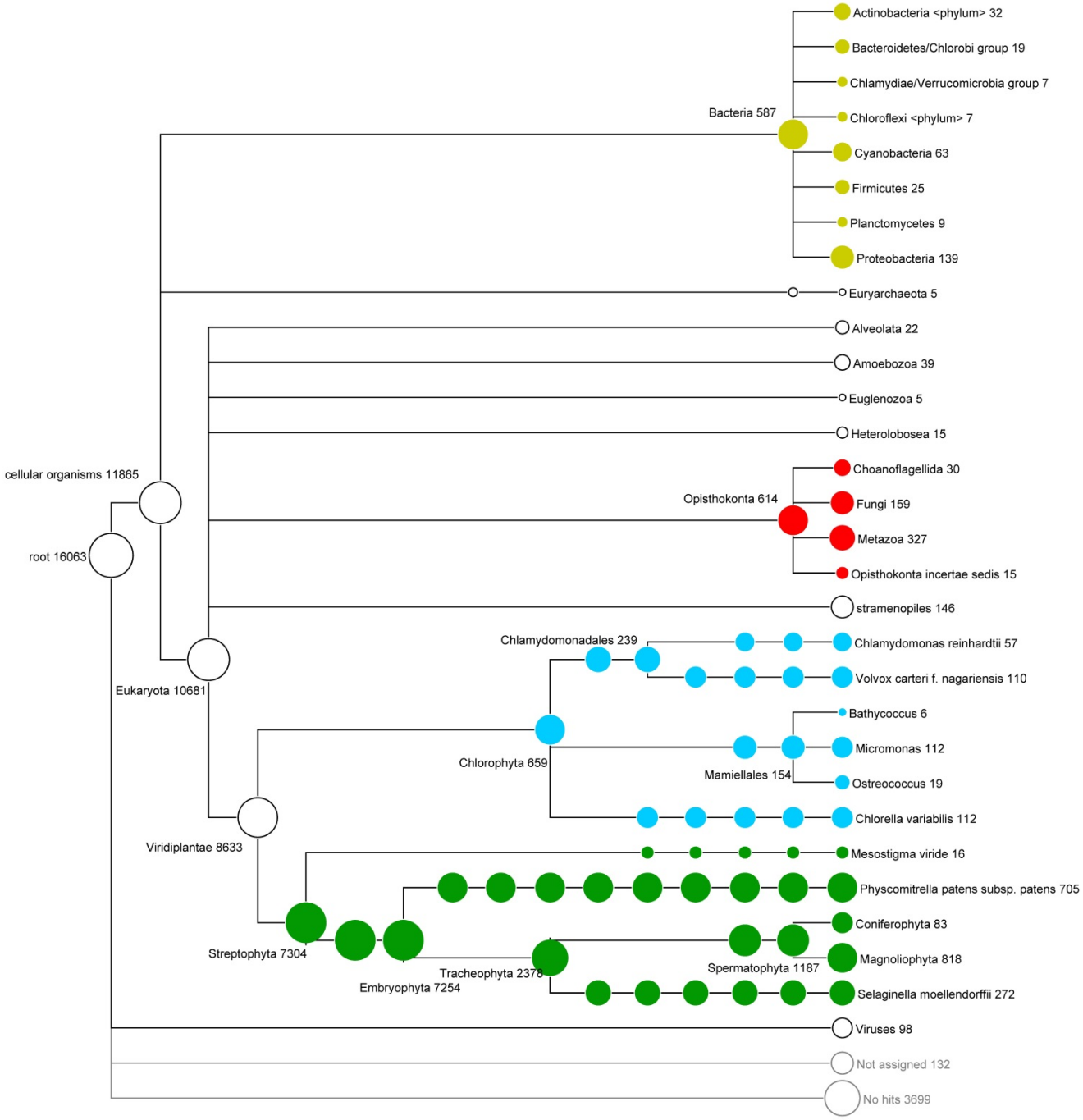
Supplementary Figure 76-5. Scaffold map (scaffold0330 - scaffold0524).
Genes line : Red dots indicate positions of start codon of predicted gene. Synteny blocks line : Green dots indicate synteny blocks area (≥ 1k bp).
Cover rate line : Coverage plots of remapping with 454 reads were shown (brown). Each scaffold was separated by blue bar.



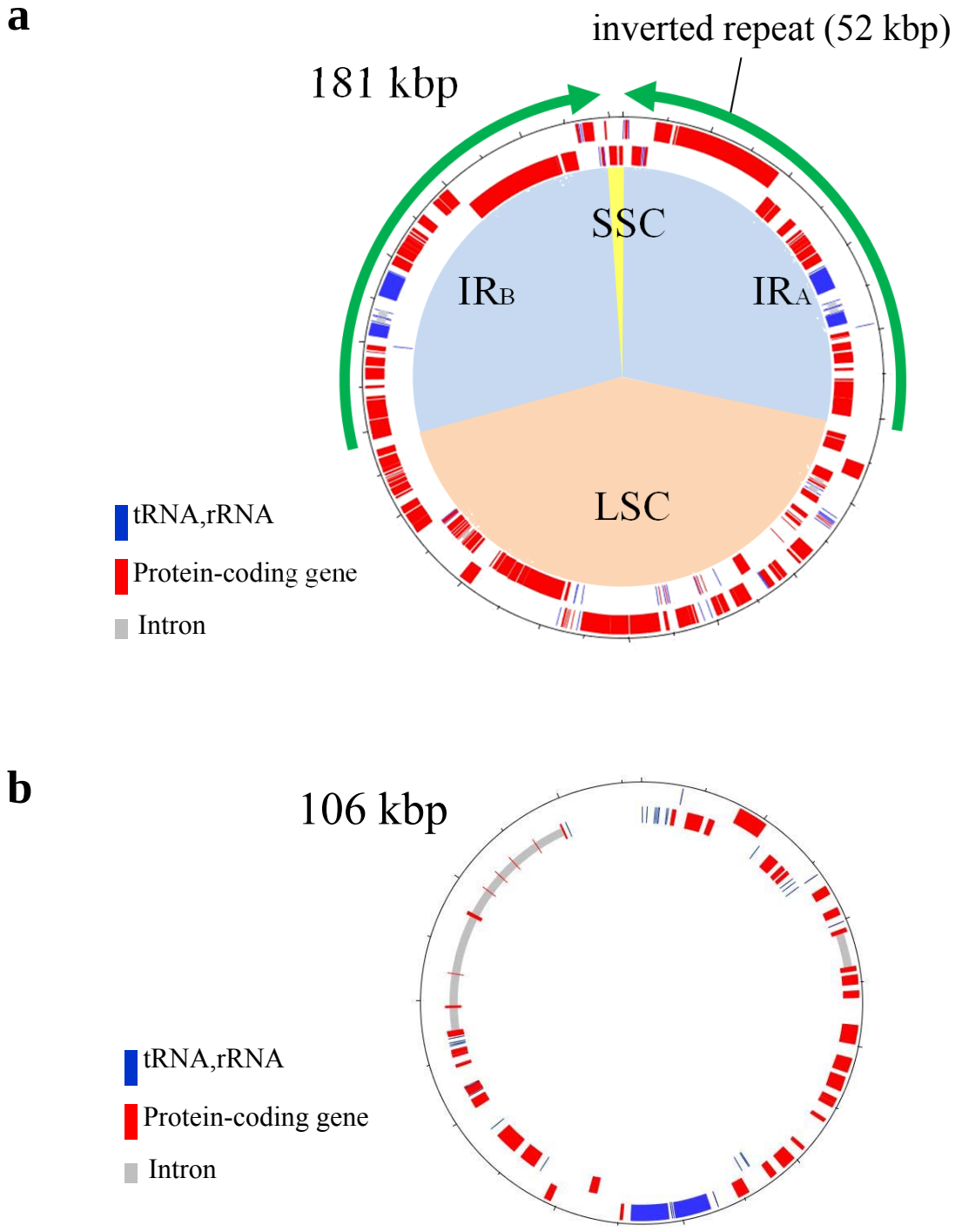
Supplementary Figure 76-6. Scaffold map (scaffold0524 - scaffold0886).
Genes line : Red dots indicate positions of start codon of predicted gene. Synteny blocks line : Green dots indicate synteny blocks area (≥ 1k bp).
Cover rate line : Coverage plots of remapping with 454 reads were shown (brown). Each scaffold was separated by blue bar.



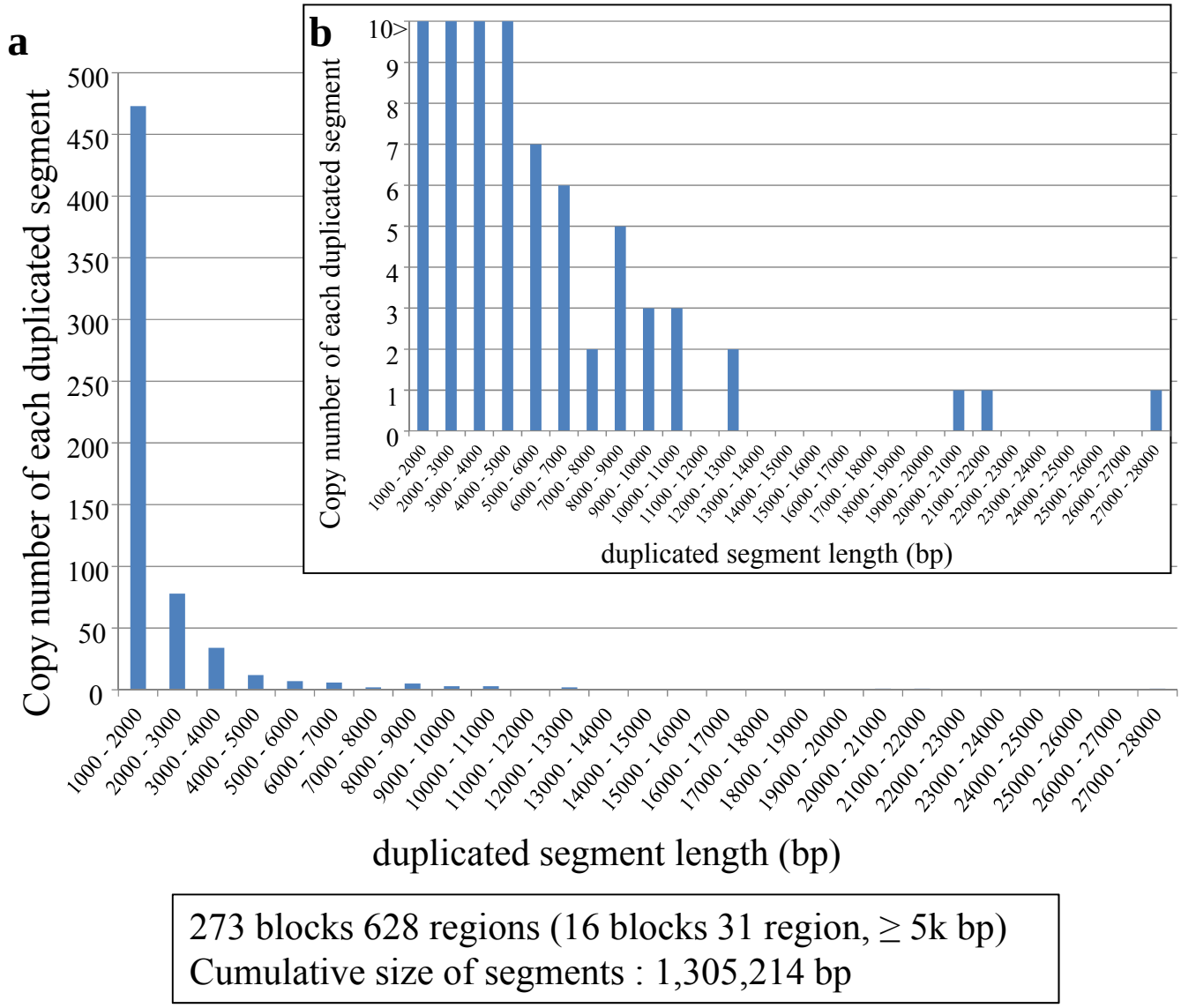
Supplementary Figure 76-7. Scaffold map (scaffold0886 – scaffold1812).
Genes line : Red dots indicate positions of start codon of predicted gene. Synteny blocks line : Green dots indicate synteny blocks area (≥ 1k bp).
Cover rate line : Coverage plots of remapping with 454 reads were shown (brown). Each scaffold was separated by blue bar.



Supplementary Figure 77. MEGAN analysis of 16,063 nuclear genes based on a BLASTP with NCBI-NR are presented in the comparative tree views of MEGAN.
The number of sequences assigned to the corresponding were summarized number of nodes.

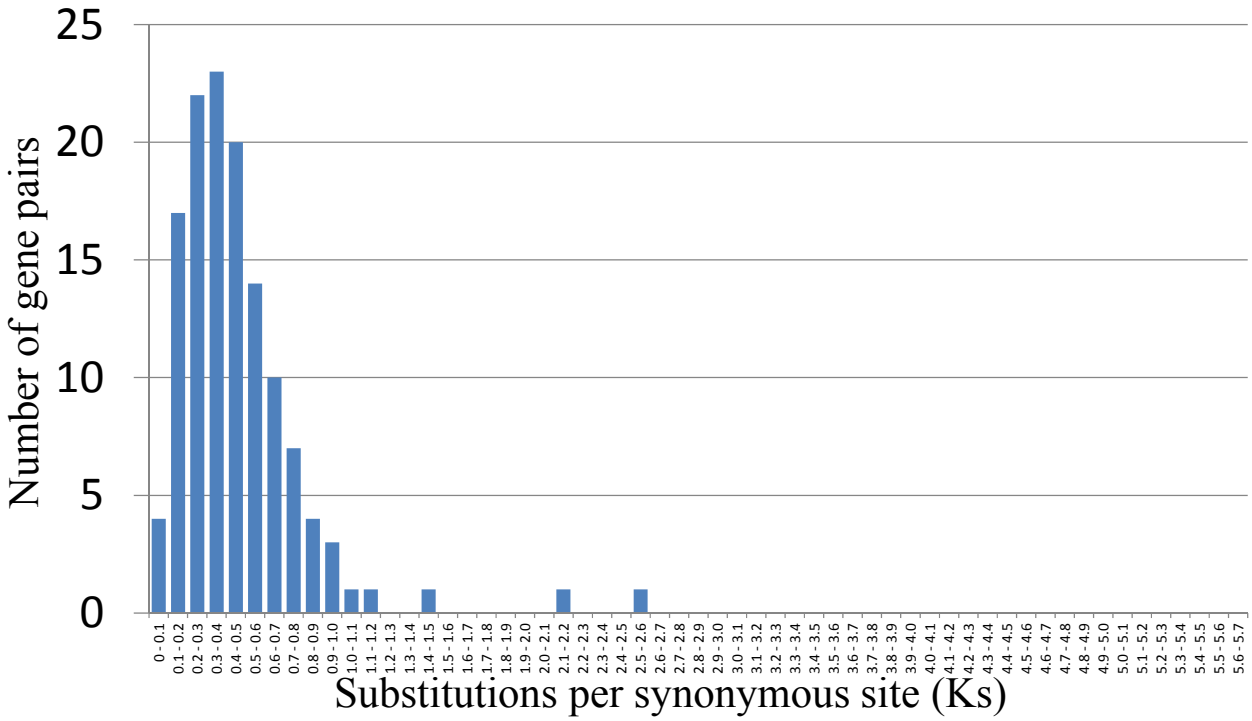


Supplementary Figure 78. Circular maps for organellar genomes. Protein-coding genes (red), tRNAs and rRNAs (blue), and introns (grey) are indicated. (a) Map of the chloroplast genome. Green arrows indicate inverted repeat regions. (b) Map of the mitochondrial genome.



Supplementary Figure 79. Size distribution of duplicated chromosome segment copy number. (**a**) The copy number of each duplicated chromosome segment is reported on the y-axis and the corresponding cumulative size is reported on x-axis. (b) Low copy number region (<10) of panel (a) is magnified.

a



b

K. flaccidum gene ID	Spices	Syntenic paralog or ortholog gene ID	Ks
kfl01110_0020	*K. flaccidum*	kfl00106_0020	0.606
kfl01110_0020	*Coleochaete sp.*	JO249077.1	4.893
kfl00215_0220	*K. flaccidum*	kfl00215_0170	0.298
kfl00215_0220	*Coleochaete sp.*	JO248949.1	4.855
kfl00394_0030	*K. flaccidum*	kfl00240_0210	0.652
kfl00394_0030	*K. flaccidum*	kfl00106_0130	2.510
kfl00394_0030	*Chaetosphaeridium globosum*	JO168263.1	5.124
kfl00394_0030	*Coleochaete sp.*	JO250815.1	4.669
kfl00394_0030	*Nitella hyalina*	JO299964.1	3.732
kfl00394_0030	*Penium margaritaceum*	JO230496.1	4.030
kfl00456_0020	*K. flaccidum*	kfl00049_0100	0.197
kfl00456_0020	*Nitella hyalina*	JO316377.1	3.015
kfl00456_0020	*Chaetosphaeridium globosum*	JO158516.1	3.190
kfl00456_0020	*Coleochaete sp.*	JO249079.1	3.594
kfl00633_0010	*K. flaccidum*	kfl00038_0010	0.462
kfl00633_0010	*Nitella hyalina*	JO282723.1	3.973
kfl00633_0010	*Coleochaete sp.*	JO249789.1	4.834
kfl00633_0010	*Spirogyra pratensis*	JO184038.1	4.141
kfl00633_0010	*Chaetosphaeridium globosum*	JO159830.1	4.311

Supplementary Figure 80. (**a**) Ks distribution of *K. falccidum* 129 duplicate gene pairs. (b) Ks value of syntenic paralog in *K. flaccidum* (6 pairs) and syntenic ortholog between *K. flaccidum* and the late diverging charophyta (13 pairs).

Supplementary Table1. *K. flaccidum* genome and transcripts statistics

Statistical analysis of reads			
Roche 454 GS FLX Titanium	Single run	4,103,759	reads
	3 Kbp mate pair	1,483,727	reads
Illumina Genome Analyzer IIx	75 bp paired	54,525,596	reads
Statistics for assembly of the nuclear genome			
estimated genome size		117.1 ± 21.8 Mbp	(Supplementary Fig. 1)
chromosome number		n = 22	Chaudhary B.R. and Sarma Y.S.R.K. (1978)
Number of scaffolds		1,812	scaffolds
Total scaffolds length		103,921,766	bp
Scaffold N50 size (N50)		134,930	bp (229 scaffolds)
Contig N50 size		56,148	bp
Peak coverage		40	
Genomic G+C content (%)		52.4	%
gene dencity		154.6	genes / 1 Mbp
Telomere nucleotide sequences		(TTTTAGGG)n	
Sequencing error (checked by sanger sequence of 456 fosmid-end 277,625 bp)			
	Insertion	1 error / 18,508	bp
	Deletion	0 error	
	Substitution	1 error / 10,282	bp
Putative assemble error (checked by 382 fosmid-end, 191 clones)			
abnormal fosmid clone(Chimera clone or miss assemble.)		1 clone / 191	clones
Statistical analysis of organellar genomes			
Chloroplast (scaffold kfl01813)		181,482	bp
	Protein-coding genes	117	genes
	tRNA	36	genes
	rRNA	6	genes
Mitochondrion (scaffold kfl01814)		106,468	bp
	Protein-coding genes	35	genes
	tRNA	29	genes
	rRNA	3	genes
Statistical analysis of transcripts			
Roche 454 GS FLX Titanium	Single run	1,457,422	reads
	assembled sequences	18,735	sequences
Statistical analysis of transcripts mapping to the nuclear genome			
18,735 *K. flaccidum* ESTs in this paper		18,382	ESTs were mapped to genome (98.1 %)
24,923 *K. flaccidum* ESTs in public database, Timme, R.E. et al. (2012)		23,943	ESTs were mapped to genome (96.1 %)
Statistical analysis of predicted protein-coding nuclear genes			
Predicted protein-coding genes		16,063	genes
at least partially supported by assembled transcripts		10,731	genes (66.2%)
fully supported by assembled transcripts		7,376	genes (45.5%)
Number of introns		93,007	
Number of introns per CDS (mean / median)		5.79 / 5	
intron length (mean / median)		498 / 427	bp
Number of single exon CDSs:		360	genes
Statistical analysis of non-coding RNA			
tRNA (Supplementary Data 10.)		98	copies
miRNA (Supplementary Data 11.)		72	copies
snRNA (Supplementary Data 11.)		28	copies
snoRNA (Supplementary Data 11.)	CD box	55	copies
	HACA box	2	copies
	other sno RNA	28	copies

Supplementary Table 2. Statistics of gene family analysis (Fig. 3a and Supplementary Fig. 8)

	Dataset 1 (Fig. 3a)				Dataset 2 (Supplementary Fig. 8)			
organism	both	Algae	Plant	unique	both	Algae	Plant	unique
Cyanidioschyzon merolae	2,668	367	84	1,895	2,679	352	79	1,904
Chondrus crispus	2,761	539	211	5,860	2,861	538	217	5,755
Ectocarpus siliculosus	4,477	1,356	84	10,658	4,482	1,370	142	10,581
Phaeodactylum tricornutum	4,002	1,280	61	5,059	4,053	1,259	87	5,009
Ostreococcus tauri	3,797	1,163	92	2,673	3,376	1,107	109	3,291
Micromonas strain RCC299	4,743	1,577	128	3,655	4,777	1,599	124	3,544
Chlorella variabilis NC64A	4,400	1,197	272	3,922	4,533	1,163	226	3,869
Chlamydomonas reinhardtii	4,431	3,285	41	9,980	4,722	2,557	105	7,028
Volvox carteri f. nagariensis	4,108	2,919	42	7,902	4,428	2,582	87	7,339
Klebsormidium flaccidum	6,105	1,070	1,238	7,650	6,170	908	1,170	7,815
Physcomitrella patens subsp. patens	11,225	383	4,454	16,211	11,381	317	4,273	19,838
Selaginella moellendorffii	8,740	662	4,272	8,599	15,656	927	6,988	11,176
Oryza sativa subsp. japonica	11,940	65	10,244	16,800	10,672	108	8,509	9,103
Populus trichocarpa	15,961	90	15,903	9,381	15,535	198	14,519	10,171
Arabidopsis thaliana	11,891	30	10,867	4,628	16,634	35	13,263	5,241

Supplementary Table 3. Statistics of gene family analysis (Fig. 4a, Supplementary Fig. 3 and Supplementary Fig. 10)

	Dataset 1 (Fig. 4a and Supplementary Fig. 3)					Dataset 2 (Supplementary Fig. 3 and Supplementary Fig. 10)				
#Taxon	Total number of genes	Number of Gene families	Number of of the gene families of paralogues	Number of of the genes of paralogues	Number of singleton	Total number of genes	Number of Gene families	Number of of the gene families of paralogues	Number of of the genes of paralogues	Number of singleton
Cyanidioschyzon merolae	5,014	4,144	505	1,375	3,639	5,014	4,145	506	1,375	3,639
Chondrus crispus	9,371	7,078	833	3,126	6,245	9,371	7,083	830	3,118	6,253
Ectocarpus siliculosus	16,589	12,359	1,482	5,712	10,877	16,589	12,340	1,494	5,743	10,846
Ostreococcus tauri	7,725	6,324	712	2,113	5,612	7,883	6,727	632	1,788	6,095
Chlorella variabilis NC64A	9,791	7,457	1,016	3,350	6,441	9,791	7,445	1,024	3,370	6,421
Micromonas strain RCC299	10,103	8,007	959	3,055	7,048	10,044	7,996	940	2,988	7,056
Phaeodactylum tricornutum	10,402	7,239	1,346	4,509	5,893	10,408	7,228	1,356	4,536	5,872
Chlamydomonas reinhardtii	17,737	13,841	1,492	5,388	12,349	14,412	11,143	1,363	4,632	9,780
Volvox carteri f. nagariensis	14,971	11,971	1,158	4,158	10,813	14,436	10,911	1,283	4,808	9,628
Klebsormidium flaccidum	16,063	11,257	1,808	6,614	9,449	16,063	11,253	1,810	6,620	9,443
Oryza sativa subsp. *japonica*	39,049	20,240	4,638	23,447	15,602	28,392	14,400	3,663	17,655	10,737
Selaginella moellendorffii	22,273	10,423	3,295	15,145	7,128	34,747	10,899	7,432	31,280	3,467
Arabidopsis thaliana	27,416	10,964	4,169	20,621	6,795	35,173	10,922	5,514	29,765	5,408
Physcomitrella patens subsp. *patens*	32,273	19,640	4,463	17,096	15,177	35,809	17,313	4,349	22,845	12,964
Populus trichocarpa	41,335	15,040	6,417	32,712	8,623	40,423	15,257	6,358	31,524	8,899

Supplementary Table 4. Number of Pfam domains and domain combinations (Fig. 4b and Supplementary Fig. 11).

	Dataset 1 (Fig. 4b)			Dataset 2 (Supplementary Fig. 11)		
#Taxon	Total number of genes	Number of domains	Number of domain combinations	Total number of genes	Number of domains	Number of domain combinations
Cyanidioschyzon merolae	5,014	3,309	2,599	5,014	3,309	2,599
Chondrus crispus	9,371	3,502	2,836	9,371	3,502	2,836
Ectocarpus siliculosus	16,589	5,531	5,633	16,589	5,531	5,633
Ostreococcus tauri	7,725	4,143	3,460	7,883	4,185	3,510
Chlorella variabilis NC64A	9,791	5,203	4,552	9,791	5,203	4,552
Micromonas strain RCC299	10,103	5,203	4,530	10,044	5,200	4,521
Phaeodactylum tricornutum	10,402	4,162	3,670	10,408	4,176	3,680
Chlamydomonas reinhardtii	17,737	6,419	6,731	14,412	5,296	4,736
Volvox carteri f. nagariensis	14,971	5,709	5,075	14,436	5,869	5,398
Klebsormidium flaccidum	16,063	6,298	5,757	16,063	6,298	5,757
Oryza sativa subsp. japonica	39,049	6,715	8,653	28,392	6,348	7,997
Selaginella moellendorffii	22,273	6,288	6,376	34,747	6,438	6,956
Arabidopsis thaliana	27,416	6,849	7,463	35,173	6,804	7,727
Physcomitrella patens subsp. *patens*	32,273	6,804	6,842	35,809	6,920	7,178
Populus trichocarpa	41,335	6,882	8,717	40,423	6,931	8,821

Supplementary Table 5.
Number of damains and domain combinations that are commonly found in five land plants (Fig. 4c and Supplementary Fig. 12).

	Dataset 1 (Fig. 4c)		Dataset 2 (Supplementary Fig. 12)	
#Taxon	Number of domains	Number of domain combinations	Number of domains	Number of domain combinations
commonly found in five land plants	4,894	2,801	4,676	2,708
Phaeodactylum tricornutum	2,989	1,507	2,899	1,470
Cyanidioschyzon merolae	2,635	1,331	2,556	1,298
Chondrus crispus	2,668	1,258	2,597	1,234
Ectocarpus siliculosus	3,496	1,640	3,360	1,600
Micromonas strain RCC299	3,548	1,734	3,411	1,676
Ostreococcus tauri	3,110	1,526	3,010	1,386
Chlorella variabilis NC64A	3,592	1,776	3,470	1,741
Volvox carteri f. nagariensis	3,616	1,701	3,509	1,656
Chlamydomonas reinhardtii	3,845	1,810	3,392	1,728
Klebsormidium flaccidum	4,441	2,360	4,262	2,283

Supplementary Table 6. Plant hormones measurements determined using mass spectrometry.

(n = 3)

	endo-conc (ng/g FW)	s.d.	endo-conc (ng/g DW)	s.d.
GA1	0		0	
IAA	2.921 ±	0.816	17.983 ±	3.675
ABA	1.004 ±	0.690	6.046 ±	3.982
JA	0.076 ±	0.009	0.481 ±	0.095
GA4	0		0	
JA-Ile	0		0	
SA	10.810 ±	12.054	64.299 ±	68.325
tZ	0		0	
DHZ	0		0	
iP	0.001 ±	0.001788	0.006 ±	0.010

Supplementary Table 7. Datasets of gene family and domain analysis.

#Taxon	Dataset 1(mainly JGI data)	Dataset 2(mainly refseq data)
Cyanidioschyzon merolae	Matsuzaki, M. et al. (2004)	Matsuzaki, M. et al. (2004)
Chondrus crispus	Collen, J. et al. (2013)	Collen, J. et al. (2013)
Ectocarpus siliculosus	Cock, J.M. et al. (2010)	Cock, J.M. et al. (2010)
Ostreococcus tauri	JGI v2.0	refseq 54
Chlorella variabilis NC64A	JGI v1.0	JGI v1.0
Micromonas strain RCC299	JGI phytosome299	refseq 54
Phaeodactylum tricornutum	JGI v2.0	refseq 54
Chlamydomonas reinhardtii	JGI phytosome236	refseq 54
Volvox carteri f. nagariensis	JGI phytosome199	refseq 54
Klebsormidium flaccidum	v1.0 (in this paper)	v1.0 (in this paper)
Oryza sativa subsp. *japonica*	JGI phytosome204	refseq 54
Selaginella moellendorffii	JGI phytosome91	refseq 54
Arabidopsis thaliana	JGI phytosome167	refseq 54
Physcomitrella patens subsp. *patens*	JGI phytosome152	refseq 54
Populus trichocarpa	JGI phytosome210	refseq 54

Supplementary Table 8. Best reciprocal protein hits for *K. flaccidum* with other plants proteins.

#Taxon		data source	Number of sequences	reciprocal best hits of K.flaccidum genes (16,063 genes)
Land plants	*Arabidopsis thaliana*	JGI phytosome167	27,416	6,100 genes 38.0 %
	Populus trichocarpa	JGI phytosome210	41,335	6,187 genes 38.5 %
	Oryza sativa subsp. *japonica*	JGI phytosome204	39,049	5,886 genes 36.6 %
	Selaginella moellendorffii	JGI phytosome91	22,273	6,087 genes 37.9 %
	Physcomitrella patens subsp. *patens*	JGI phytosome152	32,273	6,698 genes 41.7 %
Algae	*Chlamydomonas reinhardtii*	JGI phytosome236	17,737	5,055 genes 31.5 %
	Volvox carteri f. nagariensis	JGI phytosome199	14,971	4,560 genes 28.4 %
	Chlorella variabilis NC64A	JGI v1.0	9,791	4,613 genes 28.7 %
	Micromonas strain RCC299	JGI phytosome299	10,103	4,572 genes 28.5 %
	Ostreococcus tauri	JGI v2.0	7,725	3,653 genes 22.7 %
	Cyanidioschyzon merolae	Matsuzaki, M. et al. (2004)	5,014	2,505 genes 15.6 %
	Chondrus crispus	Collen, J. et al. (2013)	9,371	2,506 genes 15.6 %
	Phaeodactylum tricornutum	JGI v2.0	10,402	2,979 genes 18.5 %
	Ectocarpus siliculosus	Cock, J.M. et al. (2010)	16,589	3,766 genes 23.4 %

Supplementary Table 9. Best reciprocal EST hits for *K. flaccidum* with other charophyte algae ESTs.

#Taxon		data source	Number of EST sequences	reciprocal best hits of K.flaccidum genes (16,063 genes)	
Charophyta	*Mesostigma viride*	Timme, R.E. et al. (2012)	15,972	761 genes	4.7 %
	Chlorokybus atmophyticus	Timme, R.E. et al. (2012)	12,496	4,296 genes	26.7 %
	Klebsormidium flaccidum	v1.0 transcriptome (in this paper)	18,661	9,660 genes	60.1 %
	Klebsormidium flaccidum	Timme, R.E. et al. (2012)	24,923	9,361 genes	58.3 %
	Nitella hyalina	Timme, R.E. et al. (2012)	40,615	3,120 genes	19.4 %
	Chaetosphaeridium globosum	Timme, R.E. et al. (2012)	24,200	3,425 genes	21.3 %
	Coleochaete sp.	Timme, R.E. et al. (2012)	18,386	4,247 genes	26.4 %
	Spirogyra pratensis	Timme, R.E. et al. (2012)	9,587	2,714 genes	16.9 %
	Penium margaritaceum	Timme, R.E. et al. (2012)	29,220	3,160 genes	19.7 %

Supplementary Table 10. Predicted proteins, rRNAs and tRNAs in organellar genomes

Chloroplast (Predicted proteins and rRNAs)

Photosystem I	*psaA*	*psaB*	*psaC*	*psaI*	*psaJ*										
Photosystem II	*psbA* **	*psbB*	*psbC*	*psbD*	*psbE*	*psbF*	*psbH*	*psbI*	*psbJ*	*psbK*	*psbL*	*psbM*	*psbN*	*psbT*	*psbZ*
Cytochrome b6/f	*petA*	*petB*	*petD*	*petG*	*petN*	*petL*									
ATP synthase	*atpA*	*atpB*	*atpE*	*atpF*	*atpH*	*atpI*									
Chlorophyll biosynthesis	*chlB*	*chlI*	*chlL*	*chlN*											
Rubisco	*rbcL*														
NADH oxidoreductase	*ndhA*	*ndhB*	*ndhC*	*ndhD* *	*ndhE*	*ndhF*	*ndhG*	*ndhH*	*ndhI*	*ndhJ*	*ndhK**				
Large subunit ribosomal proteins	*rpl2*	*rpl19*	*rpl20*	*rpl21*	*rpl22*	*rpl32*	*rpl33*	*rpl36* *							
Small subunit ribosomal proteins	*rps2*	*rps4*	*rps7*	*rps8*	*rps9*	*rps11*	*rps12* **	*rps14*	*rps18*	*rps19*					
RNAP	*rpoA*	*rpoB*	*rpoC1*	*rpoC2*											
Translation factor	*tufA*														
Other proteins	*ccsA*	*cemA* *	*clpP* *	*odpB*	*matK*										
Proteins of unknown function	*ycf1*	*ycf4*	*ycf12*	*ycf20*	*ycf62*	*ycf66*									
	ORF183	*ORF502*	*ORF619*	*ORF1731*	*ORF263*	*ORF147*	*ORF75*	*ORF180*	*ORF405*	*ORF274*	*ORF91*	*ORF167*			
Ribosomal RNAs	*rrn23*	*rrn16*	*rrn5*												

red symbols : Two gene copies due to the IR

* predicted intron containing

** predicted trans-splicing

Mithochondria (Predicted proteins and rRNAs)

Complex I (NADH dehydrogenase)	*nad1*	*nad2*	*nad3*	*nad4*	*nad4L*	*nad5*	*nad6*	*nad7*	*nad9*
Complex II (succinate dehydrogenase)	*sdh3*	*sdh4*							
complex III (ubichinol cytochrome c reductase)	*cob*								
Complex IV (cytochrome c oxidase)	*cox1* *	*cox2**	*cox3*						
Complex V (ATP synthase)	*atp1*	*atp4*	*atp6*	*atp8* **	*atp9*				
Cytochrome c biogenesis	*ccmB*	*ccmC*	*ccmF* *						
Ribosomal proteins (SSU)	*rps3*	*rps4*	*rps7*	*rps11*	*rps12*	*rps14*			
Ribosomal proteins (LSU)	*rpl14*	*rpl16*							
Other genes	*tatC* (*mttB*)								
	unknown modification methylase								
	putative integrase_recombinase protein								
	putative reverse transcriptase and intronmaturase								
Ribosomal RNAs	*rrn23*	*rrn16*	*rrn5*						

* predicted intron containing

Supplementary Table 11. codon and anticodon tables in organellar genomes

Chloroplast (codon and anticodon table)

I	II																		III
	U					C				A				G					
U	U	U	U	Phe	A A A	U	C	U	Ser	A G A	U	A	U	Tyr	A U A	U	G	U	Cys	A C A	U
	U	U	C		G A A	U	C	C		G G A	U	A	C		G U A	U	G	C		G C A	C
	U	U	A	Leu	U A A	U	C	A		U G A	U	A	A	ocher	U U A	U	G	A	opal	U C A	A
	U	U	G		C A A	U	C	G		C G A	U	A	G	amber	C U A	U	G	G	Trp	C C A	G
C	C	U	U	Leu	A A G	C	C	U	Pro	A G G	C	A	U	His	A U G	C	G	U	Arg	A C G	U
	C	U	C		G A G	C	C	C		G G G	C	A	C		G U G	C	G	C		G C G	C
	C	U	A		U A G	C	C	A		U G G	C	A	A	Gln	U U G	C	G	A		U C G	A
	C	U	G		C A G	C	C	G		C G G	C	A	G		C U G	C	G	G		C C G	G
A	A	U	U	Ile	A A U	A	C	U	Thr	A G U	A	A	U	Asn	A U U	A	G	U	Ser	A C U	U
	A	U	C		G A U	A	C	C		G G U	A	A	C		G U U	A	G	C		G C U	C
	A	U	A		U A U	A	C	A		U G U	A	A	A	Lys	U U U	A	G	A	Arg	U C U	A
	A	U	G	Met	C A U	A	C	G		C G U	A	A	G		C U U	A	G	G		C C U	G
G	G	U	U	Val	A A C	G	C	U	Ala	A G C	G	A	U	Asp	A U C	G	G	U	Gly	A C C	U
	G	U	C		G A C	G	C	C		G G C	G	A	C		G U C	G	G	C		G C C	C
	G	U	A		U A C	G	C	A		U G C	G	A	A	Glu	U U C	G	G	A		U C C	A
	G	U	G		C A C	G	C	G		C G C	G	A	G		C U C	G	G	G		C C C	G

: predicted tRNA

blue symbols : *trnM, trnI, trnfM*

red symbols : multicopies

Mithochondria (codon and anticodon table)

I	II																		III
	U					C				A				G					
U	U	U	U	Phe	A A A	U	C	U	Ser	A G A	U	A	U	Tyr	A U A	U	G	U	Cys	A C A	U
	U	U	C		G A A	U	C	C		G G A	U	A	C		G U A	U	G	C		G C A	C
	U	U	A	Leu	U A A	U	C	A		U G A	U	A	A	ocher	U U A	U	G	A	opal	U C A	A
	U	U	G		C A A	U	C	G		C G A	U	A	G	amber	C U A	U	G	G	Trp	C C A	G
C	C	U	U	Leu	A A G	C	C	U	Pro	A G G	C	A	U	His	A U G	C	G	U	Arg	A C G	U
	C	U	C		G A G	C	C	C		G G G	C	A	C		G U G	C	G	C		G C G	C
	C	U	A		U A G	C	C	A		U G G	C	A	A	Gln	U U G	C	G	A		U C G	A
	C	U	G		C A G	C	C	G		C G G	C	A	G		C U G	C	G	G		C C G	G
A	A	U	U	Ile	A A U	A	C	U	Thr	A G U	A	A	U	Asn	A U U	A	G	U	Ser	A C U	U
	A	U	C		G A U	A	C	C		G G U	A	A	C		G U U	A	G	C		G C U	C
	A	U	A		U A U	A	C	A		U G U	A	A	A	Lys	U U U	A	G	A	Arg	U C U	A
	A	U	G	Met	C A U	A	C	G		C G U	A	A	G		C U U	A	G	G		C C U	G
G	G	U	U	Val	A A C	G	C	U	Ala	A G C	G	A	U	Asp	A U C	G	G	U	Gly	A C C	U
	G	U	C		G A C	G	C	C		G G C	G	A	C		G U C	G	G	C		G C C	C
	G	U	A		U A C	G	C	A		U G C	G	A	A	Glu	U U C	G	G	A		U C C	A
	G	U	G		C A C	G	C	G		C G C	G	A	G		C U C	G	G	G		C C C	G

: predicted tRNA

red symbols : multicopies

Supplementary Table 12. Occurrence of repetitive sequences in the *K. flaccidum* genome.

		copy number	Coverage	Occupation
Total interspersed repeats			318355 bp	0.31 %
	Retroelements	1169	303385 bp	0.29 %
	SINEs:	2	95 bp	0 %
	Penelope	2	366 bp	0 %
	LINEs:	122	9615 bp	0.01 %
	R2/R4/NeSL	2	188 bp	0 %
	RTE/Bov-B	9	661 bp	0 %
	L1/CIN4	106	8168 bp	0.01 %
	LTR elements:	1045	293675 bp	0.28 %
	Ty1/Copia	464	141063 bp	0.14 %
	Gypsy/DIRS1	570	151222 bp	0.15 %
	DNA transposons	147	11669 bp	0.01 %
	hobo-Activator	24	1584 bp	0 %
	Tourist/Harbinger	11	1034 bp	0 %
	Other (Mirage,P-element, Transib)	7	558 bp	0 %
	Unclassified:	35	3301 bp	0 %
Local repeats				
Simple repeats:		8773	412002 bp	0.4 %
Low complexity:		1244	66703 bp	0.06 %

Supplementary Methods

Culture conditions, and DNA and RNA extraction

The *K. flaccidum* strain NIES-2285 was cultured in liquid C medium[1] or 0.1% glucose + BCDAT medium[2] under continuous light (10 μmol photons m^{-2} s^{-1}) with the cells circulated by bubbling air through the medium. Cells were harvested by vacuum filtration through filter paper (No. 3, 70 mm diameter, Advantec, Tokyo, Japan).

The genome size was estimated by measuring the fluorescence intensity of DAPI-stained nuclei and comparing mean relative fluorescence of each sample with an internal standard, *C. merolae*[3] (16.5 Mbp, GC content = 55.0%), (Supplementary Fig. 1).

Genomic DNA was extracted using the DNeasy Plant Mini kit (Cat. No. 69104, Qiagen, Hilden, Germany). Total RNA was extracted using the following acid-phenol method. The pelleted sample—collected by filtration of the cultured algae—was frozen in liquid nitrogen and ground to a fine powder in a mortar. At least six volumes of RNA extraction buffer (0.8% SDS, 25 mM Tris-HCl pH 7.6, 25 mM $MgCl_2$, 25 mM KCl)/acid phenol (1:1) were added, and the aqueous phase obtained was extracted three times with acid phenol/chloroform (1:1). The RNA was precipitated by adding an equal volume of isopropanol. Next, RNA was purified using the RNeasy Plant Mini kit or

RNeasy midi kit (Cat. No. 74104 / 75142, Qiagen). Purified DNA and RNA were sequenced using the Roche 454 GS FLX Titanium or Illumina GAIIx platform (Supplementary Table 1).

Assembly of organellar genomes

Organellar genomes were assembled using Newbler[4] version 2.0 (Roche) and Mira3[5], and were manually assembled from 1,537,397 reads generated using the Roche 454 GS FLX Titanium platform. The assembly scheme is shown in Supplementary Fig. 74. In this scheme, candidate organellar contigs were selected by TBLASTN[6] analysis following comparison with other organellar proteins from members of Charophyta. Mitochondrial sequences used in the analysis were from *Mesostigma viride*[7] (NC_008240), *Chlorokybus atmophyticus*[8] (NC_009630), *Chara vulgaris*[9] (NC_005255), and *Chaetosphaeridium globosum*[10] (NC_004118). Chloroplastic sequences used in the analysis were from *M. viride*[11] (NC_002186), *C. atmophyticus*[12] (NC_008822), *Staurastrum punctulatum*[13] (NC_008116), *Zygnema circumcarinatum*[13] (NC_008117), *C. vulgaris*[14] (NC_008097), and *C. globosum*[10] (NC_004115). Finally, sequence errors of organellar genomes were corrected by mapping of Illumina GAIIx reads using BWA0-0.6.1[15]. Assembly statistics are shown in Supplementary Table 1.

The chloroplast and mitochondrial genomes were registered as scaffold-kfl01813 and kfl01814, respectively.

Assembly of the nuclear genome and transcript sequences

All DNA sequence reads obtained using Roche 454 GS FLX Titanium and Illumina GAIIx were assembled using Newbler version 2.6 (scaffold option, masked organellar sequence) for nuclear genome assembly. All reads of Roche 454 GS FLX Titanium using RNA samples were assembled using Newbler version 2.6 (cdna option, urt option, masked organellar sequence). Assembly statistics are shown in Supplementary Table 1.

Validation of sequencing and assembly of nuclear genome

The accuracy of sequencing and assembly of the nuclear genome was checked by methods i and ii (shown below), the completeness of assembly of the nuclear genome was checked by methods ii and iii, and contamination of bacterial sequences was checked by methods iv and v.

i) Fosmid-end sequencing

The CopyControl Fosmid Library Production kit (Epicentre Biotechnologies, Madison,

WI, USA) was used to construct genomic DNA libraries. Appropriate quantities of DNA were ligated and packaged according to the manufacturer's protocol. Fosmid ends purified as 307 clones were sequenced using the Sanger method. The fosmid-end sequences including highly repetitive sequences were identified and removed by Illumina read mapping with Bowtie2[16]. The 456 fosmid-end sequences were used to validate the accuracy of sequencing and assembly (Supplementary Table 1 and Supplementary Fig. 75). Telomere nucleotide sequences (TTTTAGGG) were determined from fosmid clones (Supplementary Table 1).

ii) Remapping 454 sequences

To assess the possibility of segmental duplications of the genome, all reads sequenced from the Roche 454 sequencer were mapped to all the scaffolds using gsMapper software (Roche/454). Redundancies of the 454 reads on the genome were counted in each 1-kb region (Supplementary Fig. 76).

iii) Transcriptome mapping

We mapped assembled transcript sequences and public *K. flaccidum* ESTs to scaffolds using SPALN 2.0.4[17]. In total, 98.1% of our assembled transcript sequences and 96.1%

of public *K. flaccidum* ESTs were mapped to the nuclear sequence (Supplementary Table 1).

iv) Contamination of bacterial 16S ribosomal sequences

All scaffolds were assessed as to whether they contained prokaryotic 16S rRNA genes using BLASTN[6] (with parameters -F F -e 1e-8) against the SILVA SSU rRNA sequence database[18]. There was no obvious bacterial genome contamination.

v) Taxonomy mapping of all proteins

We performed taxonomy mapping of all predicted proteins using BLASTP[6] to the NCBI-NR database and MEGAN4[19] (Supplementary Fig. 77). Contamination with bacterial sequences in 30 scaffolds that had only proteins mapped to bacterial proteins was checked by BLASTN to the nt database. No hits to already-known bacterial genome sequences were found.

Prediction and annotation of organellar genes

Organellar genes were predicted using Glimmer3[20], GeneMarkP[21], GeneMark (a heuristic approach for gene prediction)[22], FGENESB[23], tRNAScan-SE[24], and

RNAmmer[25]. In addition, the presence of certain organellar genes was predicted using BLASTP with organellar genes of other species. These predicted genes were manually curated and annotated (Supplementary Table 10 and Supplementary Table 11). The circular-mapped organellar genomes were drawn with Artemis[26] (Supplementary Fig. 78).

Prediction and annotation of nuclear genes

Nuclear genes were predicted using Augustus 2.5.5[27]. Assembled transcript sequences were mapped to scaffolds using SPALN 2.0.4[17] to assess the likelihood that each sequence was indeed a transcript. The manually curated 309 gene models (without transposable elements, Supplementary Data 9) were used as Augustus training sets, and 16,078 genes were predicted by Augustus using transcript evidence. These genes were manually curated based on data of blast2GO[28], BLASTP, interpro[29], Gclust[30], target[31], ipsort[32], KAAS[33] and rough phylogenetic analysis (clustalW[34], MUSCLE 3.8 [35], Gblocks 0.91b[36], FastTree 2.1.4[37]) with Artemis[26]. Finally, IDs were assigned for 16,063 genes.

Prediction of transposable elements and non-coding RNAs

Predicted transposable elements in all scaffolds were identified by using RepeatMasker (version 4.0.2)[38] with WU-blast[39] against a plant-specific database (-species Viridiplantae) in Repbase (20130422)[40] (Supplementary Table 12). The tRNA genes were predicted using tRNA-scan SE 1.3.1[41] with default parameters (Supplementary Data 10). The micro RNA, small nuclear RNA and small nucleolar RNA genes were predicted using INFERNAL software (version 1.0)[42] with mapping to all scaffolds against the Rfam database (release 11.0)[43] under default parameters (Supplementary Data 11).

Analysis of genome duplication

Genome duplication events in K. flaccidum were inferred by analyses of (i) synteny blocks and (ii) duplicated genes.

i) Analysis of synteny blocks

Synteny blocks in *K. flaccidum* scaffolds were identified by Sibelia[44] (parameter set = "fine"). There were 273 segmental duplications (≥1 kbp) in 628 regions. However, there were few large duplications (≥5k bp; 16 blocks, 31 regions; Supplementary Fig. 79).

ii) Analysis for duplicated genes

Duplicated gene sets were identified by USEARCH v6[45] (uclast program, identity ≥50%, query coverage ≥30%, target coverage ≥30%). A total of 825 genes (347 clusters) were selected. Next we removed putative transposable elements and genes that were unique to *K. flaccidum* (Supplementary Table 2 : dataset 1). Putative Charophyte orthologs were selected from reciprocal TBLASTN-BLASTX[6] best hits. A total of 129 duplicate gene pairs (syntenic paralogue in *K. flaccidum*) and 13 gene pairs (syntenic orthologue, *K. flaccidum* vs. charophyte ESTs) were aligned by MUSCLE 3.8[35] and manually curated. The synonymous substitution rate (Ks) was estimated using KaKs_Calculator[46] with the method of model averaging (YN, MYN, GY, etc.) (Supplementary Fig. 80). Although only 5 genes had a best hit to other charophyte ESTs, Ks values of duplicate pairs in *K. flaccidum* were lower than those of 13 pairs between *K. flaccidum* and other charophyte ESTs. These results suggest that there is no indication of whole genome duplication, but an ancestor *of K. flaccidum* possibly had a certain level of local gene duplication.

Supplemental References

1. Ichimura, T. Sexual cell division and conjugation-papilla formation in sexual

reproduction of *Closterium strigosum*. In Proceedings of the seventh international seaweed symposium. *University of Tokyo Press*, 208–214 (1971).

2. Nishiyama, T. Tagged mutagenesis and gene-trap in the moss, *Physcomitrella patens* by shuttle mutagenesis. *DNA Res.* **7**, 9–17 (2000).

3. Matsuzaki, M. *et al.* Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. *Nature.* **428**, 653–657 (2004).

4. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature*. **437**, 376–380 (2005).

5. Chevreux, B., Wetter, T. & Suhai, S. Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. *Computer Science and Biology: Proceedings of the German Conference on Bioinformatics* **99**, 45–56 (1999).

6. Altschul, S. F. *et al*. Basic local alignment search tool. *J Mol Biol.* **215**, 403–10 (1990).

7. Turmel, M., Otis, C. & Lemieux, C. The complete mitochondrial DNA sequence of *Mesostigma viride* identifies this green alga as the earliest green plant divergence and predicts a highly compact mitochondrial genome in the ancestor of all green plants. *Mol. Biol. Evol.* **19**, 24–38 (2002).

8. Turmel, M., Otis, C., & Lemieux C. An unexpectedly large and loosely packed

mitochondrial genome in the charophycean green alga *Chlorokybus atmophyticus*. *BMC Genomics.* **8**:137. (2007).

9. Turmel, M., Otis, C. & Lemieux C. The mitochondrial genome of *Chara vulgaris*: insights into the mitochondrial DNA architecture of the last common ancestor of green algae and land plants. *Plant Cell.* **15**, 1888–903 (2003).
10. Turmel, M., Otis, C. & Lemieux, C. The chloroplast and mitochondrial genome sequences of the charophyte *Chaetosphaeridium globosum*: insights into the timing of the events that restructured organelle DNAs within the green algal lineage that led to land plants. *Proc. Natl Acad. Sci. U S A* **99**, 11275–80 (2002).
11. Lemieux, C., Otis, C. & Turmel, M. Ancestral chloroplast genome in *Mesostigma viride* reveals an early branch of green plant evolution. *Nature.* **403**, 649–52. (2000).
12. Lemieux, C., Otis, C. & Turmel, M. A clade uniting the green algae *Mesostigma viride* and *Chlorokybus atmophyticus* represents the deepest branch of the Streptophyta in chloroplast genome-based phylogenies. *BMC Biol.* **12**;5 (2007).
13. Turmel, M., Otis, C. & Lemieux, C. The complete chloroplast DNA sequences of the charophycean green algae *Staurastrum* and *Zygnema* reveal that the chloroplast genome underwent extensive changes during the evolution of the Zygnematales.

BMC Biol. **3**:22 (2005).

14. Lemieux, C., Otis, C. & Turmel, M. The chloroplast genome sequence of *Chara vulgaris* sheds new light into the closest green algal relatives of land plants. *Mol Biol Evol.* **23**, 1324–1338. (2006).
15. Li, H., & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* **25**, 1754–1760 (2009).
16. Langmead, B., & Salzberg, S. Fast gapped-read alignment with Bowtie 2. *Nature Methods.*, **9**, 357–359 (2012).
17. Wata, H. & Gotoh, O. Comparative analysis of information contents relevant to recognition of introns in many species. *BMC Genomics* **12**, 45 (2011).
18. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–596, (2013).
19. Huson, D.H., Mitra, S., Weber, N., Ruscheweyh, H., & Schuster, S.C. Integrative analysis of environmental sequences using MEGAN4. *Genome Research*, **21**, 1552–1560, (2011).
20. Delcher, A. L., Bratke, K. A., Powers, E. C. & Salzberg S. L. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics,* **23**, 673–679 (2007).
21. Lukashin, A. & Borodovsky, M. GeneMark.hmm: new solutions for gene finding.

Nucleic. Acids. Res. **26**, 1107–1115 (1998).

22. Besemer, J. & Borodovsky, M., Heuristic approach to deriving models for gene finding. *Nucleic. Acids. Res.* **27**, 3911–3920 (1999).

23. Softberry, Inc (http://linux1.softberry.com/berry.phtml).

24. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved transfer RNA detection in genomic sequence. *Nucleic. Acids. Res.*, **25**, 955–964 (1997).

25. Lagesen, K. *et al.* RNammer: consistent annotation of rRNA genes in genomic sequences. *Nucleic. Acids. Res.* **35**, 3100–3108 (2007).

26. Rutherford, K. *et al.* Artemis: sequence visualization and annotation. *Bioinformatics* **16**, 944–945 (2000).

27. Stanke, M., Tzvetkova, A. & Morgenstern, B. AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biology* **7**: S11 (2006).

28. Conesa, A. *et al*. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics.* **21**, 3674–3676 (2005).

29. Hunter, S. *et al*. InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.* **40**, D306–12. (2012).

30. Sato, N. Gclust: trans-kingdom classification of proteins using automatic individual

threshold setting. *Bioinformatics.* **25**, 599–605 (2009).

31. Emanuelsson, O., Brunak, S., von Heijne, G. & Nielsen, H. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc.* **2**, 953–971 (2007).

32. Bannai, H. *et al*. Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics*, **18**, 298–305 (2002).

33. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. & Kanehisa, M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* **35**, W182–W185 (2007)..

34. Larkin, M. A. *et al*. ClustalW and ClustalX version 2.0. *Bioinformatics.* **23**, 2947–2948 (2007).

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

36. Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology* **56**, 564–577 (2007).

37. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 -- approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).

38. Smit, A.F.A., Hubley, R. & Green, P. RepeatMasker Open-3.0. (1996–2010);

http://www.repeatmasker.org

39. Gish, W. (1996–2003) ;http://blast.wustl.edu

40. Jurka, J. et al. Repbase Update, a database of eukaryotic repetitive elements. *Cytogentic and Genome Research* **110**, 462–467 (2005).

41. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence *Nucleic Acids Res*., **25**, 955–964 (1997).

42. Nawrocki, E.P., Kolbe, D.L. & Eddy, S.R. Infernal 1.0: Inference of RNA Alignments. *Bioinformatics*, 25, 1335–1337 (2009).

43. Burge, S.W. *et al*. Rfam 11.0: 10 years of RNA families Nucleic Acids Research (2012)

44. Minkin, I., Patel, A., Kolmogorov, M., Vyahhi, N. & Pham, S.K. Sibelia: A Scalable and Comprehensive Synteny Block Generation Tool for Closely Related Microbial Genomes. WABI 2013, 215–229 (2013)

45. Edgar, R.C. Search and clustering orders of magnitude faster than BLAST *Bioinformatics* **26**, 2460–2461 (2010).

46. Zhang, Z. *et al*. KaKs Calculator: Calculating Ka and Ks through model selection and model averaging. *Genomics Proteomics Bioinformatics* **4**, 259–263 (2006).