SI Appendix for

A novel genetic code and record-setting AT-richness in the highly reduced plastid genome of the holoparasitic plant *Balanophora*

Huei-Jiun Su, Todd J. Barkman, Weilong Hao, Samuel S. Jones, Julia Naumann, Elizabeth Skippington, Eric K. Wafula, Jer-Ming Hu, Jeffrey D. Palmer* and Claude W. dePamphilis*

*Corresponding authors. E-mail: jpalmer@indiana.edu and cwd3@psu.edu

This PDF includes:

www.pnas.org/cgi/doi/10.1073/pnas.1816822115

SI Materials and Methods

GC content, codon usage, and amino-acid usage. The analyses of gene-by-gene GC content in Fig. S5 and gene-wide GC content and codon usage in Table S5 were performed on the two *Balanophora* plastomes and plastid, mitochondrial, or bacterial genomes from 28 other taxa (no AT-rich nuclear genomes were analyzed because of their incompleteness and large sizes). These taxa were selected as follows (see Table S12 for full names and accession numbers): *Zinderia* has the most AT-rich bacterial genome sequenced, *Carsonella* is 2nd-most AT-rich, and *Nasuia* is 3rd-most excluding other *Carsonella* isolates. All three bacteria have highly reduced genomes and are obligate intracellular symbionts of insects. The mitochondrial genome of the yeast *Nakaseomyces* is the most AT-rich genome described to date in any organism. The yeast *Saccharomyces cerevisiae* YJM1447 has the most AT-rich of the many sequenced and annotated *Saccharomyces* mitochondrial genomes. *Rozella* has the most AT-rich mitochondrial genome in fungi apart from the yeasts. The insect *Diadegma* has the most AT-rich animal mitogenome and 3rd-most overall. *Radopholus* has the most AT-rich nematode genome. *Monosiga*, *Ichthyophthirius*, and *Acrasis* have the most AT-rich "protist" genomes and represent three different phyla. The four apicomplexans possess the most AT-rich and fully sequenced plastomes from each of four phylogenetically disparate genera of the group. The nine most AT-rich, non-apicomplexan plastomes are from species whose full-plastome GC content is less than 24%; all but one of these plastomes (that of *Bulboplastis*) are from non-photosynthetic organisms. The other five plastomes are from all available non-photosynthetic organisms whose full-plastome GC content is between 24% and 30% (many photosynthetic plastomes are also in this GC range).

GenBank accession numbers of *B. laxiflora* mitochondrial genes assembled in this study and employed in codon-usage analysis are MK144465-MK144474. *Balanophora fungosa* nuclear sequences used for this purpose are from GenBank (JQ613229, JQ613232, JQ613242, JQ613262, JQ613269) and the 1KP database (scaffolds STKY 0018172, 0079935, 0095611, 0104504, 2000678, 2000811, 2001013, 2002472, 2002620, 2002847, 2003209, 2003887, 2003966, 2004417, 2005545, 2005849, 2006043, 2006503, 2007049, 2007203, 2007307, 2007505, 2007706, 2007890, 2008143, 2008354, 2008508, 2008673, 2008894, 2009559, 2010082, 2010496, 2010783, 2011060, 2011423, 2011655, 2011904, 2012280, 2012406, 2075101, 2075212, 2075289, 2075967, 2076149, 2076342, 2076680, 2077455, 2077652, 2077741, 2078058, 2078082, 2078577, 2078768, 2078866, 2078885, 2078962, 2079157, 2079236, 2079392, 2079434).

Transcript analysis. Total RNA was extracted from *B. laxiflora* developing female inflorescence tissue using Concert Plant reagent (Invitrogen, Carlsbad, CA, USA). Eleven *B. laxiflora* plastid genes were sufficiently GC-rich to enable design of effective PCR primers (Tables S10A and S11C). Complementary DNAs (cDNAs) were generated for these genes by RT-PCR amplification and sequenced using the Sanger method and the primers listed in Table S10.

To ensure that RT-PCR products were derived from RNA and not genomic DNA contamination, several controls were devised using RNA and DNA templates with various applications of RNase, DNase, and/or reverse transcriptase as in (1). In the case of DNA controls, total DNA was treated with RNase A. In the case of RNA controls, total RNA was treated with DNase I. Both DNase-treated RNA and RNase-treated DNA were used for cDNA synthesis reactions using the Maxima First Strand cDNA Synthesis

Kit following the manufacturer's protocol (Thermo Fisher Scientific). One µg of template RNA was used from the sample treated with DNase I as measured by a Qubit Fluorometer using the Qubit RNA BR Assay Kit (Thermo Fisher Scientific) with a corresponding volume taken from the sample treated with RNase A. RNA digestions were performed in solution with 300 µg RNase A at 37°C for one hour and subsequently purified using the DNeasy Spin Column (Qiagen). DNase digestions were performed following Appendix C of the RNeasy MinElute Clean-up Handbook (Qiagen). PCR amplification of single-stranded cDNA was performed using DreamTaq Green PCR Master Mix, with primer molarity of 0.5 µM and template concentration of 0.2 ng/µl. Thermal cycling parameters varied widely (Table S11) due to the extreme A+T content of most genes, which required substantially lowered melting and extension temperatures in most cases (2). Gel electrophoresis was performed on all amplified products using a 1.5% agarose gel containing 0.5X SyberSafe dye (Life Technologies).

As an additional source of *Balanophora* cDNA sequences, the *B. fungosa* transcriptome assemblies from the 1KP project (http://www.onekp.com/public_data.html) were used as queries in BLASTn (E-value = $1e^{-10}$) searches against a database of 626 plastomes, including the *Balanophora* plastomes from this study.

Microscopy. For light microscopy, sections of basal floral bracts, inflorescence stalks, and tubers from *B. yakushimensis* were stained with either Sudan black or iodine solution. The Sudan-black solution contained 0.3% Sudan black dissolved in 70% ethanol; tissues were stained for 10 minutes at room temperature and washed with distilled water. The iodine solution consisted of 2% KI and 1% I_2; tissues were stained for 5 minutes without washing. For transmission electron microscopy, basal bract tissues of *B. laxiflora* and *B. yakushimensis* were fixed with 2.5% glutaraldehyde in 0.1 M sodium phosphate buffer (pH 7.0) at 4°C for 24 hours. After three 20-min buffer rinses, the samples were post-fixed in 1% OsO_4 in the same buffer for 4 hours at room temperature and then rinsed in three 20-min buffer changes. Samples were first dehydrated in an ethanol series and then treated with propylene oxide, embedded in Spurr's resin, and sectioned with a Leica EM UC6 or UC7 ultramicrotome. The resulting sections, of 70-90 nm thickness, were stained with uranyl acetate and lead citrate. Sections were observed using a FEI Tecnai Spirit Transmission Electron Microscope at 80 KV and photographed using a Gatan Orius CCD camera.

SI Results

The extraordinarily divergent *ycf*2 gene. In all respects, the most divergent gene in *Balanophora* plastomes is *ycf2*. This ca. 750-bp ORF is remarkably shrunken compared to all other annotated *ycf2* genes (of typically 5,100-6,900-bp length) owing to numerous large deletions across its length (Figs. S4, S12, and S13). YCF2 is only 52% identical between *B. laxiflora* and *B. reflexa*, making it the most sequence-divergent protein encoded by *Balanophora* plastomes (Table S2, Figs. S4, S12, and S13). At 2.5% and 2.2% GC in *B. laxiflora* and *B. reflexa*, respectively, *ycf2* is also the most AT-rich gene in *Balanophora*. Remarkably, GC content in the two *ycf2* genes is almost two times lower than in intergenic and intronic regions of their respective plastomes (Table S1). Because our analyses of *Balanophora ycf2* indicate that it is by far the most unusual form of the gene ever reported as likely functional, we present an extended discussion of its annotation and potential functionality in the context of what is known from other plants.

First, *ycf2* is located between *rpl*2 and *rps*7 in *Balanophora*, exactly where it should be given that gene order is highly conserved between *Balanophora* and the much larger plastomes of photosynthetic angiosperms (Fig. S1). Thus, although it is highly shrunken and divergent in sequence, this ORF is likely to be *ycf2* in terms of synteny.

Second, despite its extreme length, base-compositional, and sequence divergence, *Balanophora ycf2* contains two of three putatively functional motifs identified as conserved between land plant YCF2 and the CDC48 family of ATPases (3). The three recognized motifs constitute the only hint of a clue as to the function of the enigmatic *ycf2*, the largest and most poorly conserved plastid gene, but an essential one (4) thought to have been acquired by the plastome in the common ancestor of land plants (5). These motifs are marked on the YCF2 multiple sequence alignments shown in Figs. S4 and S12**.** Although Walker motif A is likely absent, Walker motif B is present in *Balanophora* and corresponds to a putative nucleotide-binding site. The so called "DPAL" motif is also present in *Balanophora*. Although there is no assigned role for this motif, it was identified as of probable functional importance through its evident homology to the CDC48 family (3). These two motifs that are present in *Balanophora* YCF2 also correspond to the two best-conserved regions of the entire YCF2 protein of land plants (Fig. S4).

Third, the extraordinary divergence and AT-richness of *Balanophora ycf2* is entirely in keeping with what's known about this gene in many other plastid genomes; the *Balanophora* case is simply more extreme. The exceptional nature of *ycf2* was first recognized in 1991 (6) when only two *ycf2* sequences were available, from the angiosperm *Nicotiana* and the bryophyte *Marchantia*. This prescient paper showed that at 17.9% GC, *ycf2* is the most AT-rich gene in the *Marchantia* plastome and that it shares only local homology to *Nicotiana*. A 1994 analysis (7) of the five *ycf2* sequences then available (from *Marchantia* and four angiosperms) extended these conclusions, showing that indel rates are extremely high in YCF2, with gaps occurring at 39% of alignment positions despite little variation in overall protein size (2109-2280 residues), and that amino-acid identity between *Marchantia* and the angiosperms is extremely low (27% overall) and barely greater than expected for random sequences over the first 1000-1500 residues of the protein. These observations led to the conclusion that the protein is under selective constraint in spite of its high divergence and AT bias (7). Thus, if any plastid gene could be expected to be as divergent and AT-rich as the *Balanophora* ORF, it is *ycf2*.

Fourth, although *ycf2* is extraordinarily AT-rich in *Balanophora*, there is evidence for weak purifying selection operating on it within the genus (d_N/d_S = 0.82 for the alignment shown in Fig. S4). Although this is the highest d_N/d_S for the *Balanophora* plastome genes, weak constraint on *ycf2* is typical in land plants (8-12). Furthermore, given its nearly 98% AT composition, if *Balanophora ycf2* wasn't under selective constraint, it should be riddled with stop codons and frameshifts as is expected for pseudogenes and non-coding DNA, in general. This expectation is met for non-coding plastome DNA in *Balanophora* as the 751 bp of spacer and intronic DNA in *B. laxiflora* contains an average of 27 stop codons across the six potential open reading frames (most are TAA codons, of course) (Table S4). In contrast, there is but a single canonical stop (an in-frame TAG in *B. reflexa ycf2*) in the 750 and 771 bp of *Balanophora laxiflora* and *B. reflexa ycf2* sequence, respectively. This single TAG codon, as explained in the main text, does not function as a stop codon but has been reassigned in the *Balanophora* plastome to encode Trp. Thus, in spite of the relatively high d_N/d_S estimate, *ycf2* in *Balanophora* is likely functional.

These four sets of considerations lead us to conclude that the *Balanophora* plastome contains an extraordinarily divergent but likely functional *ycf2* gene. Although the roughly 8-fold shrinkage of *ycf2* in *Balanophora,* is at first reaction very surprising, this must be viewed in the perspective of its plastid genome overall. As shown in Fig. S4 and Table S1, almost all *Balanophora* plastid protein genes are shorter than normal, with *accD*, *ycf1*, and *rps18* ranging from 38 to 53% shorter than homologs in *Schoepfia*, a hemiparasitic relative of *Balanophora*. Viewed this way, *Balanophora ycf2* is simply the most extreme point on a more-or-less continuum of divergence in length, sequence, and base composition, within this highly aberrant plastome.

Stop codon and TGG usage in the *Cytinus* plastome. We analyzed the published genome of *Cytinus hypocystis* [GenBank # KT335971, (13)] and found that its 16 protein genes all use TAA as a stop codon, i.e., there are no TAG- or TGA stops in this plastome (Table S8). Inspection of amino-acid alignments revealed, in stark contrast to the *Balanophora* situation, an absence of internal TAG (or TGA) codons and the presence of 22 internal TGG codons, all but one of which are located at sites at which TGG (tryptophan in the canonical code) is present in most or all of the diverse land plants in the alignments (Table S8). Therefore, as described in the main text, *Cytinus* appears to still use TGG as a Trp codon despite possessing the antecedent condition for a code change in which TAG (or TGA) has been reassigned from stop to Trp.

SI References

1. Naumann J*, et al.* (2016) Detecting and characterizing the highly divergent plastid genome of the nonphotosynthetic parasitic plant *Hydnora visseri* (Hydnoraceae). *Genome Biol Evol* 8:345-363.
2. Su XZ, Wu Y, Sifri CD, Wellems TE (1996) Reduced extension temperatures required for PCR amplification of extremely A+T-rich DNA. *Nucleic Acids Res* 24:1574-1575.
3. Wolfe KH (1994) Similarity between putative ATP-binding sites in land plant plastid ORF2280 proteins and the FtsH/CDC48 family of ATPases. *Curr Genet* 25:379-383.
4. Drescher A*, et al.* (2000) The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. *Plant J* 22:97-104.
5. Wicke S*, et al.* (2011) The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. *Plant Mol Biol* 76:273-297.
6. Shimada H, Sugiura M (1991) Fine structural features of the chloroplast genome: comparison of the sequenced chloroplast genomes. *Nucleic Acids Res* 19:983-995.
7. Downie SR*, et al.* (1994) Structure and evolution of the largest chloroplast gene (ORF2280): internal plasticity and multiple gene loss during angiosperm evolution. *Curr Genet* 25:367-378.
8. Sloan DB*, et al.* (2014) A recurring syndrome of accelerated plastid genome evolution in the angiosperm tribe Sileneae (Caryophyllaceae). *Mol Phylogenet Evol* 72:82-89.
9. Wicke S, Schaferhoff B, dePamphilis CW, Muller KF (2014) Disproportional plastome-wide increase of substitution rates and relaxed purifying selection in genes of carnivorous Lentibulariaceae. *Mol Biol Evol* 31:529-545.

10. Barrett CF, Davis JI (2012) The plastid genome of the mycoheterotrophic *Corallorhiza striata* (Orchidaceae) is in the relatively early stages of degradation. *Am J Bot* 99:1513-1523.
11. Raman G, Park S (2015) Analysis of the complete chloroplast genome of a medicinal plant, *Dianthus superbus* var. *longicalyncinus*, from a comparative genomics perspective. *PLoS One* 10:e0141329.
12. Zhang H, Li C, Miao H, Xiong S (2013) Insights from the complete chloroplast genome into the evolution of *Sesamum indicum* L. *PLoS One* 8:e80508.
13. Roquet C*, et al.* (2016) Understanding the evolution of holoparasitic plants: the complete plastid genome of the holoparasite *Cytinus hypocistis* (Cytinaceae). *Ann Bot* 118:885-896.
14. Kohchi T, et al. (1988) A nicked group II intron and trans-splicing in liverwort, Marchantia polymorpha, chloroplasts. Nucleic Acids Res 16:10025-10036.
15. Akashi H, Gojobori T (2002) Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. Proc Natl Acad Sci USA 99:3695-3700.
16. Sullivan MJ, Petty NK, Beatson SA (2011) Easyfig: a genome comparison visualizer. *Bioinformatics* 27:1009-1010.
17. Giege R, Sissler M, Florentz C (1998) Universal rules and idiosyncratic features in tRNA identity. *Nucleic Acids Res* 26:5017-5035.
18. Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 25:955-964.
19. Juhling F*, et al.* (2012) Improved systematic tRNA gene annotation allows new insights into the evolution of mitochondrial tRNA structures and into the mechanisms of mitochondrial genome rearrangements. *Nucleic Acids Res* 40:2833-2845.
20. Krumsiek J, Arnold R, Rattei T (2007) Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* 23:1026-1028.
21. Howe CJ, Smith A (1991) Plants without chlorophyll. *Nature* 349:109-109.
22. Barbrook AC, Howe CJ, Purton S (2006) Why are plastid genomes retained in non-photosynthetic organisms? *Trends Plant Sci* 11:101-108.
23. Wolfe KH, Morden CW, Palmer JD (1992) Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. *Proc Natl Acad Sci USA* 89:10648-10652.

Table S1. Length and GC content of plastid genes in *Balanophora* and *Schoepfia*

	accD	*clpP*	*ycf1*	*ycf2*	*rps2*	*rps3*	*rps4*	*rps7*	*rps11*	*rps12*	*rps14*	*rps18*	*rps19*	*rpl2*	*rpl14*	*rrn4.5*	*rrn16*	*rrn23*	*trnE*	CDS*	Full length
Length (bp)																					
S. jasminodora	1,485	591	5,490	6,777	711	657	606	468	417	372	303	357	279	825	369	103	1,491	2,810	73	58,791	118,743
B. laxiflora	894	600	2,943	750	558	624	588	429	372	366	210	171	249	750	357	88	1,474	2,930	75	9,861	15,505
B. reflexa	951	606	2,691	771	549	624	525	408	372	372	204	165	219	768	345	90	1,573	3,052	71	9,570	15,507
% reduction†	37.9	-	48.7	88.8	22.2	5.0	8.2	10.6	10.8	0.8	31.7	52.9	16.1	8.0	4.9	13.6	-	-	-	83.5	86.9
GC content (%)																					
S. jasminodora	34.8	40.6	30.4	37.0	37.7	33.9	39.1	40.2	44.1	41.4	42.2	34.2	34.8	43.0	37.1	53.4	56.5	55.0	57.5	38.0	38.1
B. laxiflora	17.6	19.7	4.9	2.0	4.8	5.6	9.0	5.4	11.8	20.2	8.1	4.1	6.4	14.3	10.4	13.6	24.0	19.9	29.3	8.9	12.2
B. reflexa	16.2	19.5	5.2	2.2	4.0	5.0	6.1	4.9	11.6	19.6	6.9	4.2	7.3	13.9	10.4	10.0	21.0	18.6	31.0	8.7	11.6

*CDS: Protein coding sequences. Note that these values exceed the actual total number of protein-coding nucleotides in these plastomes as a consequence of gene overlaps (see main text and Table S4).

†(1 - (average length of *Balanophora*/length of *Schoepfia*)) x 100%. Dashes indicate the three genes that are larger in *Balanophora*than in *Schoepfia*.

Table S2. Plastid gene divergence between *B. laxiflora* and *B. reflexa*[*]

	accD	*clpP*	*ycf1*	*ycf2*	*rps2*	*rps3*	*rps4*	*rps7*	*rps11*	*rps12*	*rps14*	*rps18*	*rps19*	*rpl2*	*rpl14*	*rrn4.5*	*rrn16*	*rrn23*	*trnE*
% nucleotide identity	93.6	93.5	80.2	75.2	79.9	86.5	81.7	85.8	90.2	93.1	79.8	88.1	91.4	88.6	91.2	94.1	90.5	87.9	95.8
% amino acid identity	89.6	89.5	62.4	51.7	65.7	72.7	68.2	75.2	82.0	90.0	56.1	77.4	82.9	80.4	82.5	-	-	-	-
Pairwise d_N[†]	0.06	0.05	0.22	0.32	0.22	0.14	0.20	0.15	0.07	0.05	0.23	0.11	0.08	0.09	0.08	-	-	-	-
Pairwise d_S[†]	0.40	0.49	0.50	0.39	0.43	0.30	0.28	0.30	0.35	0.52	0.36	0.31	0.25	0.77	0.20	-	-	-	-
Pairwise d_N/d_S	0.14	0.09	0.44	0.82	0.52	0.48	0.72	0.49	0.20	0.09	0.65	0.38	0.32	0.11	0.40	-	-	-	-

*Gaps were excluded from all identity calculations.

[†]TAG codons were excluded from the d_N and d_s calculations.

Table S3. Annotated start and stop codons in *Balanophora* plastomes

	Start codon*		Stop codon*	
Gene	*B. laxiflora*	*B. reflexa*	*B. laxiflora*	*B. reflexa*
accD	ATG	ATG	TAA	TAA
clpP	ATG	ATG	TAA	TAA
rpl14	ATG	ATG	TAA	TAA
rpl2	ATG	ATG	**TGA**	**TGA**
rps11	ATG	ATG	TAA	TAA
rps12	ATG	ATG	TAA	TAA
rps14	ATG	ATG	TAA	TAA
rps18	ATG	ATG	TAA	TAA
rps19	ATG	ATG	TAA	TAA
rps2	ATG	ATG	TAA	TAA
rps3	ATG	ATG	TAA	TAA
rps4	**ATA**	**ATT**	TAA	TAA
rps7	ATG	ATG	TAA	TAA
ycf1	ATG	ATG	TAA	TAA
ycf2	ATG	**ATA**	TAA	TAA

*non-ATG start codons and non-TAA stop codons are in bold.

Table S4. Intergenic regions in *Balanophora* plastomes

Region	Size*	%GC	Note	Reading frame†						Sequence
	(bp)			1	2	3	–1	–2	–3	
B. laxiflora										
rrn4.5-ycf1	219	7.8		9/1/0	7/1/0	6/0/0	8/1/2	7/0/0	1/1/0	AATTAAAATAATAATATACTATATGTTTTATTATATATTTTGGTATTTTAA AAATAAAAAATATACATTAAAAAATAAAAATCTATTTCAAATTAAATGG TAAAAATTTATTATATATAATTATTATATAATTAAAAAAATAATTTAATATC AAAATTATAAATACTTTTAGTATAGTATAAAATAATAAAAAGGAGGTTA ATAATTTAAAAAATAAAA
ycf1-rpl14	(14)	-	gene overlap	-	-	-	-	-	-	
rpl14-rps2	1	0.0		-	-	-	-	-	-	T
rps2-trnE	12	0.0		0/0/0	0/0/0	1/0/0	0/0/0	1/0/0	1/0/0	ATTTTTAATTAT
trnE-rps14	20	0.0		1/0/0	0/0/0	0/0/0	1/0/0	0/0/0	1/0/0	TTTTATTTATATATATAAAT
rps14-rps4	(7)	-	gene overlap	-	-	-	-	-	-	
rps4-accD	83	2.4		3/0/0	5/0/0	1/0/1	3/0/0	2/0/0	7/0/0	ATATTATTTATTTAATTAAATTAATAAATAATAAAAATAAATGATTATTCT TTTTTATTAATTTTATTTAATTTTTTATTATT
accD-rps18	5	0.0		0/0/0	0/0/0	0/0/0	1/0/0	0/0/0	0/0/0	TTTTA
rps18-rps12_5′	164	6.7	trans-spliced intron excluded‡	7/1/1	3/0/0	2/0/0	5/0/1	3/0/0	6/2/0	AATTAATTATAGTATTTAATATATATTTAATATTTAAAATCCAATATAAAA TATATATTTTTCTTTTTGCTACTATAATATATTCATTTTTTATTTTTTATAA ACATAATAAATTTATATTATTATTTATAATTTATTTAAATTTTGAGTATTT TTATAAATATT
rps12-clpP	13	0.0		1/0/0	1/0/0	0/0/0	2/0/0	0/0/0	0/0/0	TAATTTATAATTA
clpP-rps11	60	0.0		3/0/0	0/0/0	5/0/0	4/0/0	3/0/0	3/0/0	AATAATTAATATTTATTTAATAATTAATATTTATTTAATAATTAATATTTAT TTATATTA
rps11-rps3	8	0.0		0/0/0	0/0/0	0/0/0	0/0/0	0/0/0	1/0/0	TTATTTTT
rps3-rps19	3	0.0		0/0/0	-	-	1/0/0	-	-	AAT
rps19-rpl2	(4)	-	gene overlap	-	-	-	-	-	-	
rpl2-ycf2	23	0.0		3/0/0	0/0/0	0/0/0	0/0/0	1/0/0	0/0/0	TAATAAAAATAAATTATATAAAA
ycf2-rps7	2	0.0		-	-	-	-	-	-	AT
rps7-rps12_3′	(8)	-	gene overlap	-	-	-	-	-	-	
rps12_3′-rrn16	47	4.3	trans-spliced intron excluded‡	3/0/0	1/0/0	1/0/0	1/0/0	1/1/0	2/0/0	TTATAAGAATTAAAAAAAAAAAAATAAAACTATAAATTATTTTATAAT
rrn16-rrn23	63	1.6		1/0/0	3/0/0	0/0/0	5/0/0	1/0/0	5/0/0	TTTTTATTTTTTTTAAAAAATTAAATTATTATTATATTTATTATATTATTAA TATTAAAATTC
rrn23-rrn4.5	28	3.6		2/0/0	0/0/0	2/0/0	2/0/0	1/0/0	0/0/0	TAATTATAAATTAAAAAAGATTTTAATA
B. reflexa										
rrn4.5-ycf1	213	7.0		8/0/0	9/0/1	6/2/0	7/1/0	3/0/0	5/0/0	TTGATAAATAGAAGAATATATTTTAATATTTTAAAATATATAAAAAATAC ATTAAATAAATTACTATTTTAAATTTAATGGTAAATATATATAATATTTAT TATATAATTATTAAATAAAAAATAATATAATATTAAAGTAAATATATAAAC TTATAGTATGTATTTAATTAATATAAAATTATAAAAAAAAAATTGGAGG ATATAAAAAAAA
ycf1-rpl14	52	0.0		2/0/0	0/0/0	5/0/0	2/0/0	0/0/0	4/0/0	TATAATAAAAATTATTATTTTAAATTTAATATTTAATATTATTTTTAAAAA T
rpl14-rps2	10	0.0		0/0/0	0/0/0	0/0/0	0/0/0	0/0/0	1/0/0	AAATTTTATT
rps2-trnE	70	0.0		4/0/0	4/0/0	3/0/0	2/0/0	4/0/0	3/0/0	TTATAATTTTAAAAAATATTAATATTTTTAATATTAATTAATTAAAATAAA ATATTTAAATTTTAAATAA
trnE-rps14	42	0.0		1/0/0	1/0/0	3/0/0	2/0/0	1/0/0	2/0/0	TTTTTTAAAATATTATATAATTTTAATTAATTTTTAATATAT
rps14-rps4	16	6.3		0/0/0	1/0/0	0/0/0	0/0/0	0/0/0	0/0/0	AAAAAACAAATAATAT
rps4-accD	92	5.4		4/0/0	8/0/1	0/0/0	3/0/0	1/0/0	7/0/0	TTAATATATTTATTATAATAAATTTAATTAATGATTATCCCTTAATTCTTA TATATATTAAATATAATAATATATTATATAATATTAATAAA
accD-rps18	(15)	-	gene overlap	-	-	-	-	-	-	
rps18-rps12_5′	178	2.8	trans-spliced intron excluded‡	4/0/0	5/0/0	4/0/0	4/0/0	7/0/1	9/0/0	TATTTAATTTTTTATTAAATTTATTTATATTTTTTAATCCAATATAAAATAT ATTAATATATTAATAAATATATATTTTATTTTTGCTTTTATAATATTTTCATT ATTTATTTAATTTATAAATTATTATAATATATATATATATATATATAATTAATAT TATTATATTTTTTTTATATTT
rps12-clpP	(7)	-	gene overlap	-	-	-	-	-	-	
clpP-rps11	27	0.0		0/0/0	0/0/0	2/0/0	0/0/0	0/0/0	2/0/0	AAAAATATATATAAATTTATTAAATAT
rps11-rps3	37	2.7		2/0/0	0/0/0	2/0/0	2/0/0	3/0/0	1/0/0	TTATTATAATATAATAATTAAATTATTATTGTATTAT
rps3-rps19	6	0.0		0/0/0	0/0/0	0/0/0	0/0/0	0/0/0	0/0/0	TATTTT
rps19-rpl2	(4)	-	gene overlap	-	-	-	-	-	-	
rpl2-ycf2	29	10.4		1/0/0	1/0/0	1/0/0	1/0/0	0/1/1	0/1/0	ATAGATATGAATATATATTAATTATTTAG
ycf2-rps7	(12)	-	gene overlap	-	-	-	-	-	-	
rps7-rps12_3′	(8)	-	gene overlap	-	-	-	-	-	-	
rps12_3′-rrn16	53	1.9	trans-spliced intron excluded‡	2/0/0	1/0/0	2/0/0	2/0/0	0/0/0	4/0/0	TTATAAGAAAAAATATATTATATATAAATATTAATTTATAAATTTTTATTAA T
rrn16-rrn23	53	3.8		1/0/0	4/0/0	1/0/0	0/0/1	2/1/0	2/0/1	TTAATAAATAAAATTATTATTTTAAAATCATAAATAAAAAAAAATTTTCA ATA
rrn23-rrn4.5	18	11.1		0/0/0	1/0/0	1/0/0	1/0/0	0/0/0	1/0/0	AATTTTAGGATTTTAATA

*Numbers in parentheses are the lengths of gene overlaps.

†The values are the number of in-frame TAA/TAG/TGA codons in the sequence shown in the rightmost column.

‡The regions between *rps18* and the 5’ exon of *rps12* and between *rrn*16 and the 3’ exon of *rps12* contain the two portions of the *rps12* trans-spliced intron. We conservatively (14) assigned as intronic the 150 bp of sequence in these regions that aligns between the two *Balanophora* plastomes and which abuts the 5’ and 3’ exons of *rps12*. Owing to indels, the actual lengths of the regions assigned as intronic range from 150 to 180 bp.

Table S5. GC content and codon usage for the 30 genomes analyzed in Fig. S5

Species*	Group	Genome	No. of codons‡	% GC content‡			No. of unused codons‡
				Genome†	Protein coding	3rd pos. synony.	
Balanophora laxiflora	Angiosperm - holoparasite	Plastid	3287	12.2	8.9	1.1	21
Balanophora reflexa	Angiosperm - holoparasite	Plastid	3190	11.6	8.7	1.2	20
Plasmodium falciparum	Apicomplexan - parasite	Plastid	7418	13.1	11.0	2.1	11
Zinderia insectifolia	Proteobacterium	Bacterial	62769	13.5	13.2	2.1	1
Babesia microti	Apicomplexan - parasite	Plastid	6976	14.1	12.1	2.3	7
Carsonella ruddii	Proteobacterium	Bacterial	51205	14.0	13.3	2.6	2
Eimeria tenella	Apicomplexan - parasite	Plastid	6891	18.6	15.4	2.8	1
Toxoplasma gondii	Apicomplexan - parasite	Plastid	6158	19.3	16.4	3.9	3
Diadegma semiclausum	Animal - insect - parasite	Mitochondrial	3709	12.6	16.3	4.4	4
Nasuia deltocephalinicola	Proteobacterium	Bacterial	43435	15.2	14.3	4.4	1
Monosiga brevicollis	Choanoflagellate	Mitochondrial	8323	14.0	22.1	4.9	4
Nitzschia spp.	Diatom - nonphotosynthetic	Plastid	15673	21.9	20.4	5.3	0
Ichthyophthirius multifiliis	Ciliate - parasite	Mitochondrial	12119	16.4	15.5	5.7	0
Acrasis kona	Excavate	Mitochondrial	14149	16.8	16.0	5.9	3
Choreocolax polysiphoniae	Red alga - parasite	Plastid	19129	20.5	22.7	5.9	1
Rozella allomycis	Fungus - Cryptomycota	Mitochondrial	1926	14.5	15.1	6.0	11
Radopholus similis	Animal - nematode	Mitochondrial	3376	14.6	15.3	6.8	10
Saccharomyces cerevisiae	Fungus - yeast	Mitochondrial	2218	14.1	24.5	7.9	11
Euglena longa	Euglenid - nonphotosynthetic	Plastid	10707	20.2	21.7	9.0	0
Nakaseomyces bacillisporus	Fungus - yeast	Mitochondrial	2193	10.9	23.8	9.2	19
Helicosporidium sp.	Green alga - parasite	Plastid	9724	26.9	25.0	10.3	0
Hydnora visseri	Angiosperm - holoparasite	Plastid	5708	23.7	20.4	10.8	1
Pilostyles hamiltonii	Angiosperm - holoparasite	Plastid	995	22.7	23.2	12.9	6
Cytinus hypocistis	Angiosperm - holoparasite	Plastid	2230	29.9	29.7	13.1	3
Bulboplastis apyrenoidosa	Red alga - photosynthetic	Plastid	41592	23.3	29.9	13.0	0
Pilostyles aethiopica	Angiosperm - holoparasite	Plastid	926	24.2	24.6	14.7	8
Cynomorium coccineum	Angiosperm - holoparasite	Plastid	5515	29.9	26.0	14.7	1
Monotropa uniflora	Angiosperm - mycoheterotroph	Plastid	5123	28.0	27.9	15.3	0
Thismia tentaculata	Angiosperm - mycoheterotroph	Plastid	1027	26.6	26.2	16.8	1
Polytoma uvella	Green alga - nonphotosynthetic	Plastid	13501	23.5	39.2	35.5	0

*See Table S12 for strain number, accession number, and other information.

†The "genome" GC values include only one copy of the large, usually perfect repeats present in many plastomes, as these almost always contain rRNA genes, whose relatively GC richness will bias the full-genome GC values, especially for highly reduced genomes.

‡Protein gene duplicates were removed in the calculations for these three sets of values.

Table S6. Plastid codon usage in *Balanophora* and *Nicotiana*

Amino acid	Codon	Number of codons*			Percent codon usage*			Relative synonymous codon usage (RSCU)*		
		Nicot.	*B. lax.*	*B. ref.*	*Nicot.*	*B. lax.*	*B. ref.*	*Nicot.*	*B. lax.*	*B. ref.*
Phe	TTT	994	275	255	3.57	8.37	7.99	1.27	1.99	2.00
	TTC	575	1	0	2.07	0.03	0.00	0.73	0.01	0.00
Leu	TTA	904	304	289	3.25	9.25	9.06	1.83	5.77	5.80
	TTG	605	2	1	2.17	0.06	0.03	1.22	0.04	0.02
	CTT	644	7	7	2.31	0.21	0.22	1.30	0.13	0.14
	CTC	231	0	0	0.83	0.00	0.00	0.47	0.00	0.00
	CTA	397	3	2	1.43	0.09	0.06	0.80	0.06	0.04
	CTG	189	0	0	0.68	0.00	0.00	0.38	0.00	0.00
Ile	ATT	1119	288	284	4.02	8.76	8.90	1.44	1.33	1.30
	ATC	479	0	1	1.72	0.00	0.03	0.62	0.00	0.00
	ATA	727	363	368	2.61	11.04	11.54	0.94	1.67	1.69
Met	ATG	665	33	29	2.39	1.00	0.91	1.00	1.00	1.00
Val	GTT	543	8	13	1.95	0.24	0.41	1.43	1.19	1.73
	GTC	199	1	0	0.72	0.03	0.00	0.52	0.15	0.00
	GTA	569	18	17	2.04	0.55	0.53	1.49	2.67	2.27
	GTG	213	0	0	0.77	0.00	0.00	0.56	0.00	0.00
Ser	TCT	630	41	41	2.26	1.25	1.29	1.73	2.08	2.28
	TCC	351	0	0	1.26	0.00	0.00	0.96	0.00	0.00
	TCA	429	48	49	1.54	1.46	1.54	1.18	2.44	2.72
	TCG	223	0	0	0.80	0.00	0.00	0.61	0.00	0.00
	AGT	424	29	18	1.52	0.88	0.56	1.17	1.47	1.00
	AGC	126	0	0	0.45	0.00	0.00	0.35	0.00	0.00
Pro	CCT	448	26	27	1.61	0.79	0.85	1.52	2.26	2.51
	CCC	223	1	0	0.80	0.03	0.00	0.76	0.09	0.00
	CCA	345	19	16	1.24	0.58	0.50	1.17	1.65	1.49
	CCG	161	0	0	0.58	0.00	0.00	0.55	0.00	0.00
Thr	ACT	547	23	21	1.97	0.70	0.66	1.56	1.77	1.87
	ACC	269	0	0	0.97	0.00	0.00	0.77	0.00	0.00
	ACA	433	29	24	1.56	0.88	0.75	1.23	2.23	2.13
	ACG	154	0	0	0.55	0.00	0.00	0.44	0.00	0.00
Ala	GCT	632	17	13	2.27	0.52	0.41	1.77	2.34	1.86
	GCC	252	0	0	0.91	0.00	0.00	0.71	0.00	0.00
	GCA	404	12	15	1.45	0.37	0.47	1.13	1.66	2.14
	GCG	141	0	0	0.51	0.00	0.00	0.39	0.00	0.00
Tyr	TAT	791	304	321	2.84	9.25	10.06	1.60	1.97	1.97
	TAC	199	5	5	0.72	0.15	0.16	0.40	0.03	0.03
TER	TAA	51	14	14	0.18	0.43	0.44	1.56	1.87	1.87
	TGA	23	1	1	0.08	0.03	0.03	0.70	0.13	0.13
	TAG†	24	18	16	0.09	0.55	0.50	0.73	2.00	2.00
Trp	TGG	496	0	0	1.78	0.00	0.00	1.00	0.00	0.00
His	CAT	507	21	22	1.82	0.64	0.69	1.55	2.00	2.00
	CAC	149	0	0	0.54	0.00	0.00	0.45	0.00	0.00
Gln	CAA	735	44	42	2.64	1.34	1.32	1.49	2.00	1.95
	CAG	249	0	1	0.89	0.00	0.03	0.51	0.00	0.05
Asn	AAT	1052	488	447	3.78	14.85	14.01	1.53	1.99	1.99
	AAC	327	3	2	1.18	0.09	0.06	0.47	0.01	0.01
Lys	AAA	1106	568	567	3.97	17.28	17.77	1.48	1.98	1.99
	AAG	389	6	4	1.40	0.18	0.13	0.52	0.02	0.01
Asp	GAT	903	40	36	3.25	1.22	1.13	1.60	2.00	1.95
	GAC	226	0	1	0.81	0.00	0.03	0.40	0.00	0.05
Glu	GAA	1081	67	62	3.88	2.04	1.94	1.48	2.00	1.97
	GAG	380	0	1	1.37	0.00	0.03	0.52	0.00	0.03
Cys	TGT	234	23	22	0.84	0.70	0.69	1.46	2.00	1.91
	TGC	86	0	1	0.31	0.00	0.03	0.54	0.00	0.09
Arg	CGT	345	15	15	1.24	0.46	0.47	1.22	2.20	2.09
	CGC	106	1	0	0.38	0.03	0.00	0.38	0.15	0.00
	CGA	409	1	1	1.47	0.03	0.03	1.45	0.15	0.14
	CGG	130	0	0	0.47	0.00	0.00	0.46	0.00	0.00
	AGA	513	24	27	1.84	0.73	0.85	1.82	3.51	3.77
	AGG	192	0	0	0.69	0.00	0.00	0.68	0.00	0.00
Gly	GGT	587	40	41	2.11	1.22	1.29	1.25	1.67	1.78
	GGC	211	0	0	0.76	0.00	0.00	0.45	0.00	0.00
	GGA	752	55	50	2.70	1.67	1.57	1.60	2.29	2.17
	GGG	328	1	1	1.18	0.03	0.03	0.70	0.04	0.04

*Red indicates values of zero.

†TAG is used as Trp in *Balanophora* plastomes.

Table S7. Nitrogen usage and energy cost as a function of amino acid frequency in the *B. laxiflora* and *Nicotiana* plastomes

Amino acid	Nitrogen atoms*	Energy cost†	% frequency‡	
			B. laxifora	*Nicotiana*
Phe	1	52.0	**8.4**	5.6
Leu	1	27.3	9.7	10.7
Ile	1	32.3	**19.9**	8.4
Met	1	34.3	1.0	2.4
Val	1	23.3	0.8	5.5
Ser	1	11.7	3.6	7.8
Pro	1	20.3	1.4	4.2
Thr	1	18.7	1.6	5.0
Ala	1	11.7	0.9	5.1
Tyr	1	50.0	**9.4**	3.6
His	3	38.3	0.6	2.4
Gln	2	16.3	1.3	3.5
Asn	2	14.7	**15.0**	5.0
Lys	2	30.3	**17.5**	5.4
Asp	1	12.7	1.2	4.1
Glu	1	15.3	2.0	5.3
Cys	1	24.7	0.7	1.2
Trp	2	74.3	0.6	1.8
Arg	4	27.3	1.3	6.1
Gly	1	11.7	2.9	6.7
N index§			1.4	1.4
E index¶			29.5	25.0

*Number of nitrogen atoms present in each amino acid.

†The energetic cost of synthesizing each amino acid, which ranges from 12 to 74 high-energy phosphate bonds (15).

‡Boldface indicates amino acids used at a higher frequency in *B. laxiflora* than in *Nicotiana.*

§The N index is the sum of the products of the number of nitrogen atoms in each amino acid and the % frequency of that amino acid.

¶The E index is the sum of the products of the energy cost of each amino acid and the % frequency of that amino acid.

Table S8. Stop and Trp codons in the 16 annotated and putatively functional protein genes in the *Cytinis* plastome.

Gene	Stop codon	# Trp codons	
		Total	Conserved*
*clp*P	TAA	2	2
*rpl*2	TAA	2	2
*rpl*14	TAA	0	0
*rpl*16	TAA	3	3
*rpl*20	TAA	1	1
*rpl*22	TAA	0	0
*rpl*36	TAA	0	0
*rps*2	TAA	4	4
*rps*3	TAA	3	3
*rps*4	TAA	0	0
*rps*7	TAA	1	1
*rps*8	TAA	1	1
*rps*11	TAA	2	1
*rps*12	TAA	0	0
*rps*14	TAA	2	2
*rps*19	TAA	1	1

*Trp residues in *Cytinus* that are located at sites at which Trp is present in most or all of the diverse land plants in the amino-acid alignments on which this table is based.

Table S9. Features of highly compact plastomes*

Taxon	Group	Lifestyle	Intergenic spacer (%)	Overlapping protein genes†		Shrunken proteins	No. of cis-spliced introns	%GC
				% overlap‡	No. of genes			
Balanophora laxiflora	angiosperms	parasite	4.8	53.3	15	yes	0	12.2
Helicosporidium sp.	green algae	parasite	4.8	34.7	23	?	1	26.9
Balanophora reflexa	angiosperms	parasite	5.8	66.7	15	yes	0	11.6
Cyanidioschyzon merolae	red algae	photosynthetic	5.9	40.0	197	?	0	37.6
Babesia microti§	apicomplexans	parasite	6.2	30.0	30	?	0	14.1
Prototheca zopfii	green algae	parasite	6.6	10.5	19	?	0	27.0
Cynomorium coccineum	angiosperms	parasite	7.2	22.2	18	?	4	29.9
Sciaphila thaidanica	angiosperms	mycoheterotroph	8.1	78.6	14	?	2	30.5
Hydnora visseri	angiosperms	parasite	11.6	23.5	17	yes	2	23.7
Dictyopteris divaricata	brown algae	photosynthetic	13.1	11.6	138	?	0	30.7
Lepidodinium chlorophorum	dinoflagellates	photosynthetic	13.3	34.9	63	?	3	34.6

*Plastomes were included if they have ≥ 10% overlapping protein genes and/or ≤ 10% intergenic spacer DNA.

†We followed the standard practice for bacterial genomes and included only protein genes in the overlap analyses (overlaps with rRNA and tRNA genes are extremely rare).

‡The percentage of all protein genes that overlap at one or both ends with another protein gene.

§*Babesia* was selected to represent the many apicomplexan plastomes that meet our inclusion criteria. All other plastomes that meet these criteria are included in the table.

Table S10. PCR primers used in this study

Name	Sequence
A . Primers for *B. laxiflora* plastome validation and RT-PCR analysis	
Blax accD 1F*	ATGAATATTTGTGAACAATGTG
Blax accD 1R*	AAAAGCTATATATGTATTTGGTTC
Blax accD 2F	ATGTTTATAAAAAAATAGTATTTAAATTAT
Blax accD 2R	CTTGGTACAATAATATCAAATATTC
Blax clpP 1F	TCTTCAGCTTCAAGTATAAATTC
Blax clpP 1R*	ATGCAATATATAAAACCTAATATACG
Blax clpP 2F*	CAATTATACCATAATTTTTAGCTTC
Blax clpP 2R	ATGCCCATAGGTATTCCT
Blax rpl14 1F*	ATATTTTAATATAACTGATAATACAGG
Blax rpl14 1R	TTTTTACTTTTTGGATTATTTTC
Blax rpl2 1F	TTATAACCCTAAAAAGTAGATGG
Blax rpl2 1R	ATTAGTAAAATCAGCAGGATG
Blax rps11 1F	AAAATATCTCTACCAAAATTTAATC
Blax rps11 1R*	GAAGTTATTTATTTTTCTTCTTCTG
Blax rps12-5' 1F	GTCCTCAAAGAAAAGGAATTTG
Blax rps12-3' 1R	ACCAAATTCTGCTTTACG
Blax rps12-3' 2R*	CTCCATTTGTATCTAAAACACCTC
Blax rps3 1F	TTTAATACTTTTTAATGGAATTTTAC
Blax rps3 1R	ATGATAAATAAAATAAATCCAAT
Blax rps4 1F	TAATGGATAATAAATTAAATATTTTGG
Blax rps4 1R*	ATTTCAACAATTCCTGCATC
Blax rrn16 1F*	TCAGGATTAACGCTTGTG
Blax rrn16 1R*	TTATATTCACCACAGTATATCTTACC
Blax rrn23 1F	AATATATAACATAAATCTAAAAATTTCC
Blax rrn23 1R	TTTTTACCTATTATCTATCAATTATTC
Blax rrn23 2F	AAAGTAAAAATATAAAAACAATTAGAAG
Blax rrn23 2R	AATAATTTTATTACTTAATATTCTTTCAG
Blax ycf1 1F*	GTATTATTTTTGGTTATTTTTTTAG
Blax ycf1 1R*	TATCTTTATATTTTCTTGAACG
Blax ycf1 2R	TAATATTTTAATTTTTATACCTTCC
B. Primers for *B. reflexa* plastome validation	
Bref ycf1 F1	TTTTGGAAAAA ATATATACCTAATTT T
Bref ycf1 F2	GAACAA CAA GAA AAT GAG GAA
Bref ycf1 R1	AAAATTAGGTA ATATATTTTTCCAAA A
Bref ycf1 R2	TATATCATGATCAAAATCTGATTG
Bref rpl14 F1	TATTTTAATATAATTGATAATACAGGGA
Bref rpl14 R1	AATAATTGCAGTATTTTTACTATATTT
Bref rps4 F1	TAATGGATAATAAATAAAATAATTTGAA
Bref rps4 R1	TATTTCAACAATTTCTGCATCT
Bref rps18intron R1	GAAAATATTATAAAAGCAAAATAAAATAT
Bref accD F1	GTGAACCAAATACATATATAGCATT
Bref clpP R1	GAATTTTTATTAGAAGCTGAAGA
Bref rps11 F1	AACTTCTCCTATTAAATTAGTAAC
Bref rpl2 F1	CTATTTGGATCATATTCTATTGC
Bref rpl2 R1	TTACAAAAAAGGTAAAAATTCAT
Bref rpl2 F2	ATTTTTATGACCTATATTTCTATATTTTA
Bref rps19 R1	TGGAGGAGGTAAAGGAAA
Bref rrn16 F1	TTACTTTTCCACCTCTAACTAAAAC
Bref rrn16 F2	CGGATAATCAACCACACTGAGA
Bref rrn16 R1	TCATATCATAAGAGGTGTTTTAGA
Bref rrn16 R2	TCTCAGTGTGGTTGATTATCCG
Bref rrn23 1F1	GTCAAGTCATTATGCTCTTT
Bref rrn23 F2	AACATAAAATTAAAAATTCCCGAA
Bref rrn23 F3	AGTCAAGACTTAAGATTTATTCAAA
Bref rrn23 1R1	CGATTTATCTACCTGTATTGG
Bref rrn23 R2	AATTTAGATTTTTTAGTATATATTAGCT

* Primers also used for *B. reflexa* plastome validation

Table S11. PCR primer pairs and temperatures used for *Balanophora* plastome validation and RT-PCR analysis of *B. laxiflora*

Forward Primer	Reverse Primer	PCR temperatures (melt, anneal, extend)
A. Primers for *B. laxiflora* plastome validation		
Blax accD 1F	Blax rps11 1R	95.0, 49.0, 63.0
Blax accD 1F	Blax rps12-3' 1R	95.0, 49.0, 63.0
Blax accD 1F	Blax rps3 1R	95.0, 49.0, 63.0
Blax rpl14 1F	Blax accD 1R	95.0, 49.0, 63.0
Blax rpl2 1F	Blax rps12-3' 1R	95.0, 49.0, 63.0
Blax rpl2 1F	Blax rrn16 1R	95.0, 49.0, 63.0
Blax rpl2 1F	Blax rrn23 1R	95.0, 49.0, 63.0
Blax rpl2 1F	Blax rrn23 2R	95.0, 49.0, 63.0
Blax rps11 1F	Blax rpl2 1R	95.0, 49.0, 63.0
Blax rrn16 1F	Blax rrn23 1R	95.0, 49.0, 63.0
Blax rrn23 1F	Blax ycf1 1R	95.0, 49.0, 63.0
Blax rrn23 2F	Blax rpl14 1R	95.0, 49.0, 63.0
Blax rrn23 2F	Blax ycf1 1R	95.0, 49.0, 63.0
B. Primers for *B. reflexa* plastome validation		
Blax accD1F	Blax clpP 1R	95.0, 49.0, 63.0
Blax clpP 2F	Blax rps11 1R	95.0, 49.0, 63.0
Blax rpl14 1F	Blax rps4 1R	95.0, 49.0, 63.0
Blax rrn16 1F	Blax rrn16 R	95.0, 49.0, 63.0
Blax ycf1 1F	Blax ycf1 1R	95.0, 49.0, 63.0
Bref accD F1	Bref rps18intron R1	95.0, 47.5, 63.0
Bref accD F1	Bref clpP R1	95.0, 47.5, 63.0
Bref rpl14F1	Bref rps4 R1	95.0, 47.5, 63.0
Bref rpl2 F1	Bref rpl2 R1	95.0, 47.5, 63.0
Bref rpl2 F2	Blax rps12-3' 2R	95.0, 47.5, 63.0
Bref rpl2 F2	Bref rrn16 R1	95.0, 47.0, 63.0
Bref rps11 F1	Bref rps19R1	95.0, 47.5, 63.0
Bref rps4 F1	Blax accD 1R	95.0, 47.5, 63.0
Bref rrn16 F1	Bref rrn16 R2	95.0, 47.0, 63.0
Bref rrn16 F2	Blax rrn16 R1	95.0, 48.0, 63.0
Bref rrn23 1F1	Bref rrn23 1R1	95.0, 47.5, 63.0
Bref rrn23 2F1	Blax ycf1 R1	95.0, 47.5, 63.0
Bref rrn23 F2	Bref rrn23 1R1	95.0, 47.0, 63.0
Bref rrn23 F3	Bref rrn23 R2	95.0, 47.0, 63.0
Bref ycf1 F1	Bref ycf1 R2	95.0, 47.0, 63.0
Bref ycf1 F2	Bref rpl14 R1	95.0, 47.5, 63.0
C. RT-PCR amplification of *B. laxiflora*		
Blax accD 1F	Blax accD 1R	95.0, 49,0, 72.0
Blax accD 2F	Blax accD 2R	95.0, 49.0, 63.0
Blax clpP 1F	Blax clpP 1R	95.0, 49,0, 72.0
Blax clpP 2F	Blax clpP 2R	95.0, 49.0, 63.0
Blax rpl14 1F	Blax rpl14 1R	95.0, 49,0, 72.0
Blax rpl2 1F	Blax rpl2 1R	95.0, 49,0, 72.0
Blax rps11 1F	Blax rps11 1R	95.0, 49,0, 72.0
Blax rps12-5' 1F	Blax rps12-3' 2R	95.0, 49.0, 63.0
Blax rps3 1F	Blax rps3 1R	95.0, 49.0, 63.0
Blax rps4 1F	Blax rps4 1R	95.0, 49,0, 72.0
Blax rrn16 1F	Blax rrn16 1R	95.0, 49.0, 63.0
Blax rrn23 1F	Blax rrn23 1R	95.0, 49.0, 63.0
Blax rrn23 2F	Blax rrn23 2R	95.0, 49.0, 63.0
Blax ycf1 1F	Blax ycf1 1R	95.0, 49.0, 63.0
Blax ycf1 1F	Blax ycf1 2R	95.0, 49.0, 63.0

Table S12. Genomes used in this study

Genome	Group	Taxon	Accession	Analysis*					
Plastid	Angiosperm - autotroph	*Amborella trichopoda*	NC_005086.1	P			A		
Plastid	Angiosperm - autotroph	*Arabidopsis thaliana*	NC_000932.1	P			A		
Plastid	Angiosperm - autotroph	*Carica papaya*	NC_010323.1	P			A		
Plastid	Angiosperm - autotroph	*Magnolia denudata*	NC_018357.1	P			A		
Plastid	Angiosperm - autotroph	*Nicotiana tabacum*	NC_001879.2	P	C		A		
Plastid	Angiosperm - autotroph	*Nymphaea alba*	AJ627251.1	P			A		
Plastid	Angiosperm - autotroph	*Oryza sativa* Japonica	NC_001320.1	P			A		
Plastid	Angiosperm - autotroph	*Solanum lycopersicum*	NC_007898.3	P			A		
Plastid	Angiosperm - autotroph	*Vitis vinifera*	DQ424856.1	P			A		
Plastid	Angiosperm - hemiparasite	*Olax imbricata*	KX816863	P			A		
Plastid	Angiosperm - hemiparasite	*Osyris alba*	NC_027960.1					G	
Plastid	Angiosperm - hemiparasite	*Schoepfia jasminodora*	KX775962	P			A	G	
Plastid	Angiosperm - hemiparasite	*Viscum album*	NC_028012					G	
Plastid	Angiosperm - hemiparasite	*Ximenia americana*	GQ997860-GQ997931	P			A		
Plastid	Angiosperm - holoparasite	*Balanophora laxiflora*	KX784265	P	C	B	A	G	H
Plastid	Angiosperm - holoparasite	*Balanophora reflexa*	KX784266	P		B	A	G	H
Plastid	Angiosperm - holoparasite	*Conopholis americana*	NC_023131.1					G	
Plastid	Angiosperm - holoparasite	*Cuscuta gronovii*	NC_009765.1					G	
Plastid	Angiosperm - holoparasite	*Cuscuta obtusiflora*	NC_009949.1					G	
Plastid	Angiosperm - holoparasite	*Cynomorium coccineum*	KX270752			B		G	H
Plastid	Angiosperm - holoparasite	*Cytinus hypocistis*	KT335971			B		G	
Plastid	Angiosperm - holoparasite	*Hydnora visseri*	NC_029358.1			B		G	H
Plastid	Angiosperm - holoparasite	*Phelipanche purpurea*	NC_023132.1					G	
Plastid	Angiosperm - holoparasite	*Pilostyles aethiopica*	KT981955			B		G	
Plastid	Angiosperm - holoparasite	*Pilostyles hamiltonii*	KT981956			B		G	
Plastid	Angiosperm - mycoheterotroph	*Epipogium aphyllum*	NC_026449.1					G	
Plastid	Angiosperm - mycoheterotroph	*Epipogium roseum*	NC_026448.1					G	
Plastid	Angiosperm - mycoheterotroph	*Monotropa uniflora*	NC_035582.1			B			
Plastid	Angiosperm - mycoheterotroph	*Sciaphila thaidanica*	MG757197					G	H
Plastid	Angiosperm - mycoheterotroph	*Thismia tentaculata*	KX171421			B		G	
Plastid	Gymnosperm - autotroph	*Ginkgo biloba*	JN867583.1	P			A		
Plastid	Gymnosperm - autotroph	*Pinus contorta*	NC_011153.4	P			A		
Plastid	Lycophyte - autotroph	*Selaginella moellendorffii*	NC_013086.1	P			A		
Plastid	Bryophyte - autotroph	*Marchantia polymorpha*	NC_001319.1				A		
Plastid	Bryophyte - autotroph	*Physcomitrella patens*	NC_005087.1	P			A		
Plastid	Euglenid - nonphotosynthetic	*Euglena longa*	NC_002652.1			B			
Plastid	Green alga - parasite	*Helicosporidium* sp.	NC_008100.1			B		G	H
Plastid	Green alga - parasite	*Prototheca zopfhii*	NC_037450.1						H
Plastid	Green alga - nonphotosynthetic	*Polytoma uvella*	KX828177			B			
Plastid	Red alga - parasite	*Choreocolax polysiphoniae*	NC_026522.1			B			
Plastid	Red alga - autotroph	*Bulboplastis apyrenoidosa* NIES-2742	NC_034787.1			B			
Plastid	Red alga - autotroph	*Cyanidioschyzon merolae*	NC_004799.1						H
Plastid	Diatom - nonphotosynthetic	*Nitzschia spp.*	NC_028737.1			B			
Plastid	Dinoflagellate - autotroph	*Lepidodinium chlorophorum*	NC_027093.1						H
Plastid	Brown alga - autotroph	*Dictyopteris divaricata*	NC_036804.1						H
Plastid	Apicomplexan - parasite	*Babesia microti*	NC_034636.1			B			H
Plastid	Apicomplexan - parasite	*Eimeria tenella*	NC_004823.1			B		G	
Plastid	Apicomplexan - parasite	*Plasmodium falciparum*	X95275-X95276		C	B		G	
Plastid	Apicomplexan - parasite	*Toxoplasma gondii*	U87145			B		G	
Mitochondrial	Excavate	*Acrasis kona* ATCC MYA-3509	NC_026286.1			B			
Mitochondrial	Ciliate - parasite	*Ichthyophthirius multifiliis* G5	NC_015981.1			B			
Mitochondrial	Choanoflagellate	*Monosiga brevicollis*	NC_004309.1			B			
Mitochondrial	Insect - parasite	*Diadegma semiclausum*	NC_012708.1			B			
Mitochondrial	Roundworm	*Radopholus similis*	NC_013253.1			B			
Mitochondrial	Fungus	*Rozella allomycis*	NC_021611.1			B			
Mitochondrial	Yeast	*Nakaseomyces bacillisporus*	NC_012621.1		C	B			
Mitochondrial	Yeast	*Saccharomyces cerevisiae* YJM1447	CP006552			B			
Bacterial	Proteobacteria	Carsonella ruddii CE isolate Thao2000	SAMN02641648			B			
Bacterial	Proteobacteria	*Escherichia coli* str. K-12 substr. MG1655	U00096.3						
Bacterial	Proteobacteria	*Nasuia deltocephalinicola*	SAMN06919537			B			
Bacterial	Proteobacteria	Zinderia insecticola CARI	CP002161.1		C	B			

*P, phylogenomic and rate analyses (Figs. 4 and S10); C, codon and amino acid usage for five AT-rich genomes (Fig. S7); B, GC content of individual genes and whole gene sets for 30 AT-rich genomes (Fig. S5 and Table S5); A, alignments (Figs. 3B, S4, S12, and S13); G, plastome gene content (Fig. S6); H, features of highly compact plastomes (Table S9).

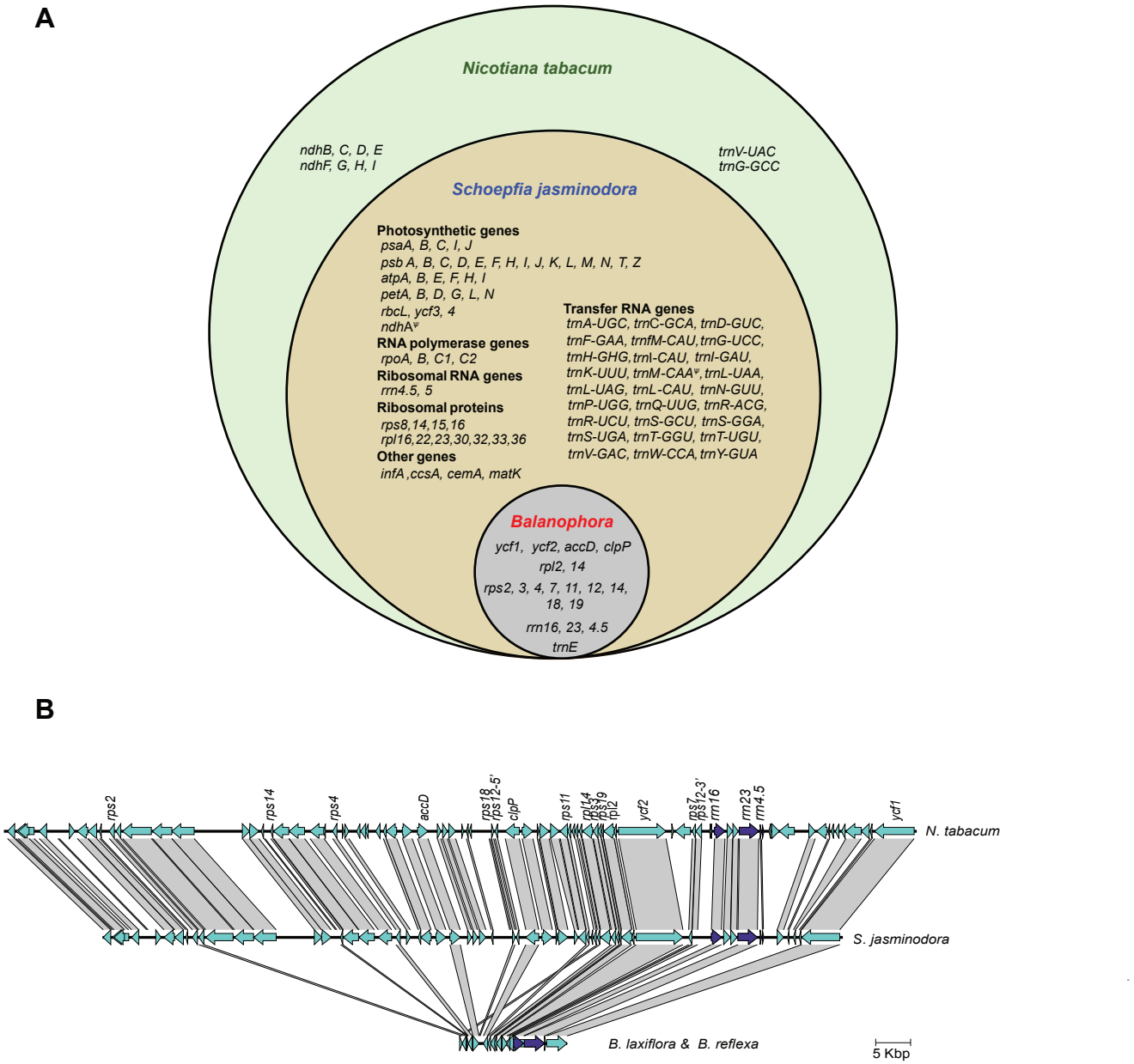
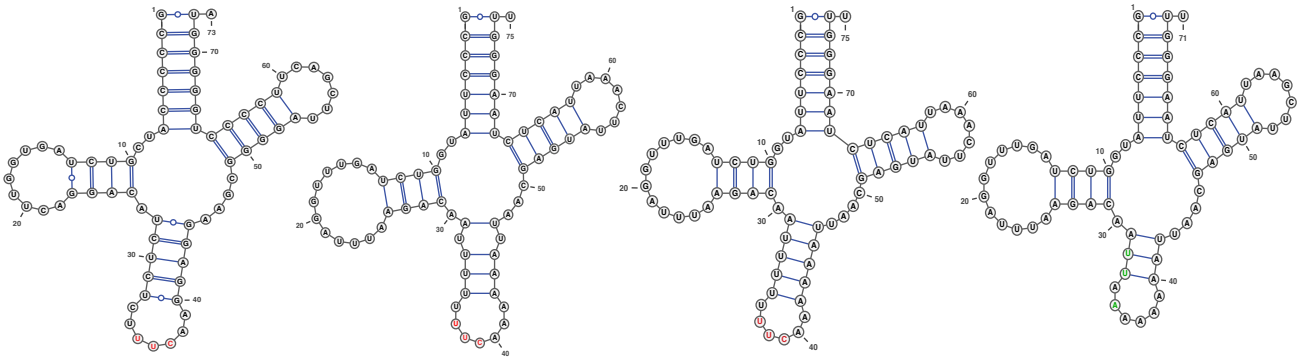


Figure S1. Gene content and order in the plastomes of *Nicotiana*, *Schoepfia*, and *Balanophora*. (*A*) Plastid gene content in *Balanophora* is a highly reduced subset of that in *Schoepfia* and *Nicotiana*. All genes listed in the gray, *Balanophora* portion of the figure are present in all three genomes. With the exception of *infA*, all genes listed in the tan, *Schoepfia* portion are shared with the *Nicotiana* genome. The genes listed in the green, *Nicotiana* portion are restricted to its genome. (*B*) Gene-order is colinear between the *N. tabacum* and *S. jasminodora* plastomes (one copy of the IR has been removed) and nearly colinear between these two plastomes and the two *Balanophora* plastomes, with the only difference being the relative location of *rpl14*. This colinearity plot was made using Easyfig v2.2.2 (16).

A

```
                >>>>>>>..>>>>.............<<<<.>>>>>........<<<<<....>>>>>.......<<<<<<<<<<.
Nicotiana       GCCCCCAUCGUCUAGU---GGUUUAGGACAUCUCUCUUUCAAGGAGGCAGCGGGGAUUCGAAUUCCCCUGGGGGUA
Schoepfia       ......................C.....................A.............C.............
B. lax.         ....UU..G.......UUG.A....A....AU.-.U.....AA.AUU.A..A.U....A....A.U..AA....U
B. ref.         ....UU..G.......UUG.A....A....AU.-----AA..AA.AUU.A..A.U.........A.U..AA....U
```

B



Schoepfia (by both tRNAscan-SE and MiTFi)

B. laxiflora (by tRNAscan-SE)

B. laxiflora (by MiTFi)

B. reflexa (by MiTFi)

Figure S2. Primary sequence and secondary structure of plastome-encoded *trn*E. (*A*) Primary sequence alignment. Carets indicate regions of secondary base pairing in the *Nicotiana* and *Schoepfia* sequences, with red marking the acceptor arm, blue the D-loop arm, orange the anticodon-loop arm, and green the T-loop arm. The UUC anticodon sequence is boxed. Magenta lettering marks the eight nucleotides in the *Nicotiana* and *Schoepfia* sequences that correspond to sequence determinants for correct charging of $tRNA^{Glu}$(UUC) in *E. coli* (17). All but one of these determinants is present in *B. laxiflora.* Green lettering marks the UUnA sequence that we propose is used in glutamate charging in *B. reflexa* in place of the normal UUnA charging determinants. (*B*) Secondary structure of $tRNA^{Glu}$(UUC)-like molecules from *Schoepfia*, *B. laxiflora*, and *B. reflexa* as predicted by tRNAscan-SE v1.21 (18) and MiTFi v0.1 (19). Note that tRNAscan did not predict a structure for the *B. reflexa* sequence. UUC triplets located in anticodon loops are in red letters. Green letters in the *B. reflexa* structure are as in (*A*). Secondary structures were visualized using the VARNA java web start applet (http://varna.lri.fr/).

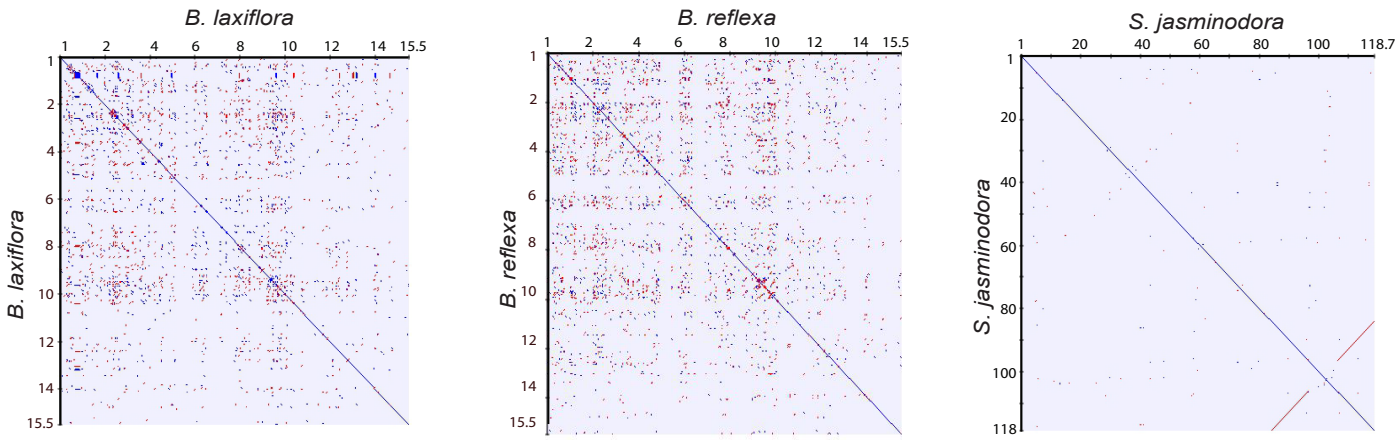


Figure S3. Self dotplots show the absence of a large inverted repeat in *Balanophora* plastomes, but the presence of one (of 12 kb in size) in *Schoepfia jasminodora*, as in the great majority of other angiosperm plastomes. Direct repeats are shown as blue dots or lines and inverted repeats as red dots or lines. These plots were made using a word size of 20 and identity of 85%. Dot plots were generated using Gepard (20).

CLPP

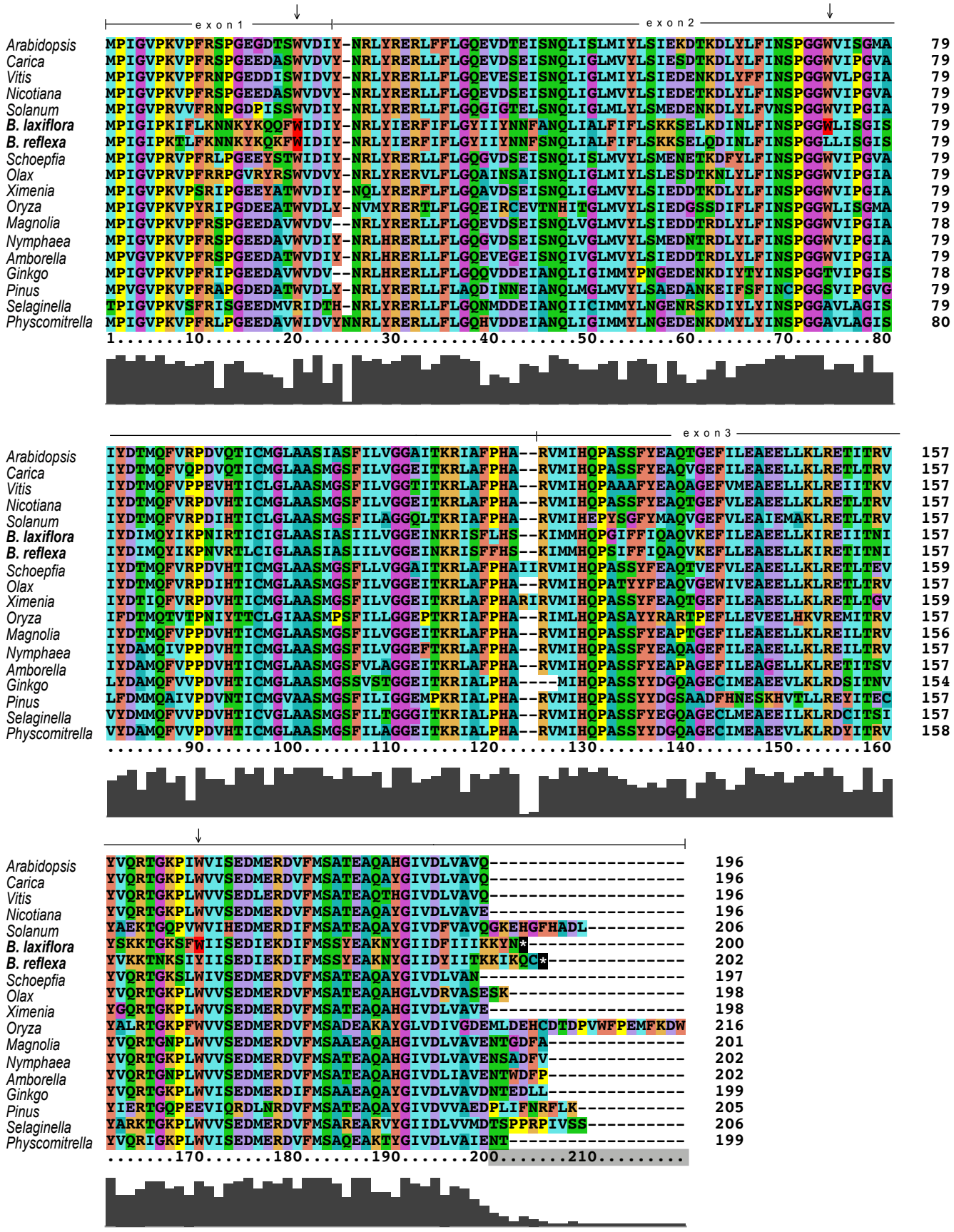


Figure S4. Inferred amino acid alignments of the 15 protein genes present in *Balanophora* plastomes. Stop codons are shown only for *Balanophora* and are shaded in gray (TGA) or black (TAA). The arrows mark internal TAG codons present in one or both *Balanophora* species and inferred to encode W; note that most or all non-*Balanophora* land plants contain TGG (W in the standard genetic code) at most of these positions (Table 1). Regions of ambiguous alignment are marked with boxes and were excluded from the phylogenetic and molecular evolutionary rate analyses (open box, only the *Balanophora* sequences were excluded; gray box, the entire region was excluded). The histograms give the percentage of the aligned sequences that share the most common amino acid at each position.

ACCD

Figure S4. (cont.) Inferred amino acid alignments of the 15 protein genes present in *Balanophora* plastomes. Stop codons are shown only for *Balanophora* and are shaded in gray (TGA) or black (TAA). The arrows mark internal TAG codons present in one or both *Balanophora* species and inferred to encode W; note that most or all non-*Balanophora* and plants contain TGG (W in the standard genetic code) at most of these positions (Table 1). Regions of ambiguous alignment are marked with boxes and were excluded from the phylogenetic and molecular evolutionary rate analyses (open box, only the *Balanophora* sequences were excluded; gray box, the entire region was excluded). The histograms give the percentage of the aligned sequences that share the most common amino acid at each position.

RPL2

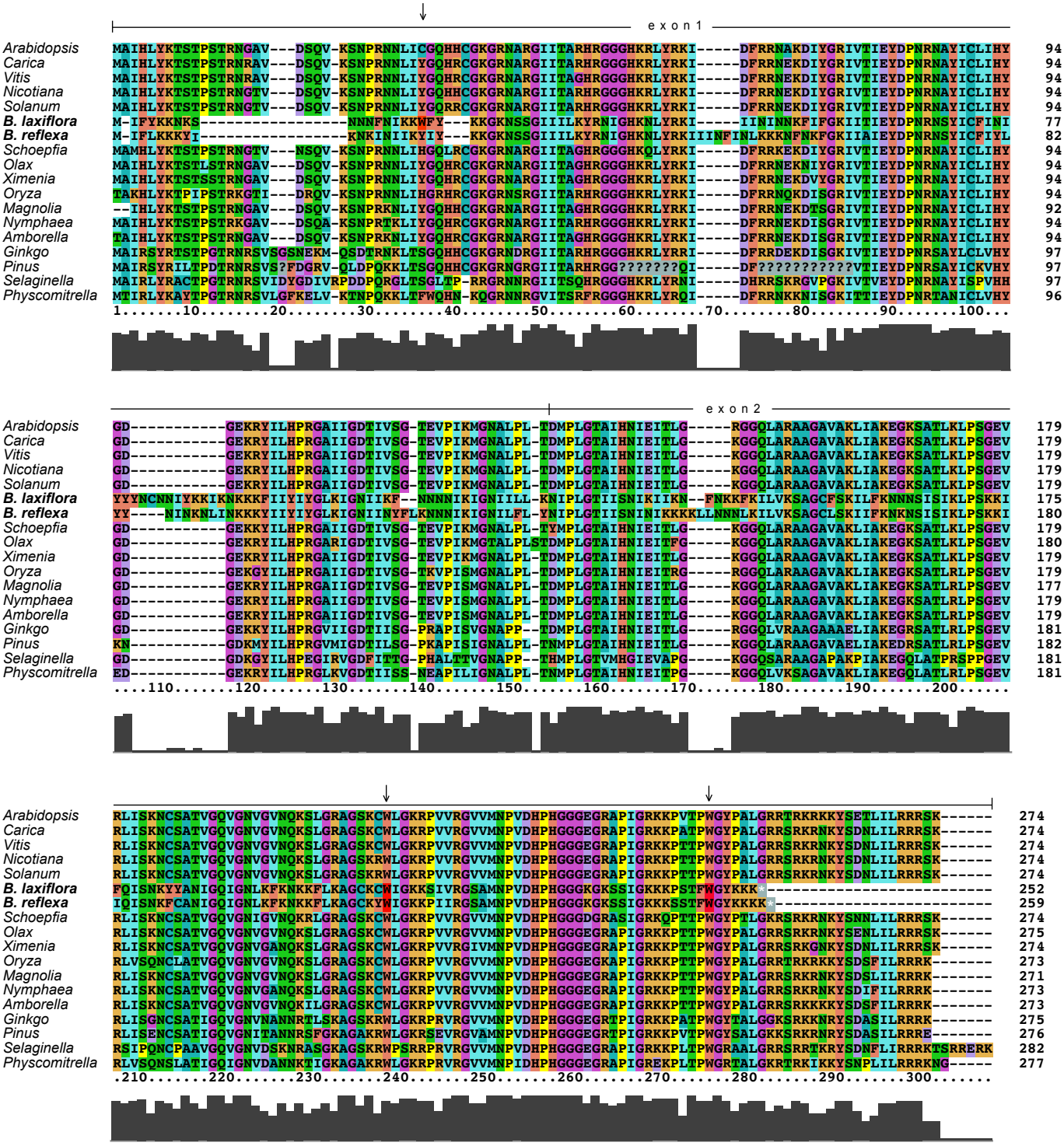


Figure S4. (cont.) Inferred amino acid alignments of the 15 protein genes present in *Balanophora* plastomes. Stop codons are shown only for *Balanophora* and are shaded in gray (TGA) or black (TAA). The arrows mark internal TAG codons present in one or both *Balanophora* species and inferred to encode W; note that most or all non-*Balanophora* and plants contain TGG (W in the standard genetic code) at most of these positions (Table 1). Regions of ambiguous alignment are marked with boxes and were excluded from the phylogenetic and molecular evolutionary rate analyses (open box, only the *Balanophora* sequences were excluded; gray box, the entire region was excluded). The histograms give the percentage of the aligned sequences that share the most common amino acid at each position.

RPL14

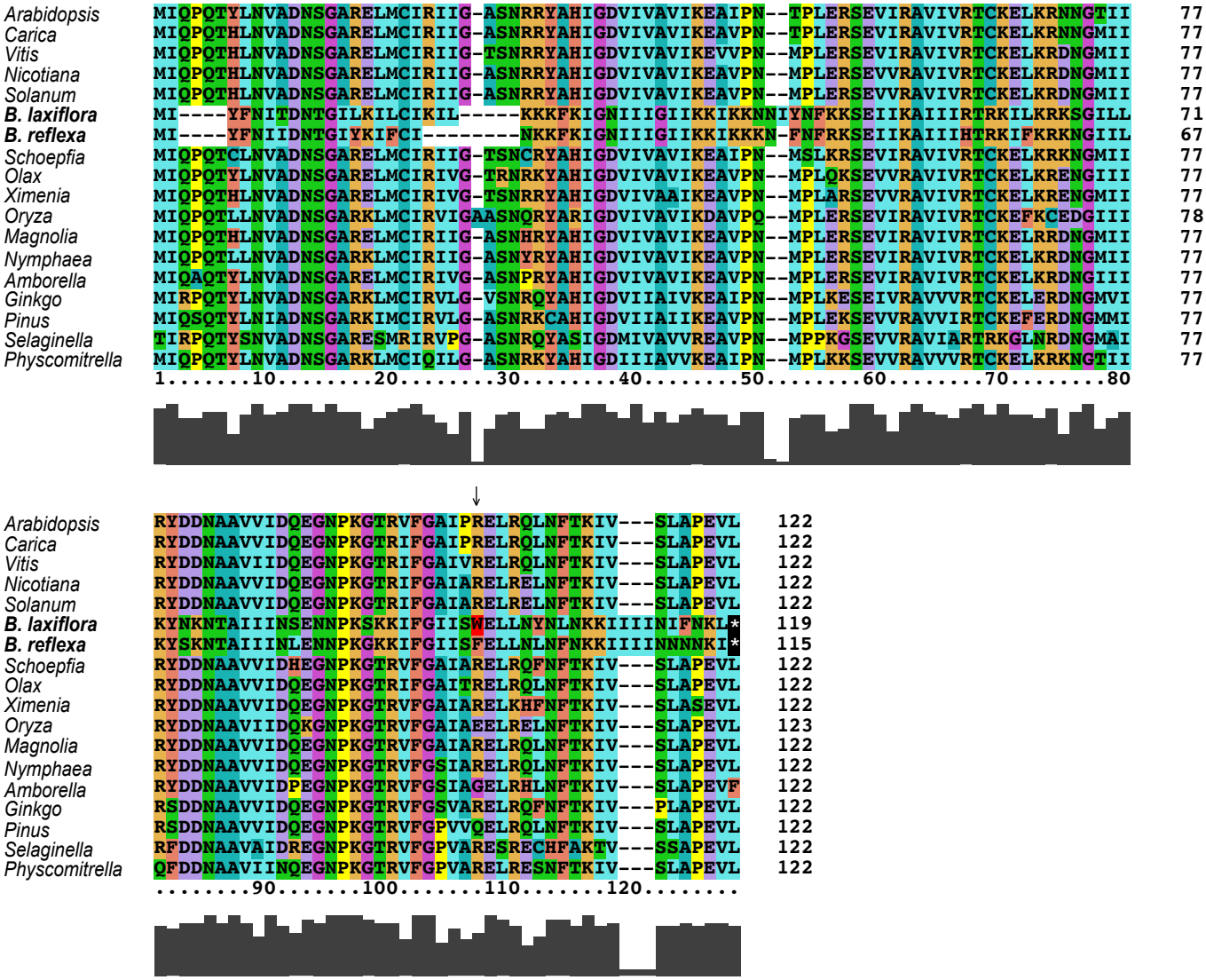


Figure S4. (cont.) Inferred amino acid alignments of the 15 protein genes present in *Balanophora* plastomes. Stop codons are shown only for *Balanophora* and are shaded in gray (TGA) or black (TAA). The arrows mark internal TAG codons present in one or both *Balanophora* species and inferred to encode W; note that most or all non-*Balanophora* and plants contain TGG (W in the standard genetic code) at most of these positions (Table 1). Regions of ambiguous alignment are marked with boxes and were excluded from the phylogenetic and molecular evolutionary rate analyses (open box, only the *Balanophora* sequences were excluded; gray box, the entire region was excluded). The histograms give the percentage of the aligned sequences that share the most common amino acid at each position.

RPS2

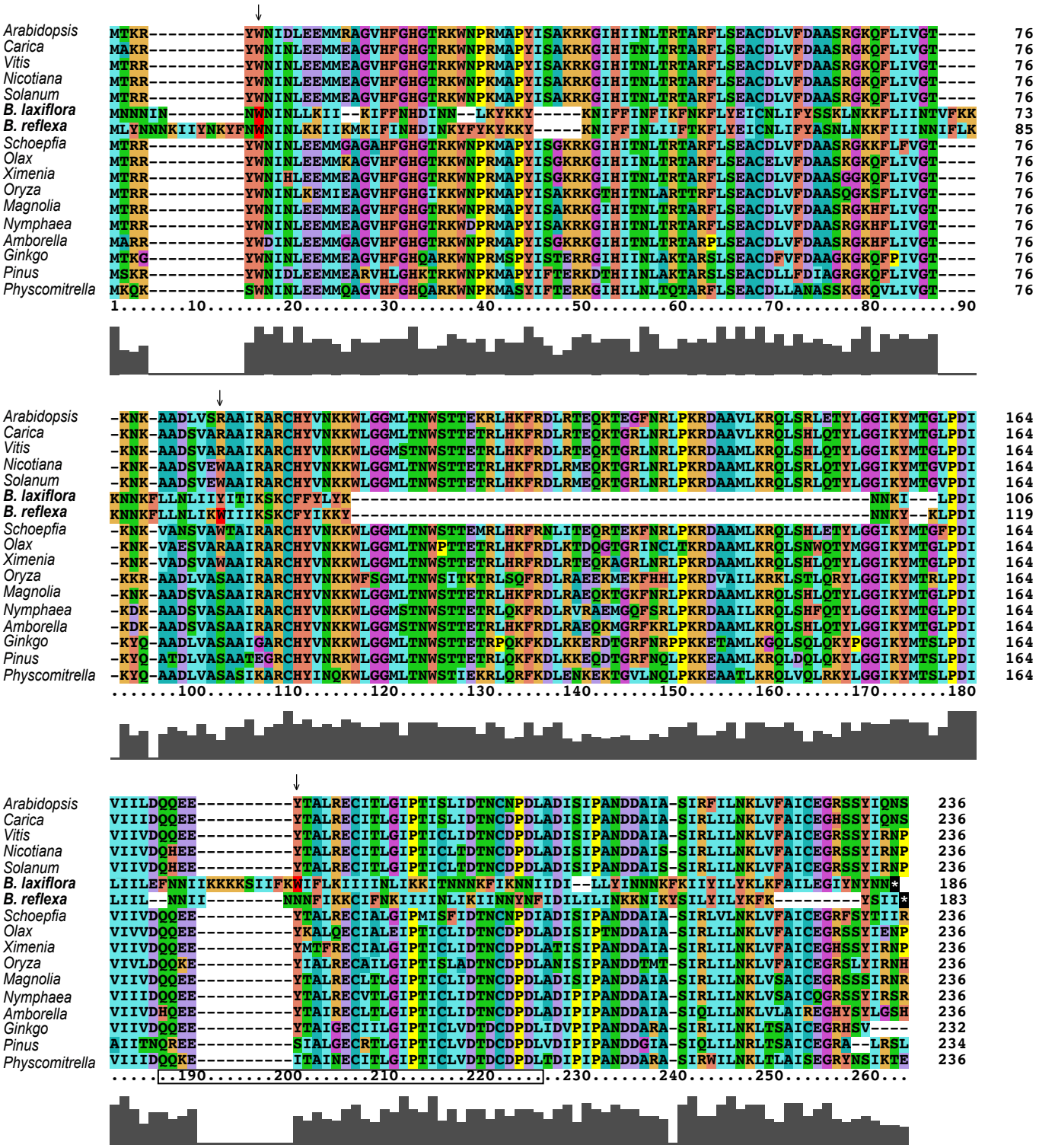


Figure S4. (cont.) Inferred amino acid alignments of the 15 protein genes present in *Balanophora* plastomes. Stop codons are shown only for *Balanophora* and are shaded in gray (TGA) or black (TAA). The arrows mark internal TAG codons present in one or both *Balanophora* species and inferred to encode W; note that most or all non-*Balanophora* and plants contain TGG (W in the standard genetic code) at most of these positions (Table 1). Regions of ambiguous alignment are marked with boxes and were excluded from the phylogenetic and molecular evolutionary rate analyses (open box, only the *Balanophora* sequences were excluded; gray box, the entire region was excluded). The histograms give the percentage of the aligned sequences that share the most common amino acid at each position.

RPS3

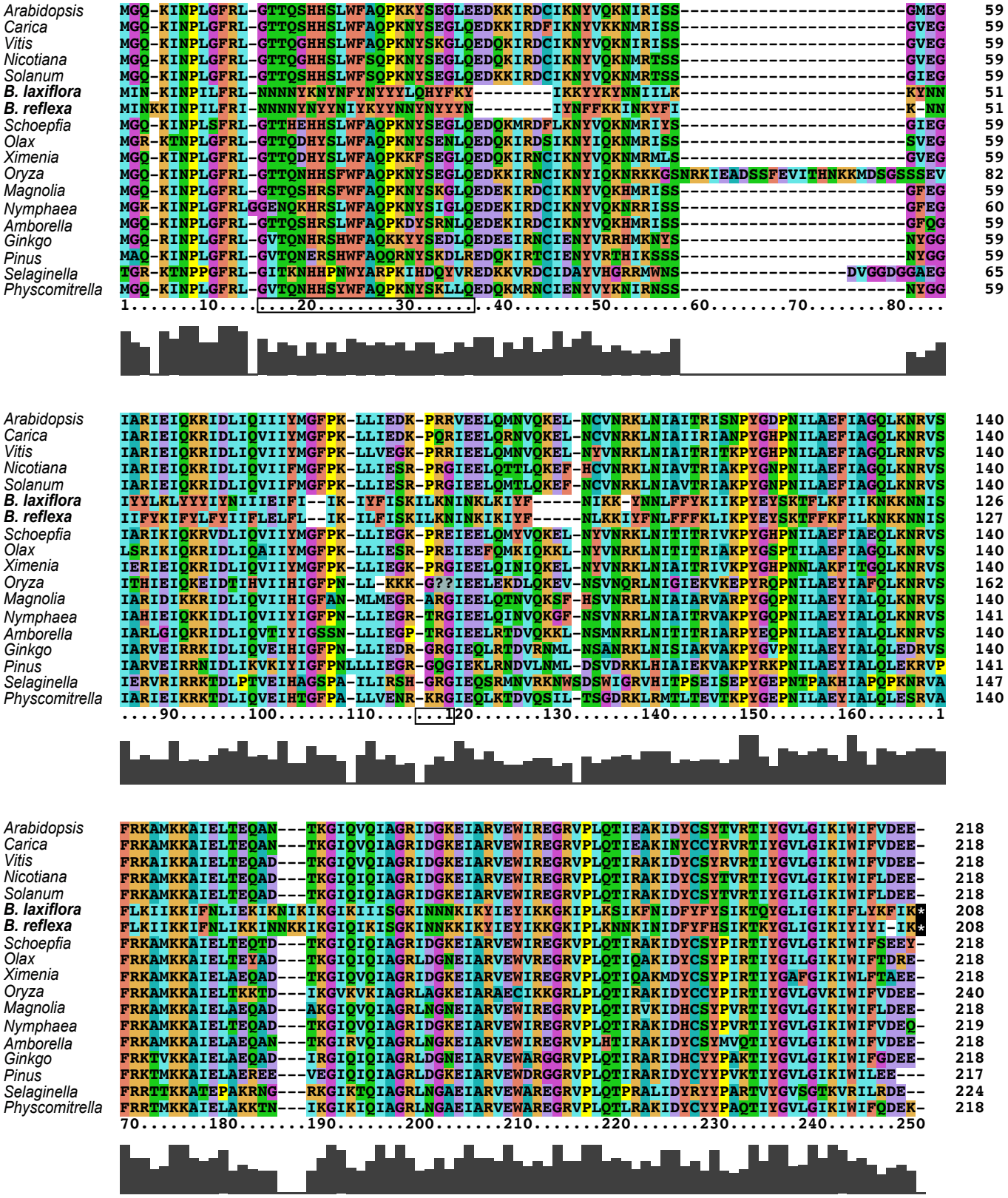


Figure S4. (cont.) Inferred amino acid alignments of the 15 protein genes present in *Balanophora* plastomes. Stop codons are shown only for *Balanophora* and are shaded in gray (TGA) or black (TAA). The arrows mark internal TAG codons present in one or both *Balanophora* species and inferred to encode W; note that most or all non-*Balanophora* and plants contain TGG (W in the standard genetic code) at most of these positions (Table 1). Regions of ambiguous alignment are marked with boxes and were excluded from the phylogenetic and molecular evolutionary rate analyses (open box, only the *Balanophora* sequences were excluded; gray box, the entire region was excluded). The histograms give the percentage of the aligned sequences that share the most common amino acid at each position.

RPS4

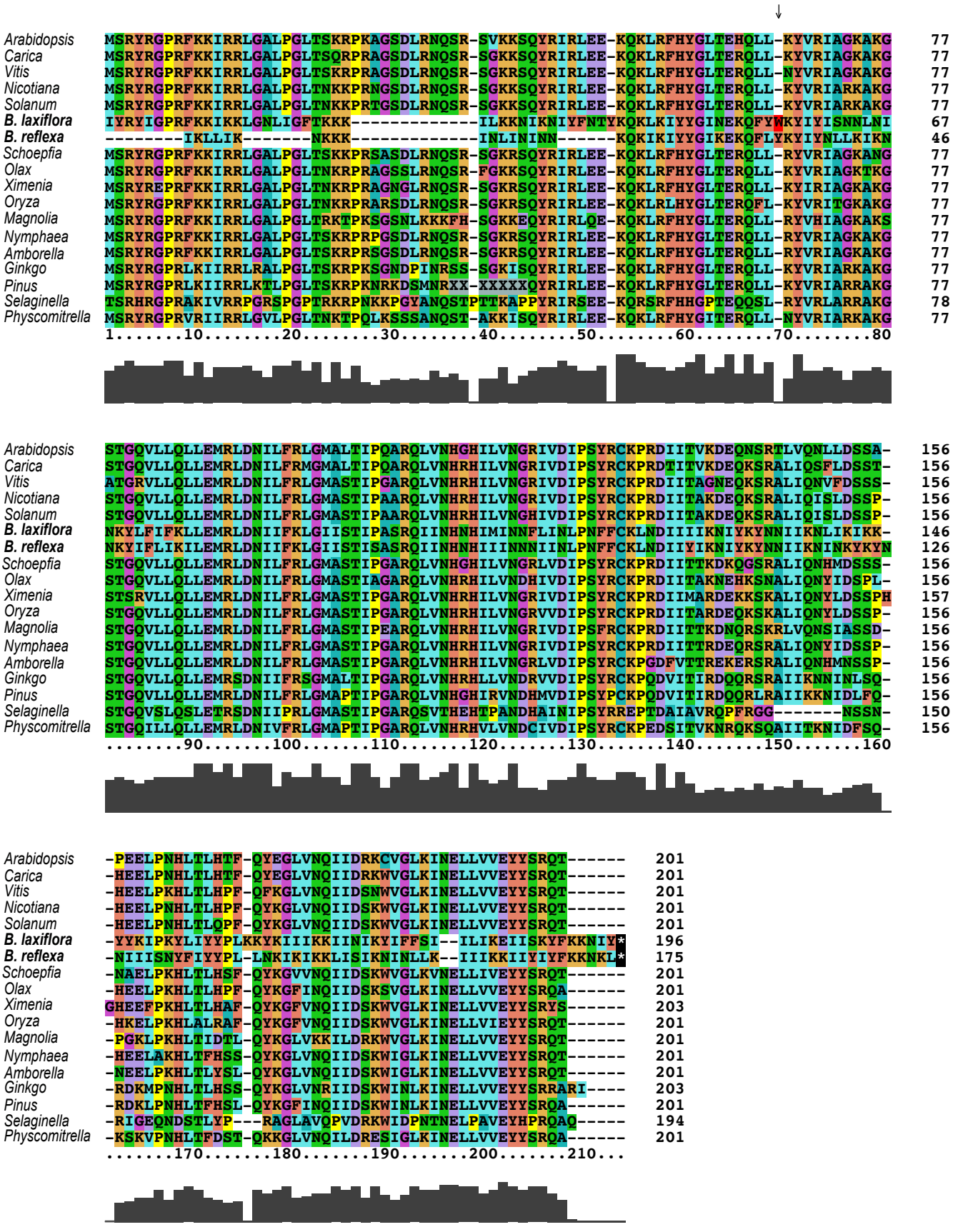


Figure S4. (cont.) Inferred amino acid alignments of the 15 protein genes present in *Balanophora* plastomes. Stop codons are shown only for *Balanophora* and are shaded in gray (TGA) or black (TAA). The arrows mark internal TAG codons present in one or both *Balanophora* species and inferred to encode W; note that most or all non-*Balanophora* and plants contain TGG (W in the standard genetic code) at most of these positions (Table 1). Regions of ambiguous alignment are marked with boxes and were excluded from the phylogenetic and molecular evolutionary rate analyses (open box, only the *Balanophora* sequences were excluded; gray box, the entire region was excluded). The histograms give the percentage of the aligned sequences that share the most common amino acid at each position.

RPS7

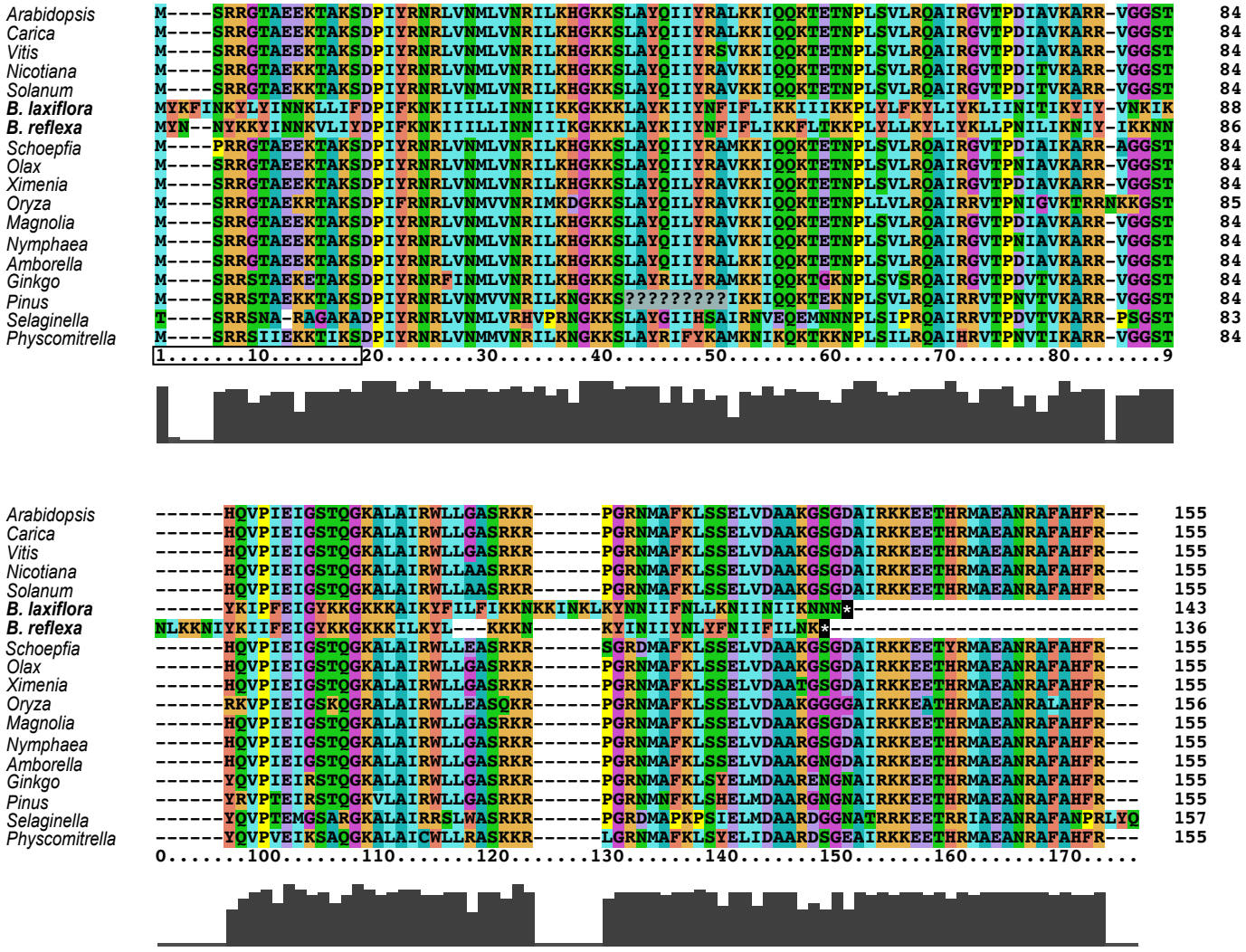


Figure S4. (cont.) Inferred amino acid alignments of the 15 protein genes present in *Balanophora* plastomes. Stop codons are shown only for *Balanophora* and are shaded in gray (TGA) or black (TAA). The arrows mark internal TAG codons present in one or both *Balanophora* species and inferred to encode W; note that most or all non-*Balanophora* and plants contain TGG (W in the standard genetic code) at most of these positions (Table 1). Regions of ambiguous alignment are marked with boxes and were excluded from the phylogenetic and molecular evolutionary rate analyses (open box, only the *Balanophora* sequences were excluded; gray box, the entire region was excluded). The histograms give the percentage of the aligned sequences that share the most common amino acid at each position.

RPS11

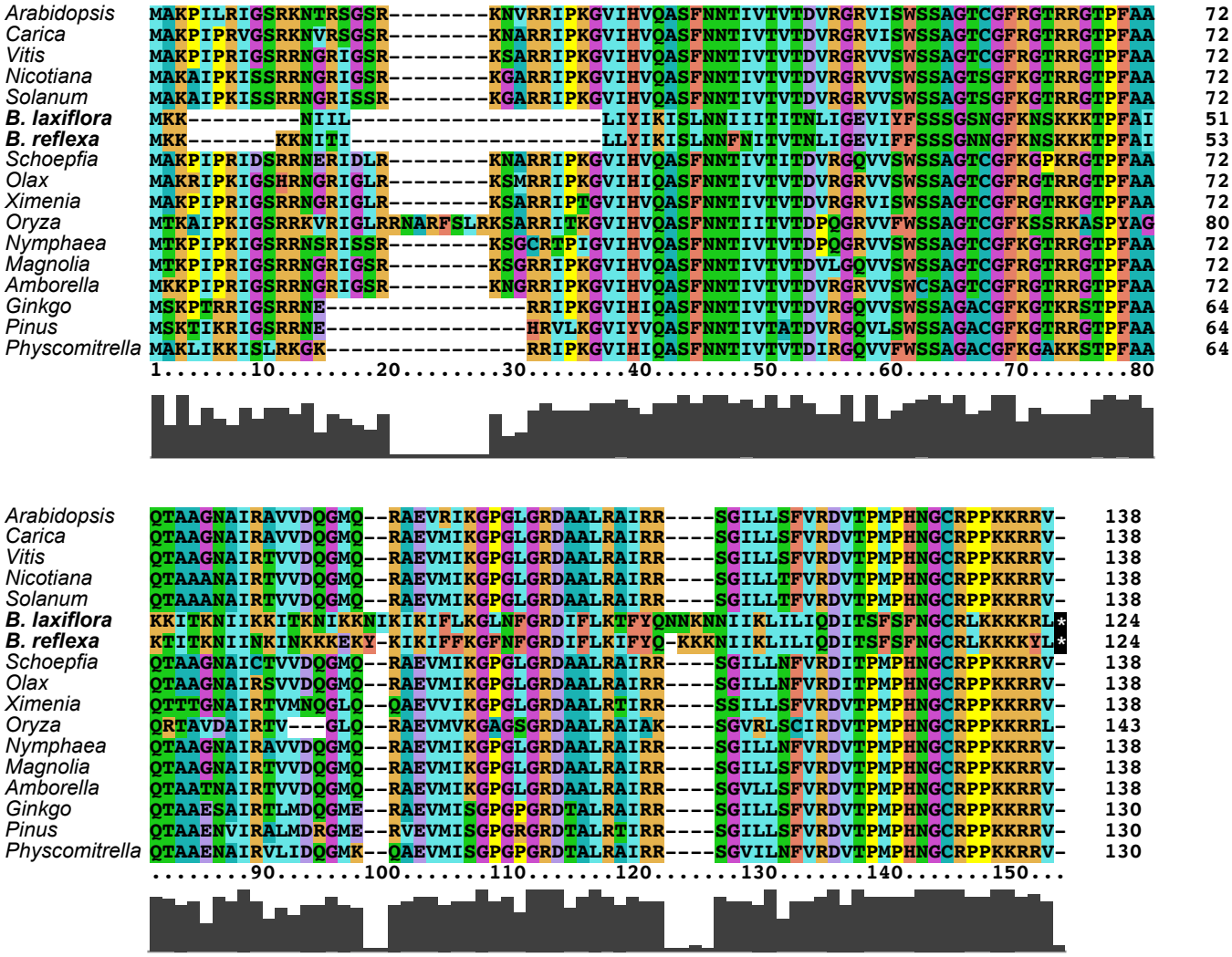


Figure S4. (cont.) Inferred amino acid alignments of the 15 protein genes present in *Balanophora* plastomes. Stop codons are shown only for *Balanophora* and are shaded in gray (TGA) or black (TAA). The arrows mark internal TAG codons present in one or both *Balanophora* species and inferred to encode W; note that most or all non-*Balanophora* and plants contain TGG (W in the standard genetic code) at most of these positions (Table 1). Regions of ambiguous alignment are marked with boxes and were excluded from the phylogenetic and molecular evolutionary rate analyses (open box, only the *Balanophora* sequences were excluded; gray box, the entire region was excluded). The histograms give the percentage of the aligned sequences that share the most common amino acid at each position.

RPS12

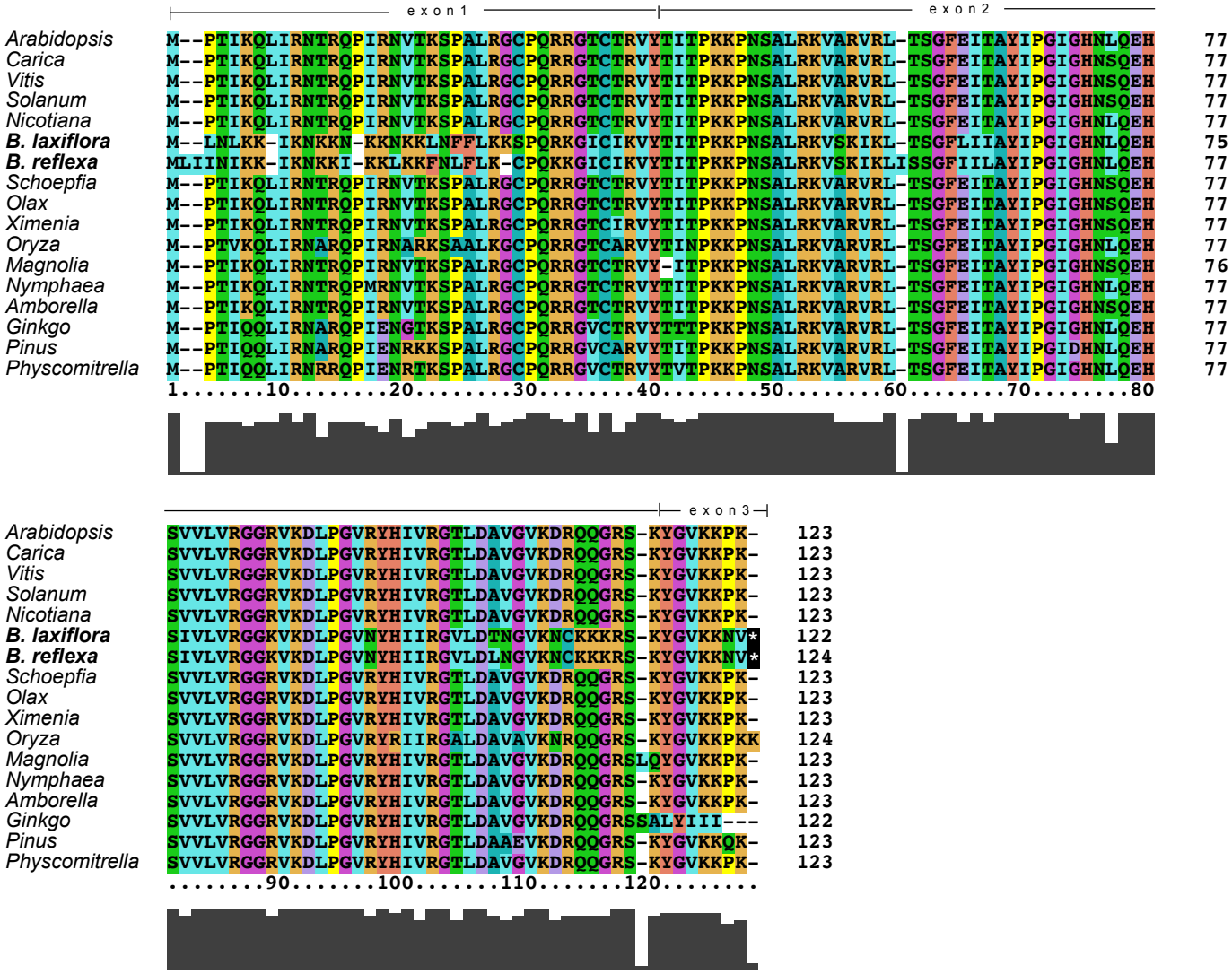


Figure S4. (cont.) Inferred amino acid alignments of the 15 protein genes present in *Balanophora* plastomes. Stop codons are shown only for *Balanophora* and are shaded in gray (TGA) or black (TAA). The arrows mark internal TAG codons present in one or both *Balanophora* species and inferred to encode W; note that most or all non-*Balanophora* and plants contain TGG (W in the standard genetic code) at most of these positions (Table 1). Regions of ambiguous alignment are marked with boxes and were excluded from the phylogenetic and molecular evolutionary rate analyses (open box, only the *Balanophora* sequences were excluded; gray box, the entire region was excluded). The histograms give the percentage of the aligned sequences that share the most common amino acid at each position.

RPS14

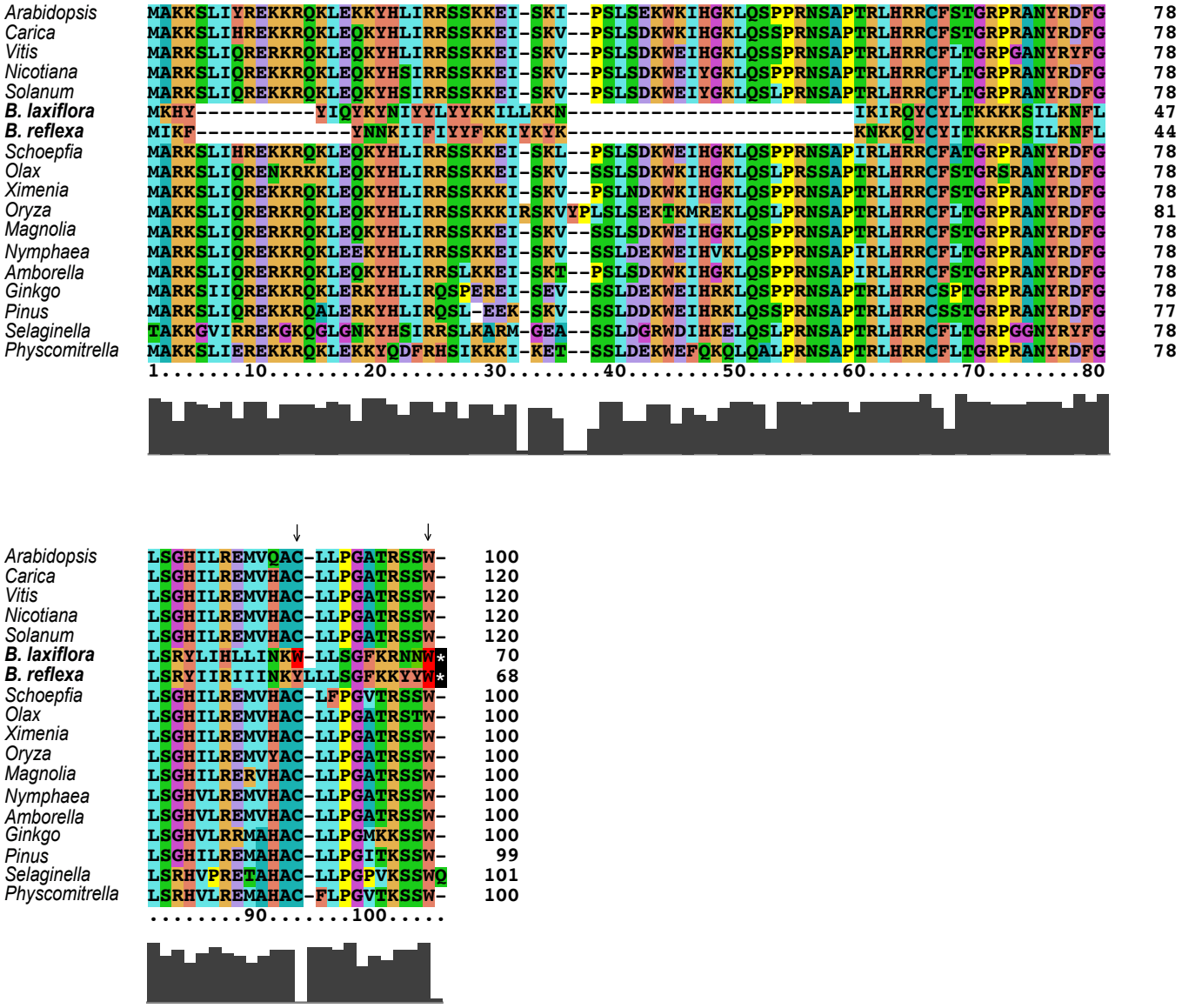


Figure S4. (cont.) Inferred amino acid alignments of the 15 protein genes present in *Balanophora* plastomes. Stop codons are shown only for *Balanophora* and are shaded in gray (TGA) or black (TAA). The arrows mark internal TAG codons present in one or both *Balanophora* species and inferred to encode W; note that most or all non-*Balanophora* and plants contain TGG (W in the standard genetic code) at most of these positions (Table 1). Regions of ambiguous alignment are marked with boxes and were excluded from the phylogenetic and molecular evolutionary rate analyses (open box, only the *Balanophora* sequences were excluded; gray box, the entire region was excluded). The histograms give the percentage of the aligned sequences that share the most common amino acid at each position.

RPS18

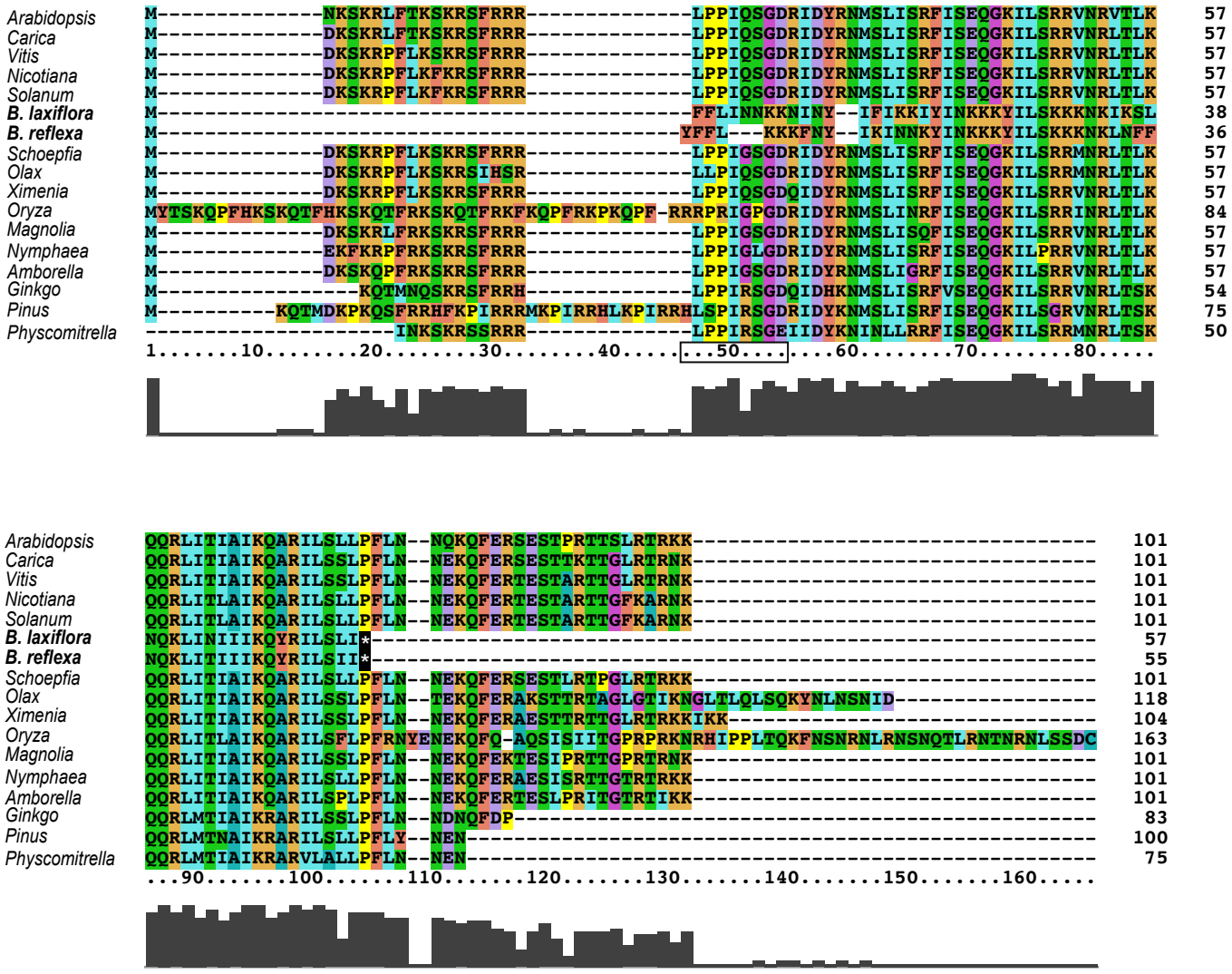


Figure S4. (cont.) Inferred amino acid alignments of the 15 protein genes present in *Balanophora* plastomes. Stop codons are shown only for *Balanophora* and are shaded in gray (TGA) or black (TAA). The arrows mark internal TAG codons present in one or both *Balanophora* species and inferred to encode W; note that most or all non-*Balanophora* and plants contain TGG (W in the standard genetic code) at most of these positions (Table 1). Regions of ambiguous alignment are marked with boxes and were excluded from the phylogenetic and molecular evolutionary rate analyses (open box, only the *Balanophora* sequences were excluded; gray box, the entire region was excluded). The histograms give the percentage of the aligned sequences that share the most common amino acid at each position.

RPS19

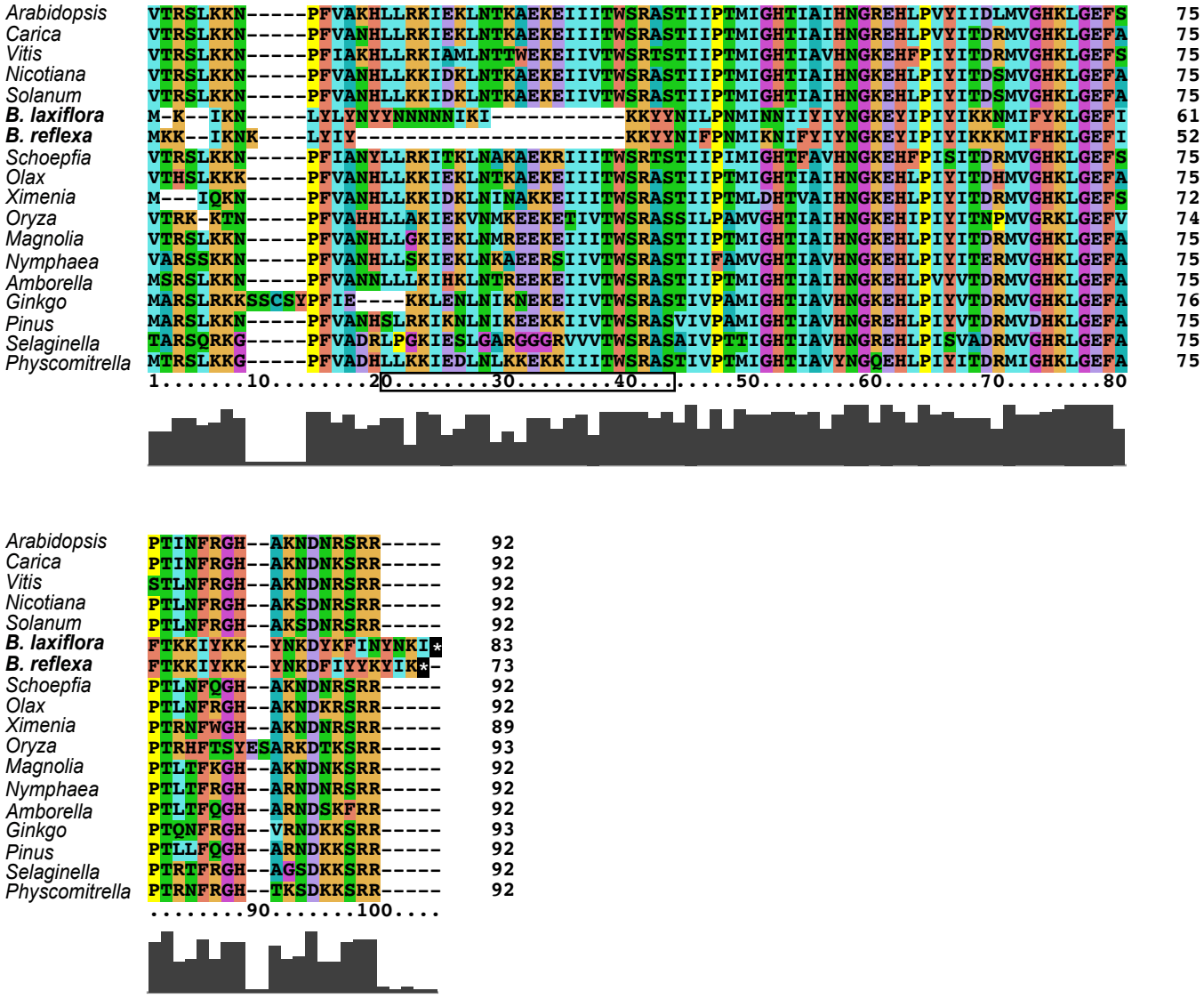


Figure S4. (cont.) Inferred amino acid alignments of the 15 protein genes present in *Balanophora* plastomes. Stop codons are shown only for *Balanophora* and are shaded in gray (TGA) or black (TAA). The arrows mark internal TAG codons present in one or both *Balanophora* species and inferred to encode W; note that most or all non-*Balanophora* and plants contain TGG (W in the standard genetic code) at most of these positions (Table 1). Regions of ambiguous alignment are marked with boxes and were excluded from the phylogenetic and molecular evolutionary rate analyses (open box, only the *Balanophora* sequences were excluded; gray box, the entire region was excluded). The histograms give the percentage of the aligned sequences that share the most common amino acid at each position.

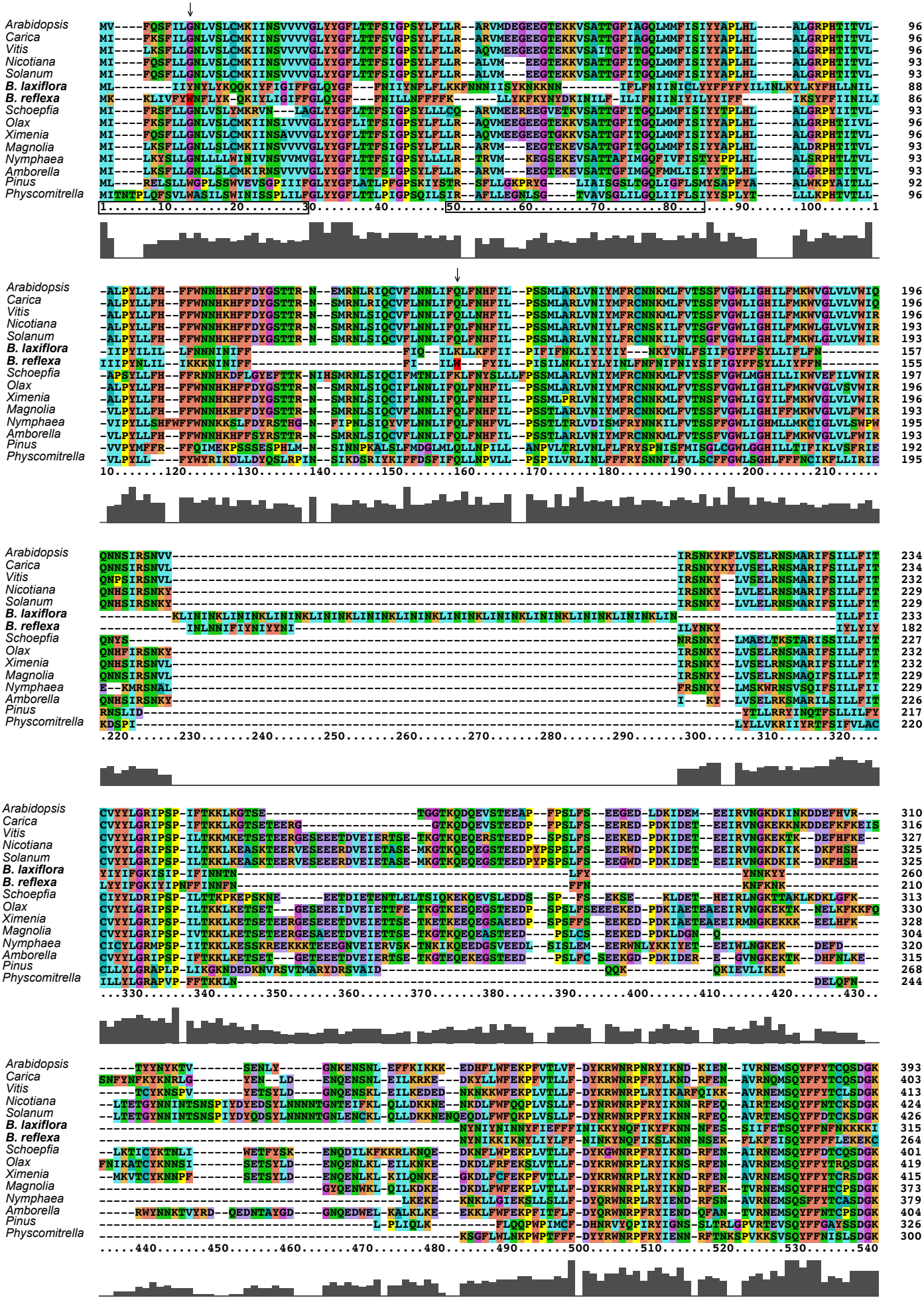


Figure S4. (cont.) Inferred amino acid alignments of the 15 protein genes present in *Balanophora* plastomes. Stop codons are shown only for *Balanophora* and are shaded in gray (TGA) or black (TAA). The arrows mark internal TAG codons present in one or both *Balanophora* species and inferred to encode W; note that most or all non-*Balanophora* land plants contain TGG (W in the standard genetic code) at most of these positions (Table 1). Regions of ambiguous alignment are marked with boxes and were excluded from the phylogenetic and molecular evolutionary rate analyses (open box, only the *Balanophora* sequences were excluded; gray box, the entire region was excluded). The histograms give the percentage of the aligned sequences that share the most common amino acid at each position.

YCF1

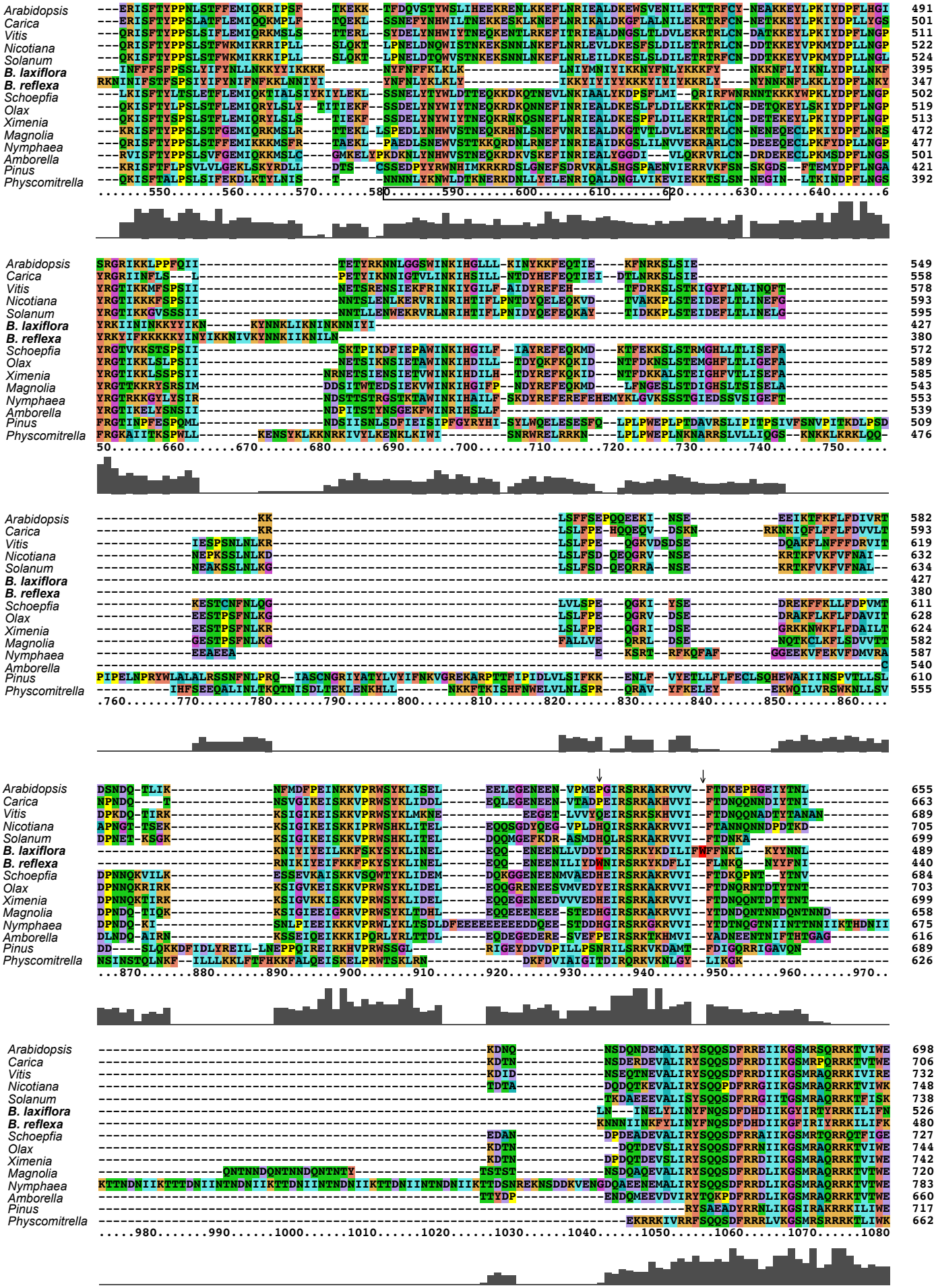


Figure S4. (cont.) Inferred amino acid alignments of the 15 protein genes present in *Balanophora* plastomes. Stop codons are shown only for *Balanophora* and are shaded in gray (TGA) or black (TAA). The arrows mark internal TAG codons present in one or both *Balanophora* species and inferred to encode W; note that most or all non-*Balanophora* land plants contain TGG (W in the standard genetic code) at most of these positions (Table 1). Regions of ambiguous alignment are marked with boxes and were excluded from the phylogenetic and molecular evolutionary rate analyses (open box, only the *Balanophora* sequences were excluded; gray box, the entire region was excluded). The histograms give the percentage of the aligned sequences that share the most common amino acid at each position.

YCF1

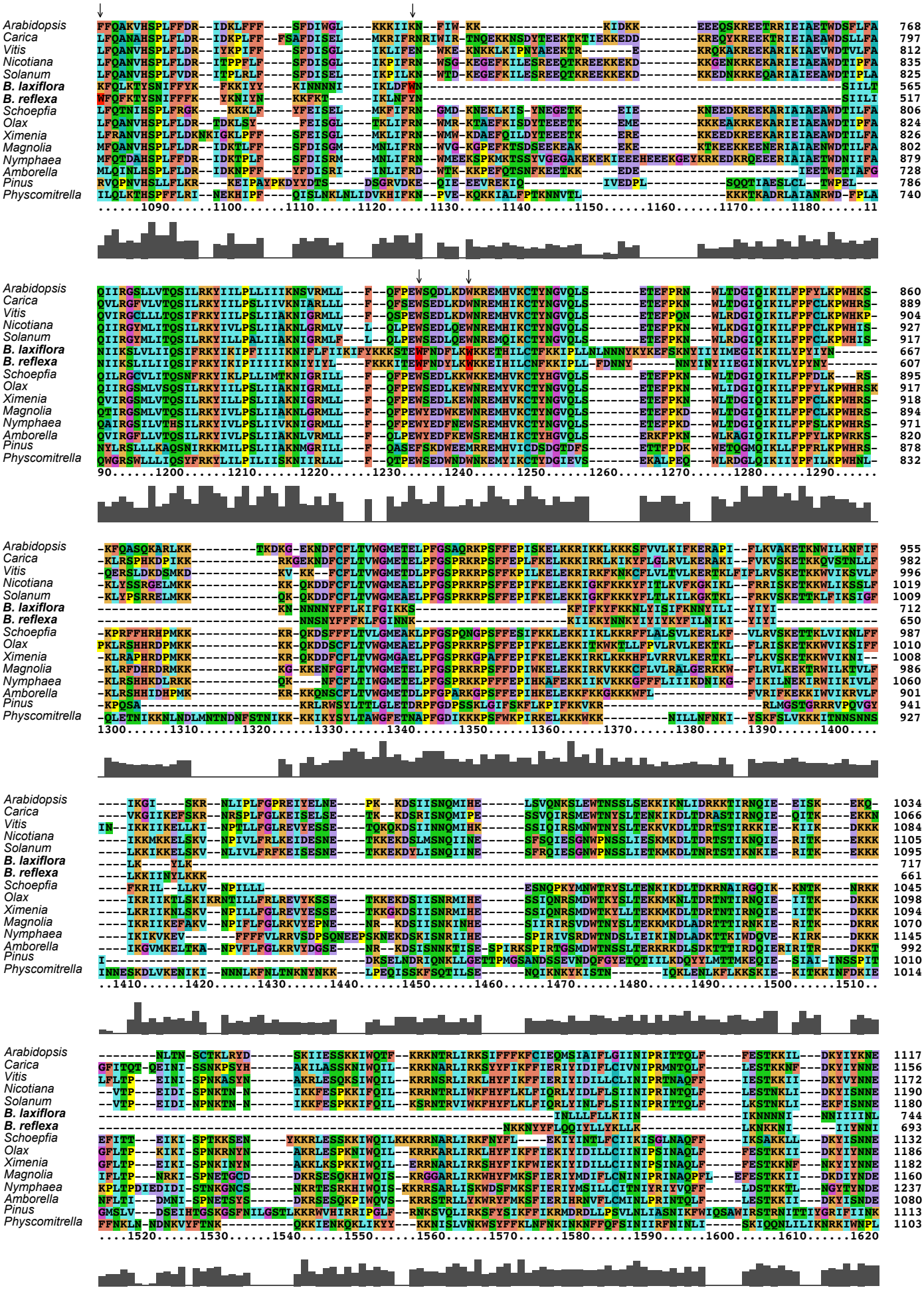


Figure S4. (cont.) Inferred amino acid alignments of the 15 protein genes present in *Balanophora* plastomes. Stop codons are shown only for *Balanophora* and are shaded in gray (TGA) or black (TAA). The arrows mark internal TAG codons present in one or both *Balanophora* species and inferred to encode W; note that most or all non-*Balanophora* land plants contain TGG (W in the standard genetic code) at most of these positions (Table 1). Regions of ambiguous alignment are marked with boxes and were excluded from the phylogenetic and molecular evolutionary rate analyses (open box, only the *Balanophora* sequences were excluded; gray box, the entire region was excluded). The histograms give the percentage of the aligned sequences that share the most common amino acid at each position.

YCF1

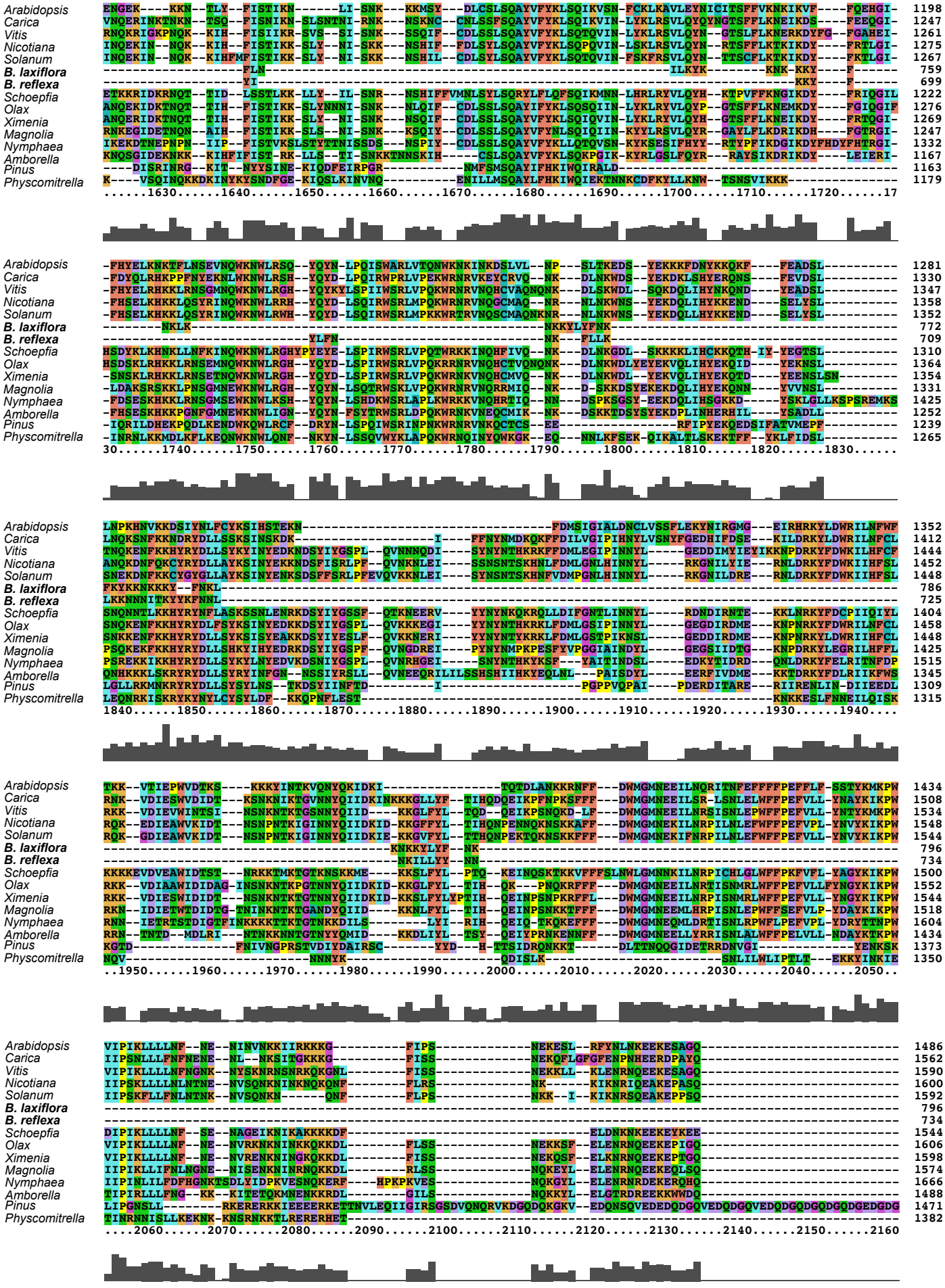


Figure S4. (cont.) Inferred amino acid alignments of the 15 protein genes present in *Balanophora* plastomes. Stop codons are shown only for *Balanophora* and are shaded in gray (TGA) or black (TAA). The arrows mark internal TAG codons present in one or both *Balanophora* species and inferred to encode W; note that most or all non-*Balanophora* land plants contain TGG (W in the standard genetic code) at most of these positions (Table 1). Regions of ambiguous alignment are marked with boxes and were excluded from the phylogenetic and molecular evolutionary rate analyses (open box, only the *Balanophora* sequences were excluded; gray box, the entire region was excluded). The histograms give the percentage of the aligned sequences that share the most common amino acid at each position.

YCF1

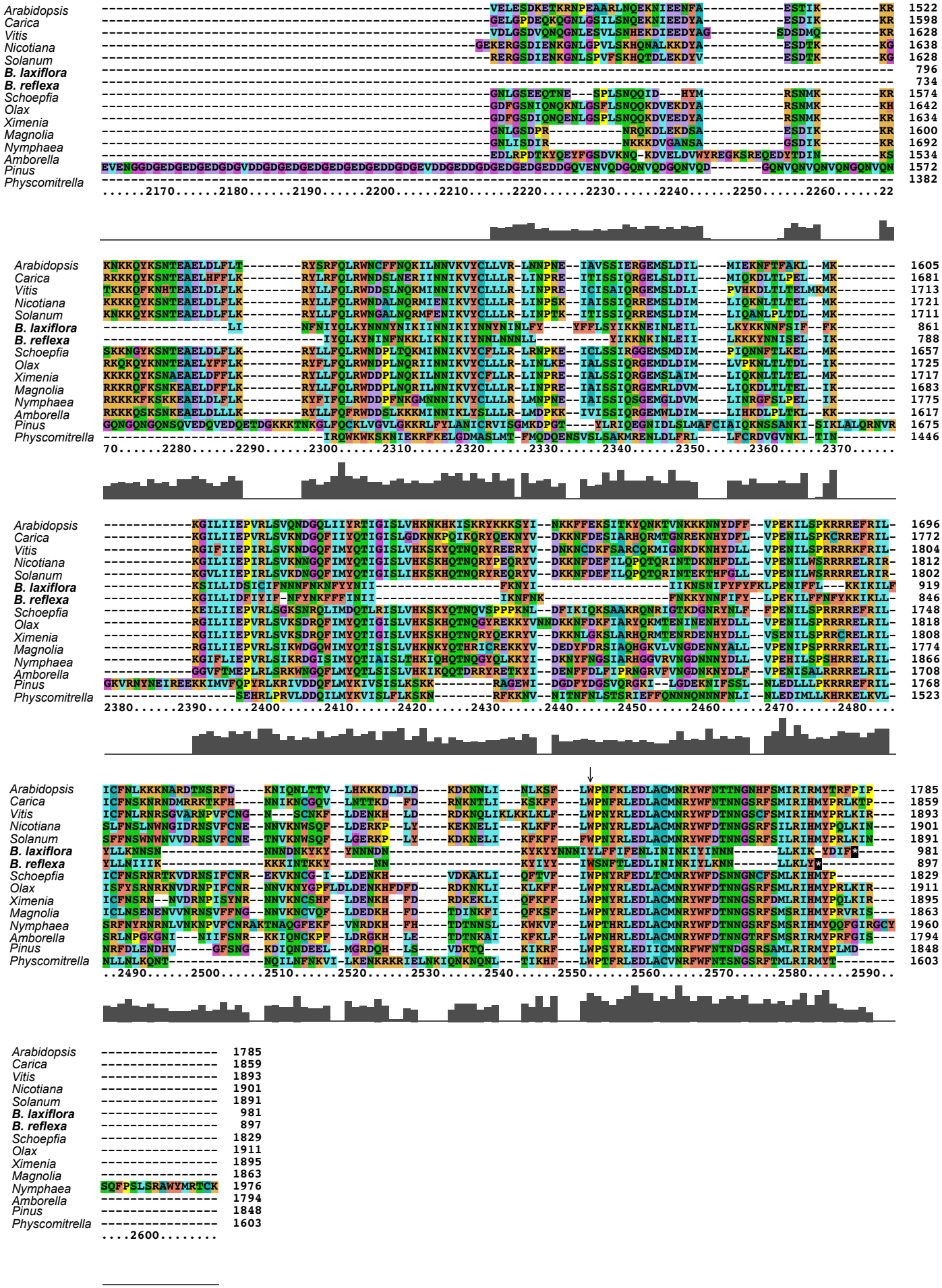


Figure S4. (cont.) Inferred amino acid alignments of the 15 protein genes present in *Balanophora* plastomes. Stop codons are shown only for *Balanophora* and are shaded in gray (TGA) or black (TAA). The arrows mark internal TAG codons present in one or both *Balanophora* species and inferred to encode W; note that most or all non-*Balanophora* land plants contain TGG (W in the standard genetic code) at most of these positions (Table 1). Regions of ambiguous alignment are marked with boxes and were excluded from the phylogenetic and molecular evolutionary rate analyses (open box, only the *Balanophora* sequences were excluded; gray box, the entire region was excluded). The histograms give the percentage of the aligned sequences that share the most common amino acid at each position.

YCF2

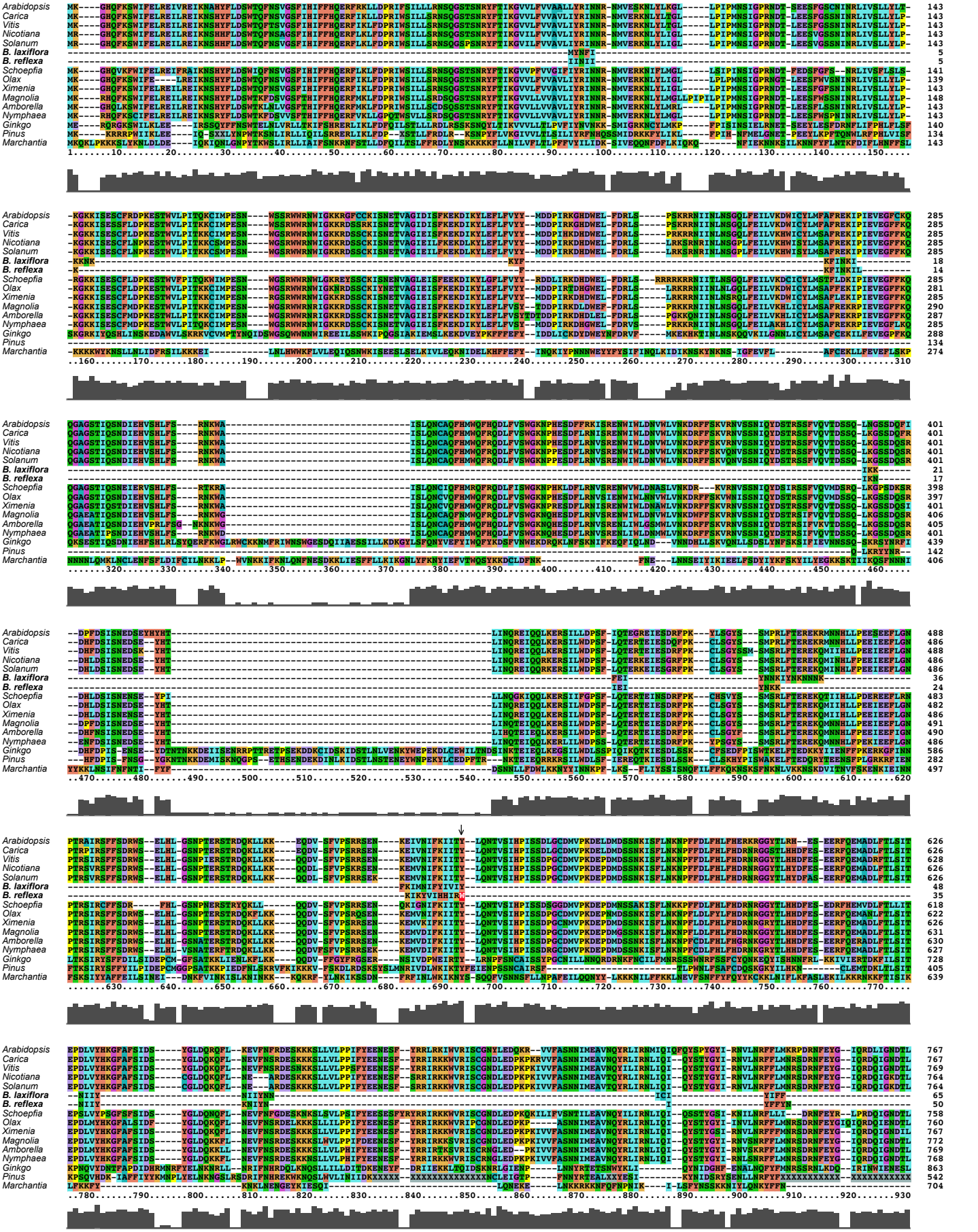


Figure S4. (cont.) Inferred amino acid alignments of the 15 protein genes present in *Balanophora* plastomes. Stop codons are shown only for *Balanophora* and are shaded in black to represent TAA. The arrow marks the internal TAG codon present in *B. reflexa* and inferred to encode W. Blue lines mark the three motifs identified in 1994 by Wolfe (3) as conserved between YCF2 and the CDC48 family of ATPases (*SI Appendix*, Results). The histograms give the percentage of the aligned sequences that share the most common amino acid at each position.

YCF2

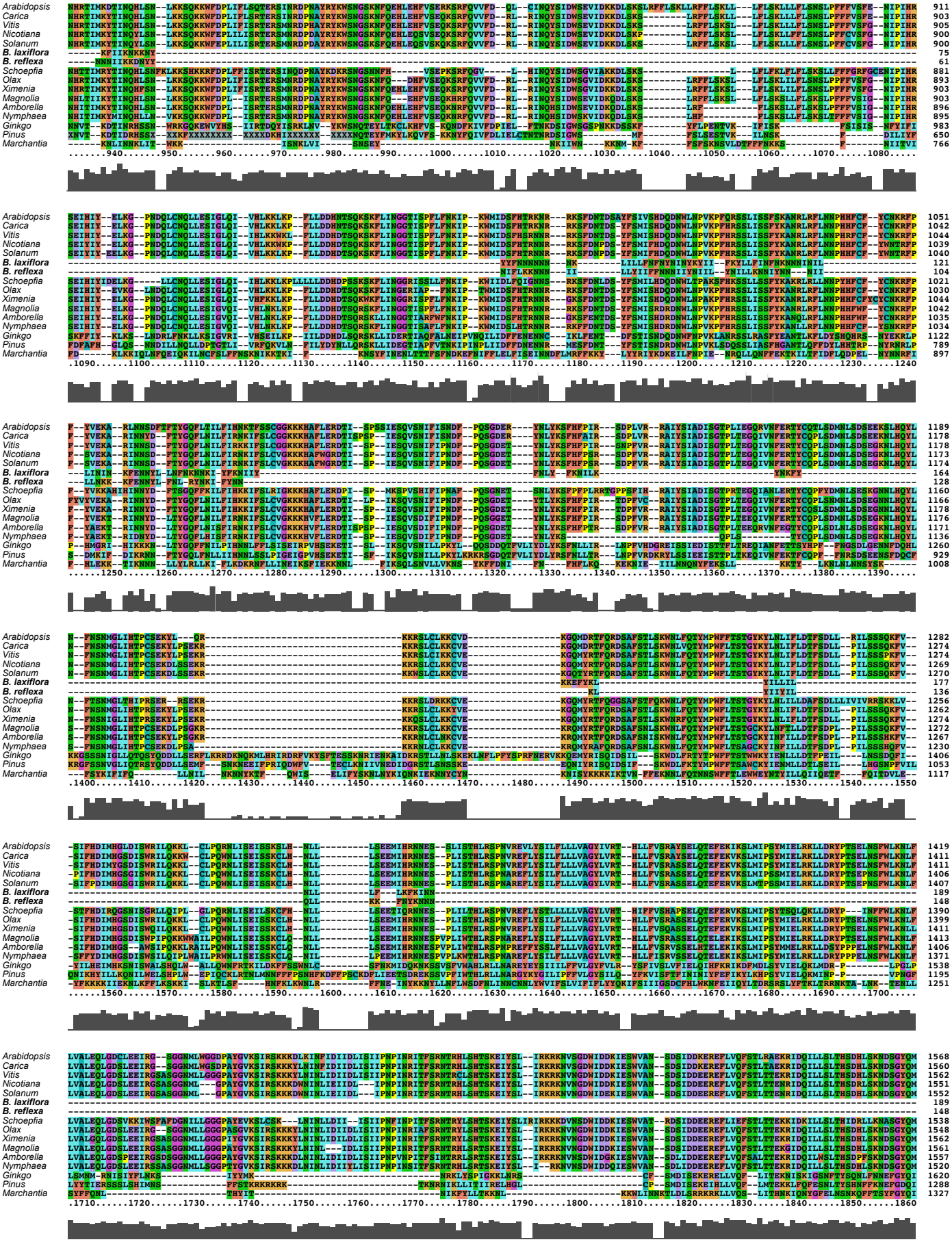


Figure S4. (cont.) Inferred amino acid alignments of the 15 protein genes present in *Balanophora* plastomes. Stop codons are shown only for *Balanophora* and are shaded in black to represent TAA. The arrow marks the internal TAG codon present in *B. reflexa* and inferred to encode W. Blue lines mark the three motifs identified in 1994 by Wolfe (3) as conserved between YCF2 and the CDC48 family of ATPases (*SI Appendix*, Results). The histograms give the percentage of the aligned sequences that share the most common amino acid at each position.

YCF2

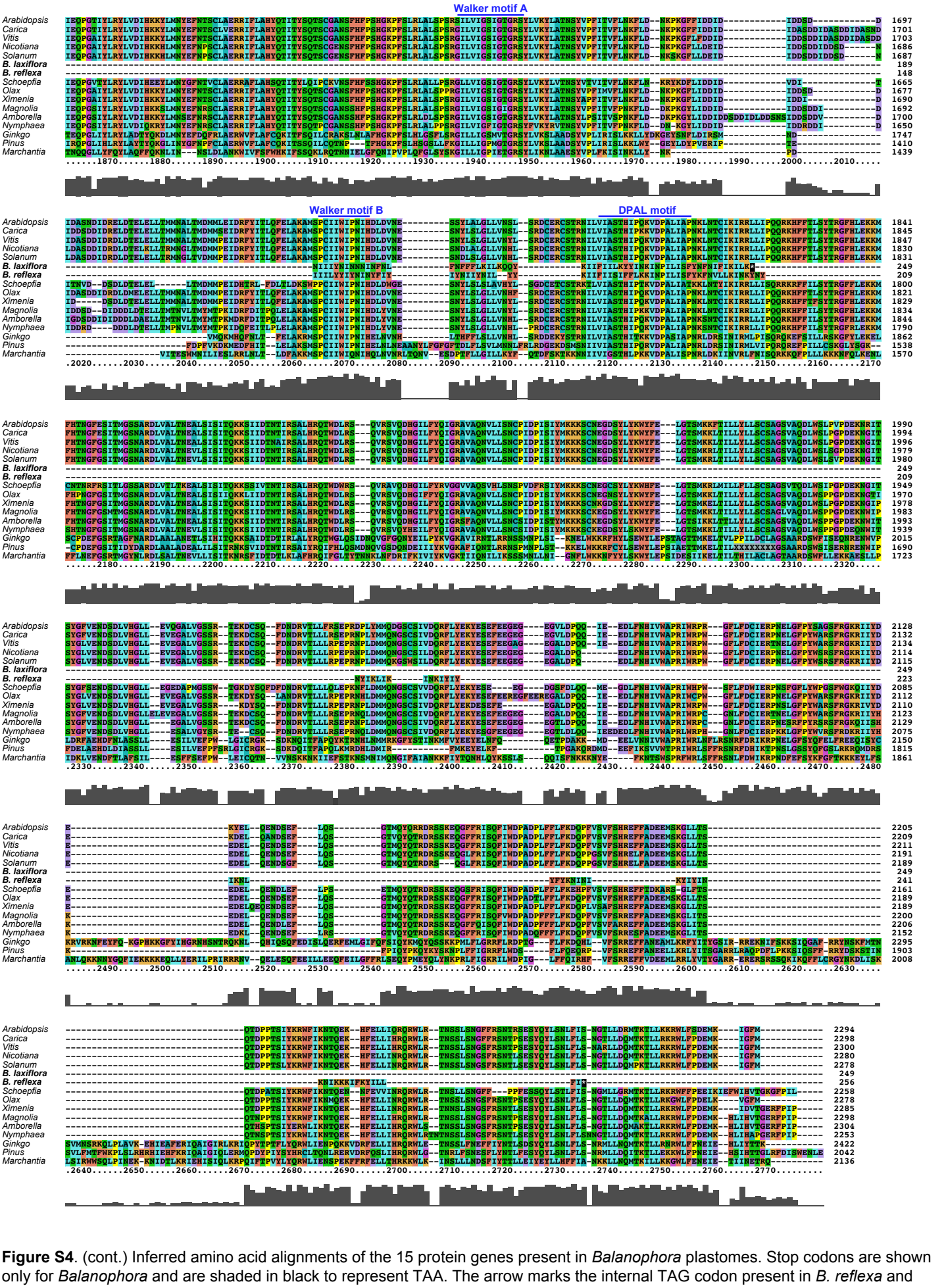


Figure S4. (cont.) Inferred amino acid alignments of the 15 protein genes present in *Balanophora* plastomes. Stop codons are shown only for *Balanophora* and are shaded in black to represent TAA. The arrow marks the internal TAG codon present in *B. reflexa* and inferred to encode W. Blue lines mark the three motifs identified in 1994 by Wolfe (3) as conserved between YCF2 and the CDC48 family of ATPases (*SI Appendix*, Results). The histograms give the percentage of the aligned sequences that share the most common amino acid at each position.

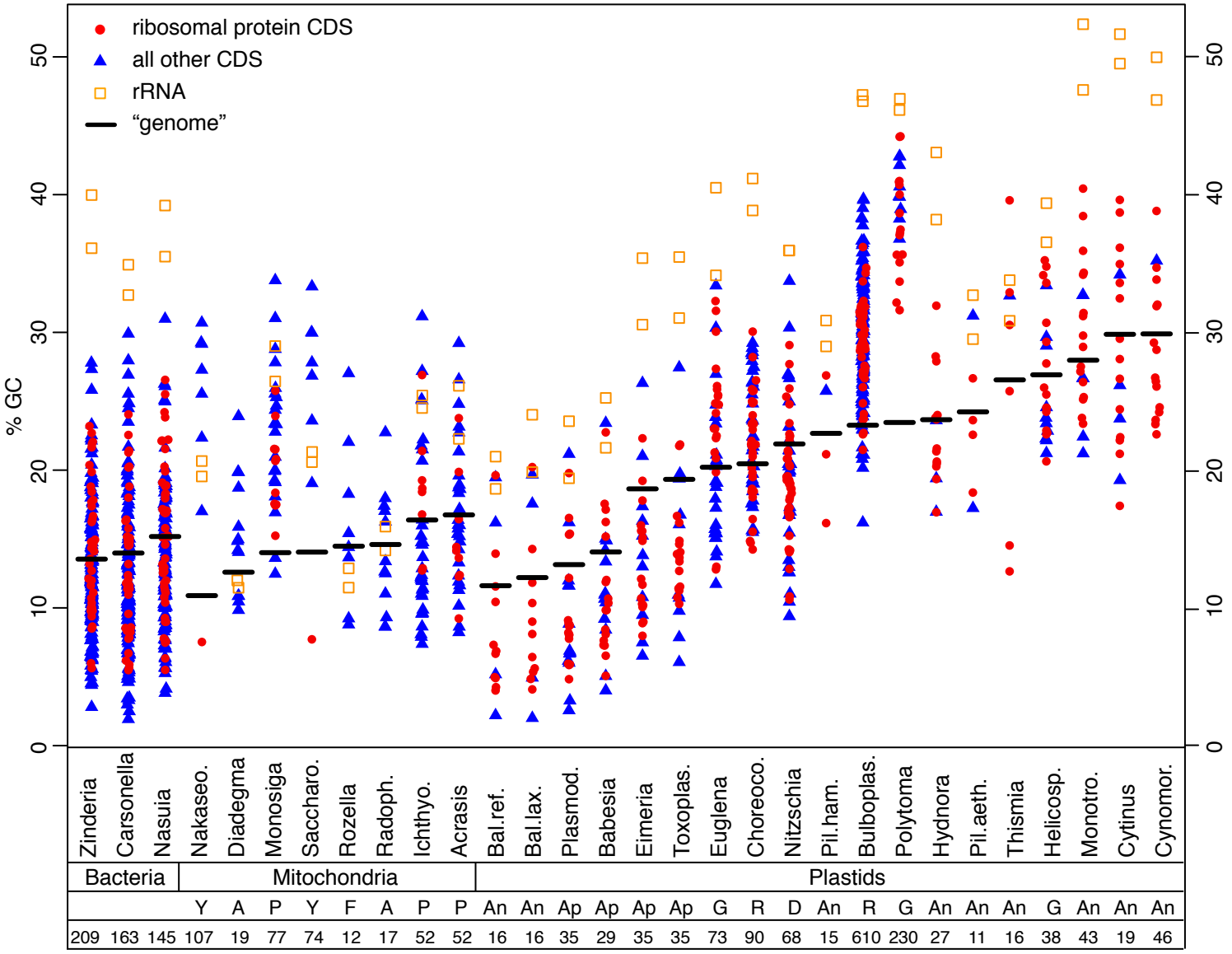


Figure S5. GC content of genes in 30 AT-rich bacterial and organellar genomes. All genes were analyzed except for tRNA, 5S rRNA, and 4.5S rRNA genes. The "genome" GC values include only one copy of the large, usually perfect repeats present in many plastomes, as these almost always contain rRNA genes, whose relative GC-richness will bias the full-genome GC values, especially for highly reduced genomes. Numbers at bottom are "genome" sizes in kb. Abbreviations: Y, yeast; A, animal; P, protist; F, fungus; An, angiosperm; Ap, apicomplexan; R, red alga; D, diatom; G, green alga or green-algal-derived. For full species names, see *SI Appendix* Table S12.

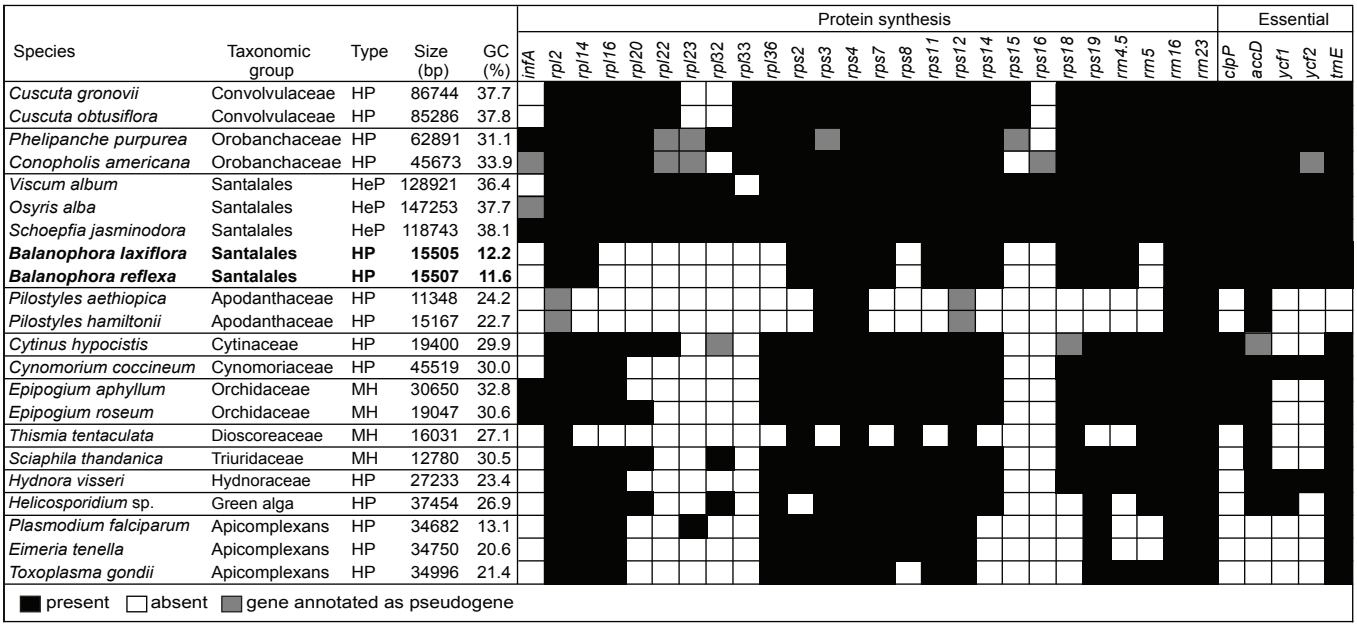


Figure S6. Plastome gene content of *Balanophora* compared with other plastome-containing heterotrophs. Abbreviations: HP, holoparasite; HeP, hemiparasite; MH, nonphotosynthetic mycoheterotrophs. The “essential” category of genes includes those five genes long suggested (21-23) to be the driving force, the *raison d'être*, for retention of the plastome in nonphotosynthetic angiosperms. Note that tRNA genes were excluded from the set of protein-synthesis genes shown for lack of space (plastomes of photosynthetic angiosperms contain 30 different tRNA genes) and because the *Balanophora* plastomes have only a single tRNA gene, *trn*E (see "Essential" category and main text).

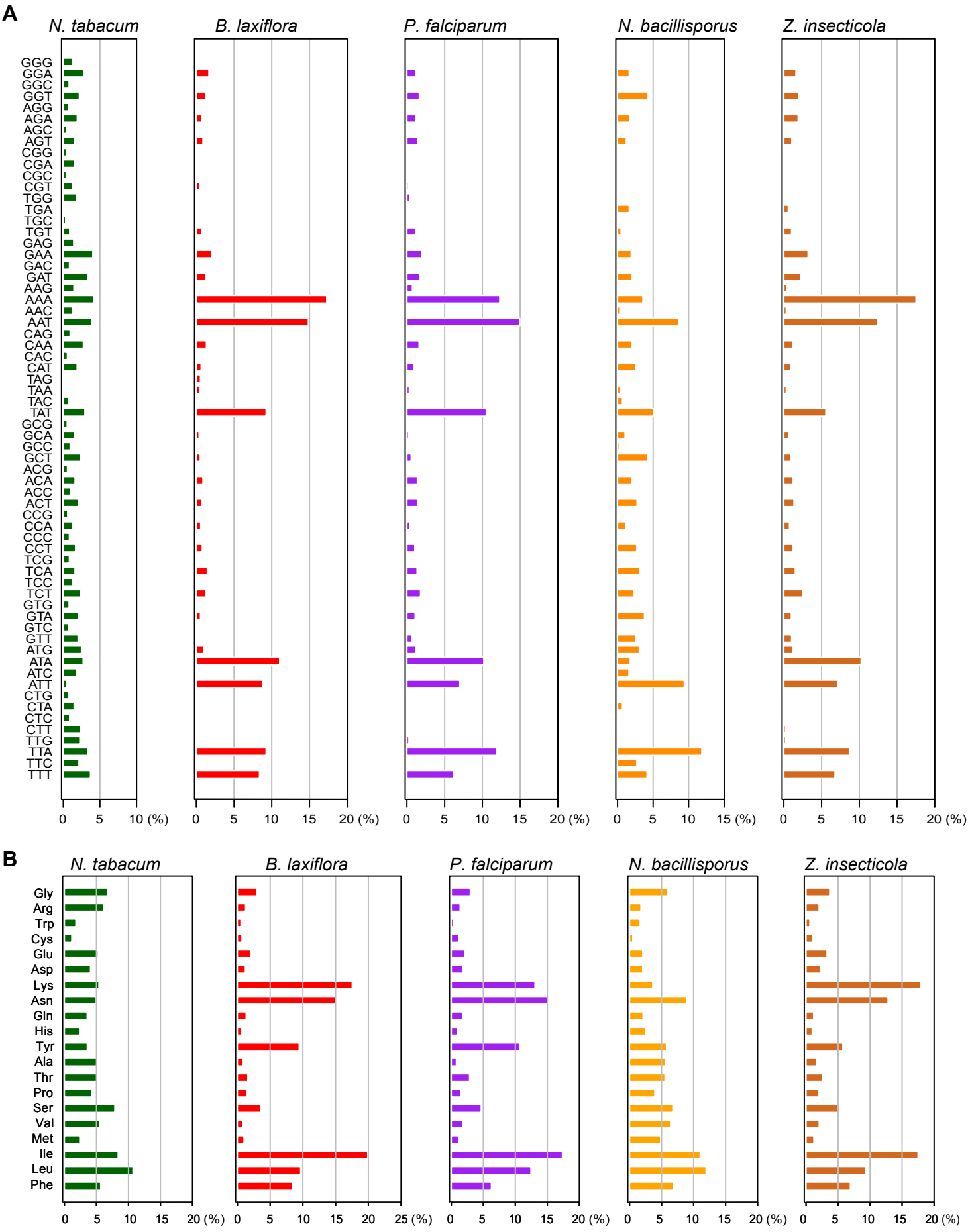


Figure S7. Codon *(A)* and amino acid *(B)* usage in the plastomes of *Nicotiana*, *Balanophora*, *Plasmodium*, the mitochondrial genome of *Nakaseomyces*, and the genome of the bacterium *Zinderia*. Percent usage frequency is shown by the bars. Amino-acid frequencies were calculated using the following translation tables: *Nicotiana* and *Plasmodium*, transl_table = 11; *Nakaseomyces*, transl_table = 3; *Zinderia*, transl_table = 4; *Balanophora*, a *Balanophora*-specific translation table that decodes TAG as Trp.

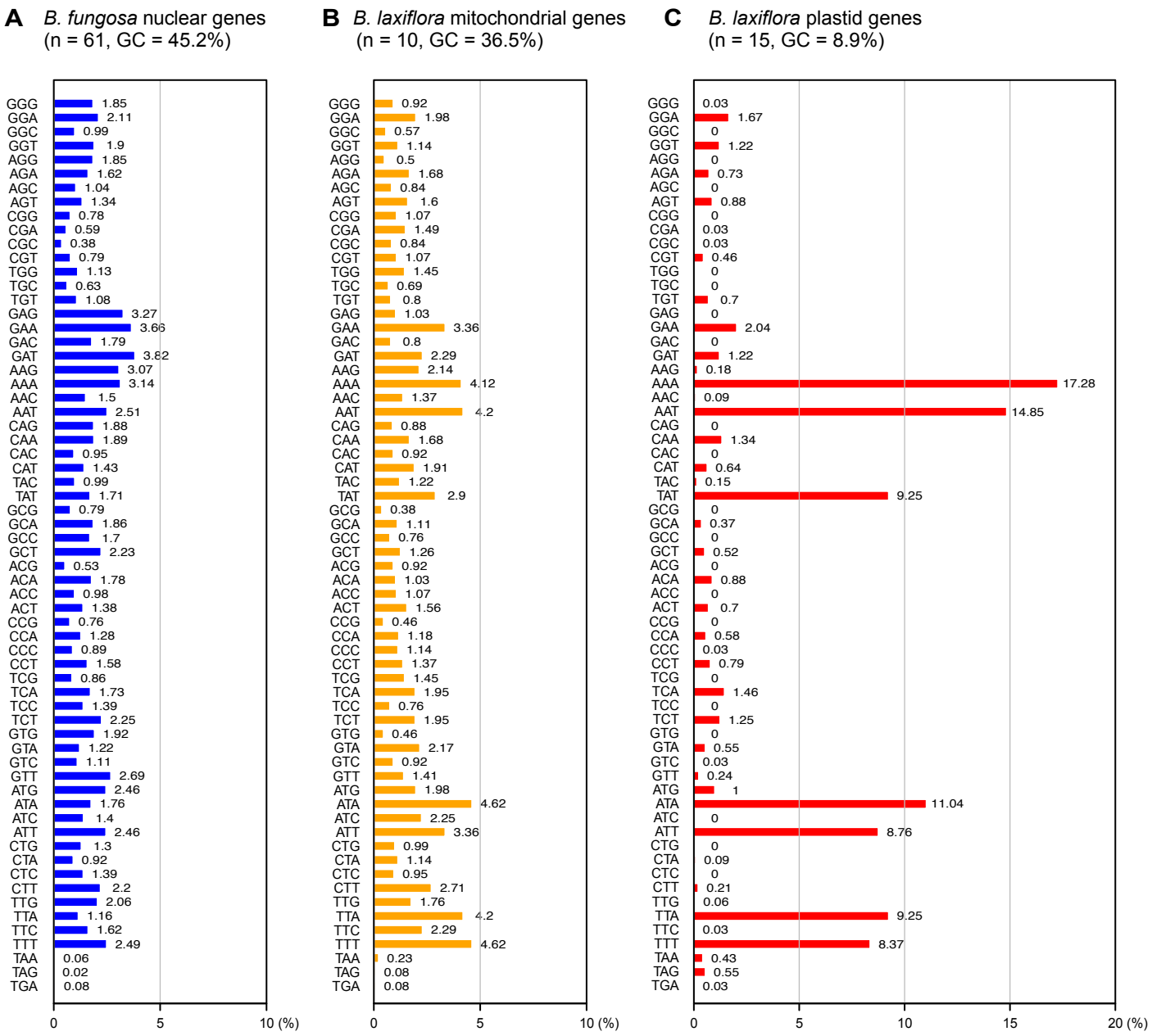


Figure S8. Comparison of codon usage in the *Balanophora* nuclear (*A*), mitochondrial (*B*), and plastid (*C*) genomes. Percent usage frequency is shown by the bars. See *SI Materials and Methods* for GenBank accession numbers for the mitochondrial and nuclear genes. Note that TAG is used as a Trp codon rather than a stop codon in the *B. laxiflora* plastid genes.

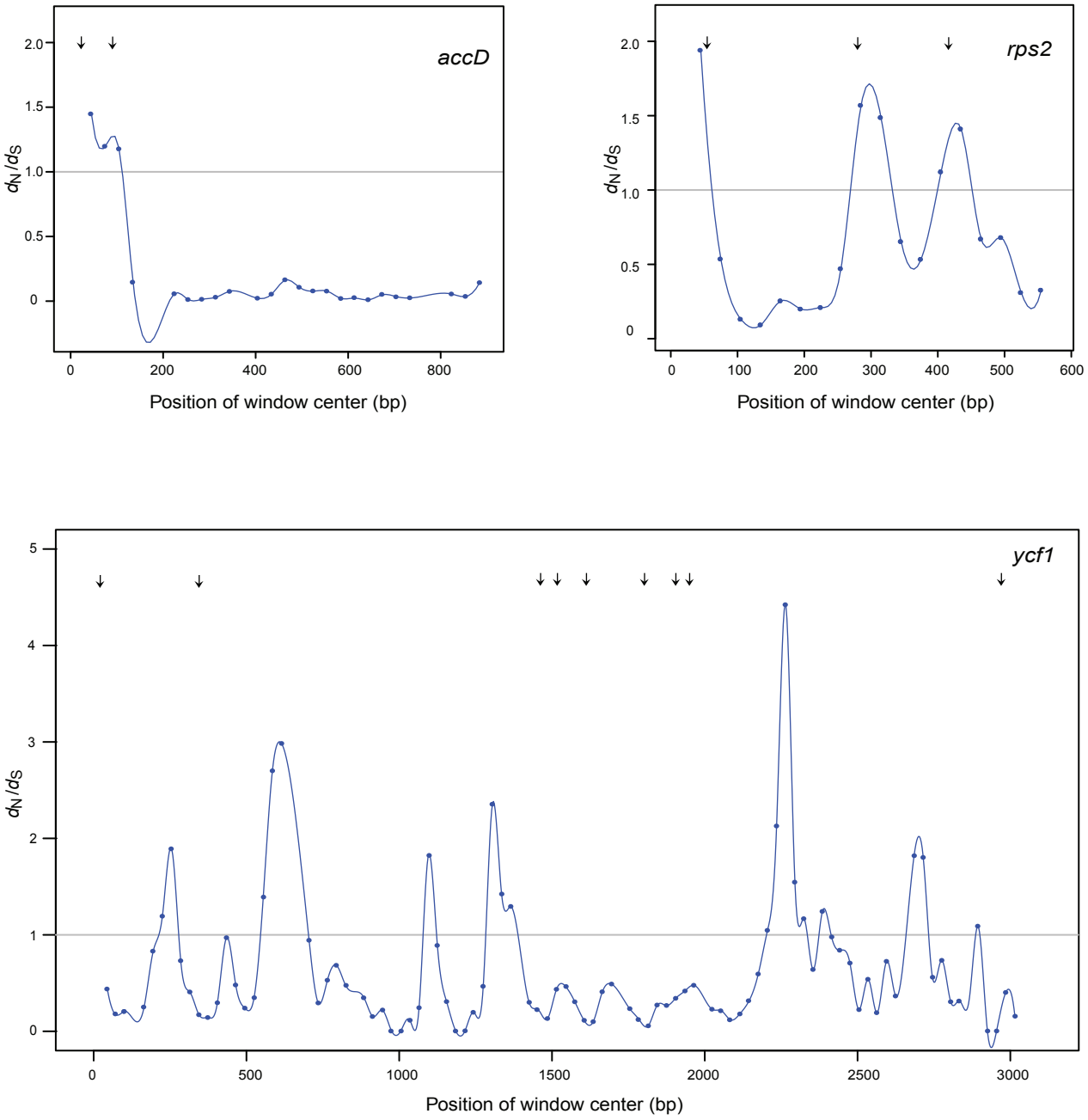


Figure S9. Pair-wise d_N/d_S ratios from sliding-window analysis (window size = 90 bp; step size = 30 bp) of the *Balanophora accD*, *rps2*, and *ycf1* genes (see Fig. 3C for the same analysis of two other *Balanophora* genes). Arrows mark internal TAG codons present in one or both *Balanophora* plastomes and inferred to encode W. d_N/d_S ratios for which $d_S < 0.01$ or $d_S = 99$ are not shown; these correspond to midpoints = 165, 195, 375, 765, and 795 bp for *accD*, and 135, 645, 675, 855, 975, 1005, 1185, 1215, 1395, 1725, 1995, 2655, 2925, and 2955 bp for *ycf*1.

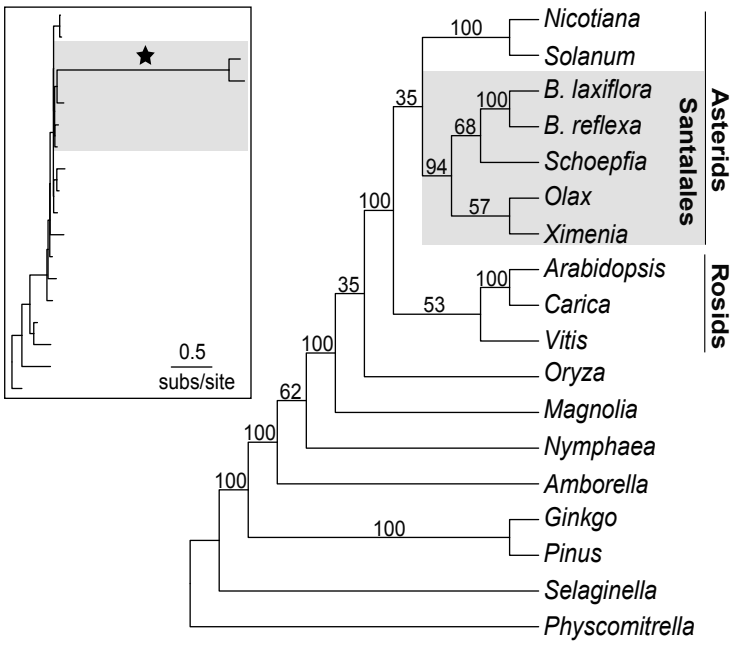


Figure S10. *Balanophora* plastid sequences are related to Santalales. The cladogram was estimated from a maximum likelihood analysis of 16 genes (all *Balanophora* genes except for *ycf*2, *rrn4.5* and *trnE*) from the 18 indicated land plants. Bootstrap support is shown above the branches. The star in the corresponding phylogram (see inset) marks the long branch leading to *Balanophora*.

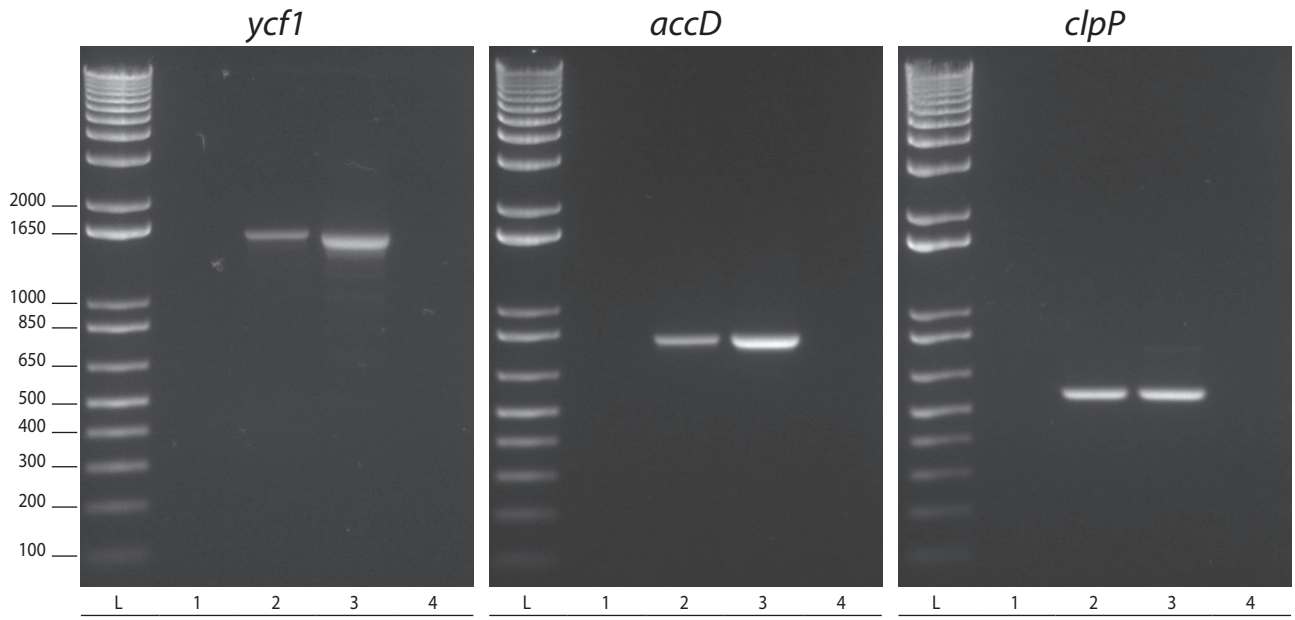


Figure S11. PCR analysis of three plastid genes and their transcripts. Templates used in PCR reactions were DNase-treated RNA (lanes 1), reverse transcribed DNase-treated RNA (2), RNase-treated DNA (3), and no template (4). The 1-kb-Plus DNA ladder was used as a length standard (lanes L).

YCF2 (angiosperms only)

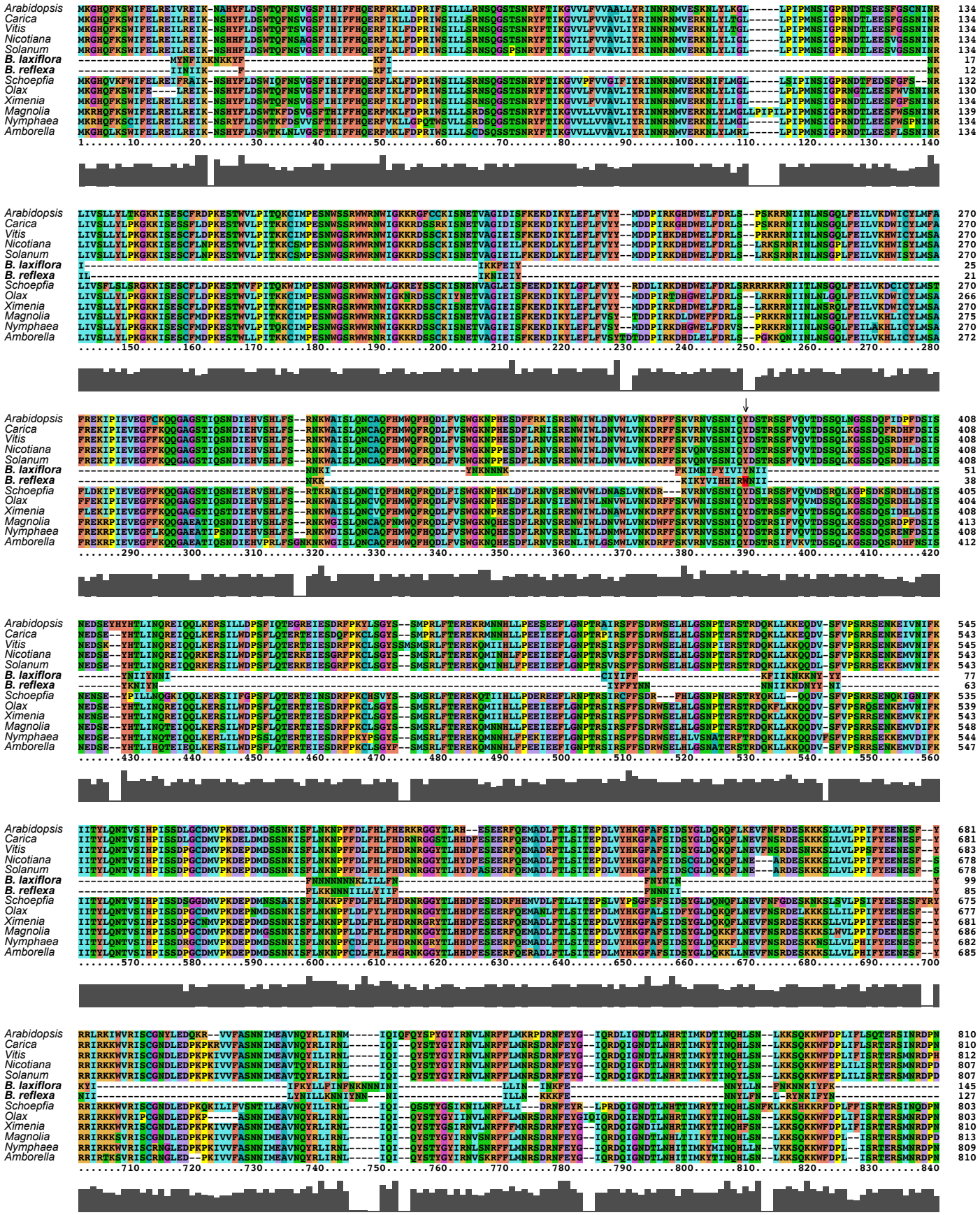


Figure S12. Inferred amino acid alignment of YCF2 from *Balanophora* and other angiosperms. Stop codons are shown only for *Balanophora* and are shaded in black to represent TAA. The arrow marks the internal TAG codon present in *B. reflexa* and inferred to encode W. Blue lines mark the three motifs identified in 1994 by Wolfe (3) as conserved between YCF2 and the CDC48 family of ATPases (*SI Appendix*, Results). The histograms give the percentage of the aligned sequences that share the most common amino acid at each position.

YCF2 (angiosperms only)

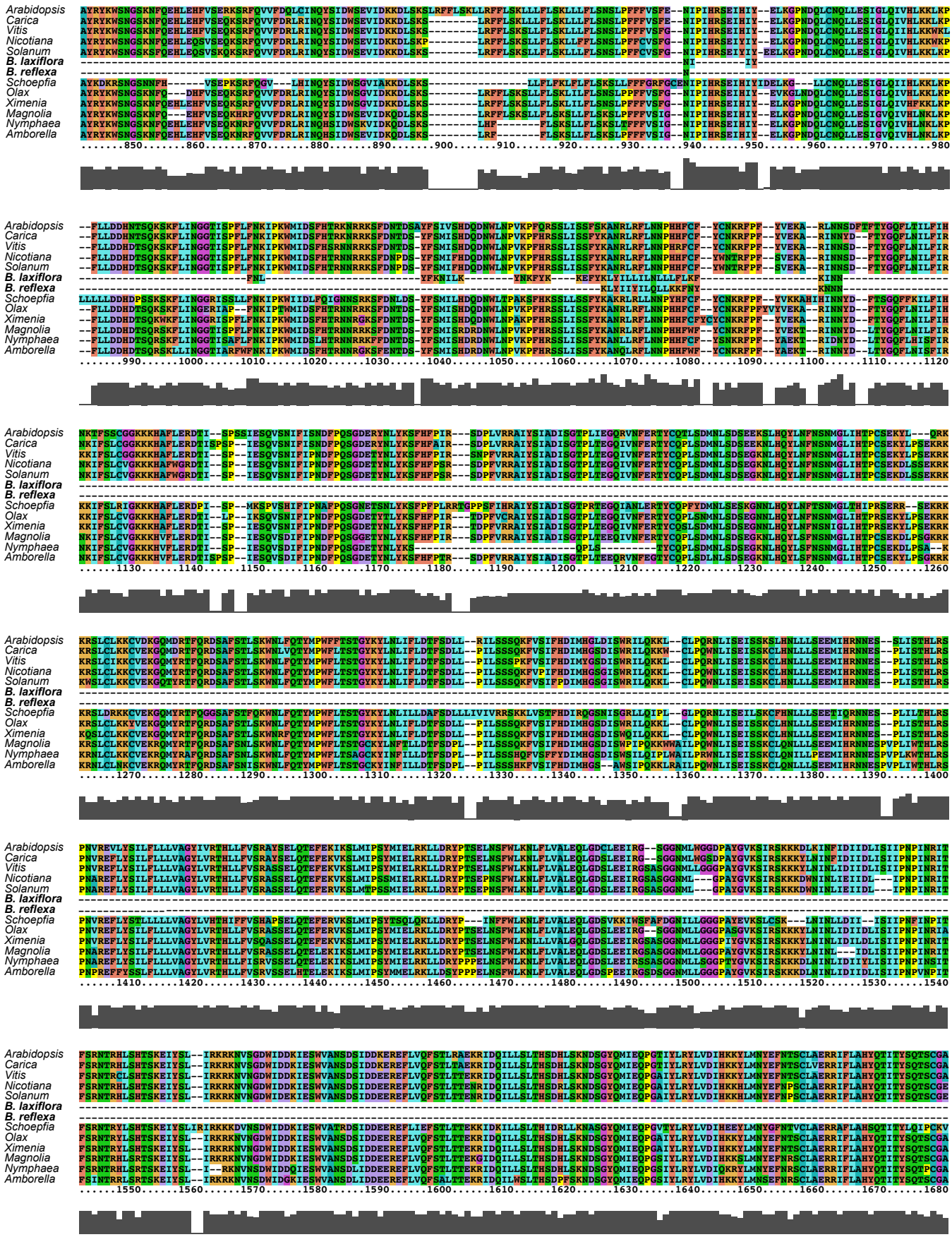


Figure S12. (cont.) Inferred amino acid alignment of YCF2 from *Balanophora* and other angiosperms. Stop codons are shown only for *Balanophora* and are shaded in black to represent TAA. The arrow marks the internal TAG codon present in *B. reflexa* and inferred to encode W. Blue lines mark the three motifs identified in 1994 by Wolfe (3) as conserved between YCF2 and the CDC48 family of ATPases (*SI Appendix*, Results). The histograms give the percentage of the aligned sequences that share the most common amino acid at each position.

YCF2 (angiosperms only)

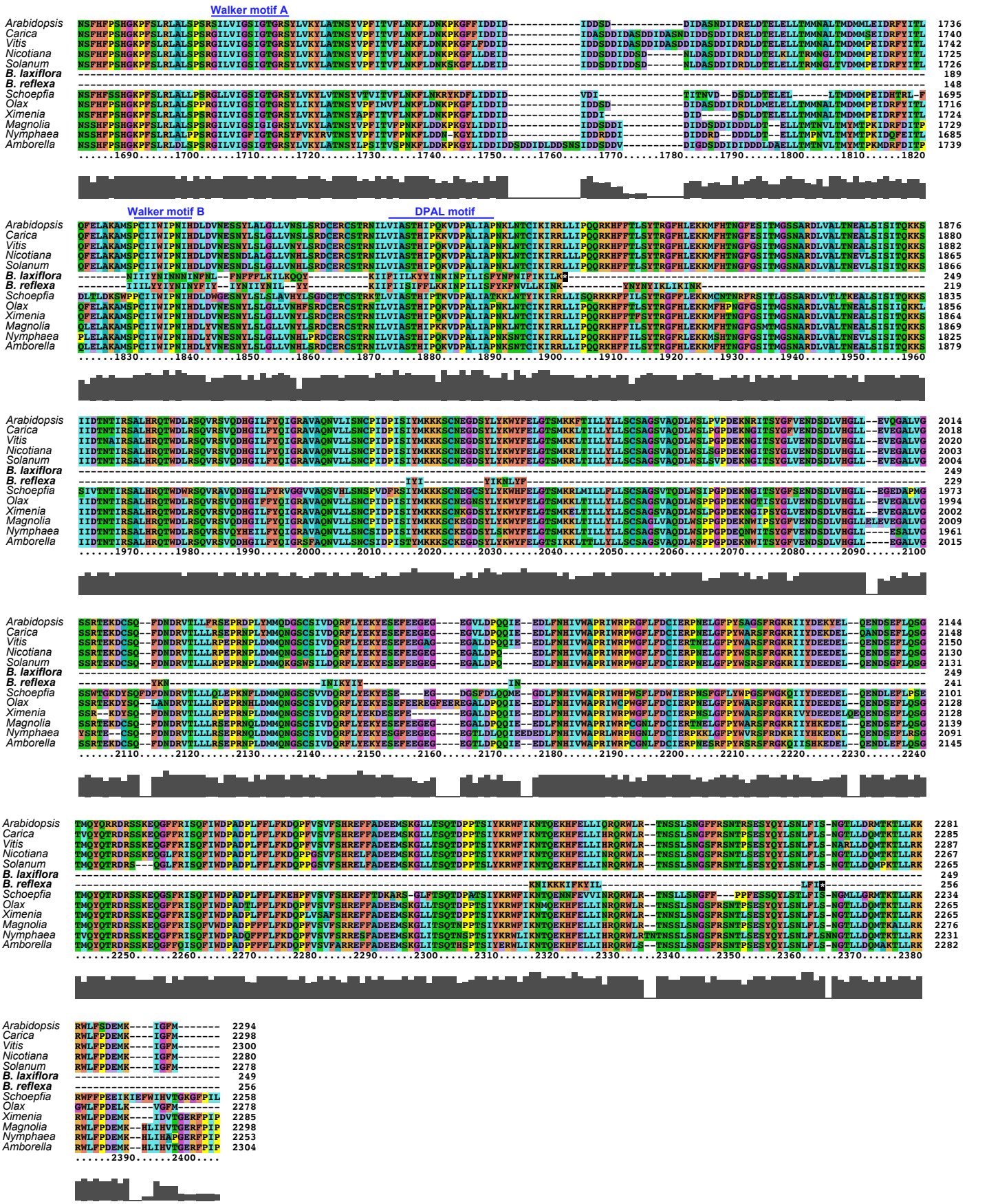


Figure S12. (cont.) Inferred amino acid alignment of YCF2 from *Balanophora* and other angiosperms. Stop codons are shown only for *Balanophora* and are shaded in black to represent TAA. The arrow marks the internal TAG codon present in *B. reflexa* and inferred to encode W. Blue lines mark the three motifs identified in 1994 by Wolfe (3) as conserved between YCF2 and the CDC48 family of ATPases (*SI Appendix*, Results). The histograms give the percentage of the aligned sequences that share the most common amino acid at each position.

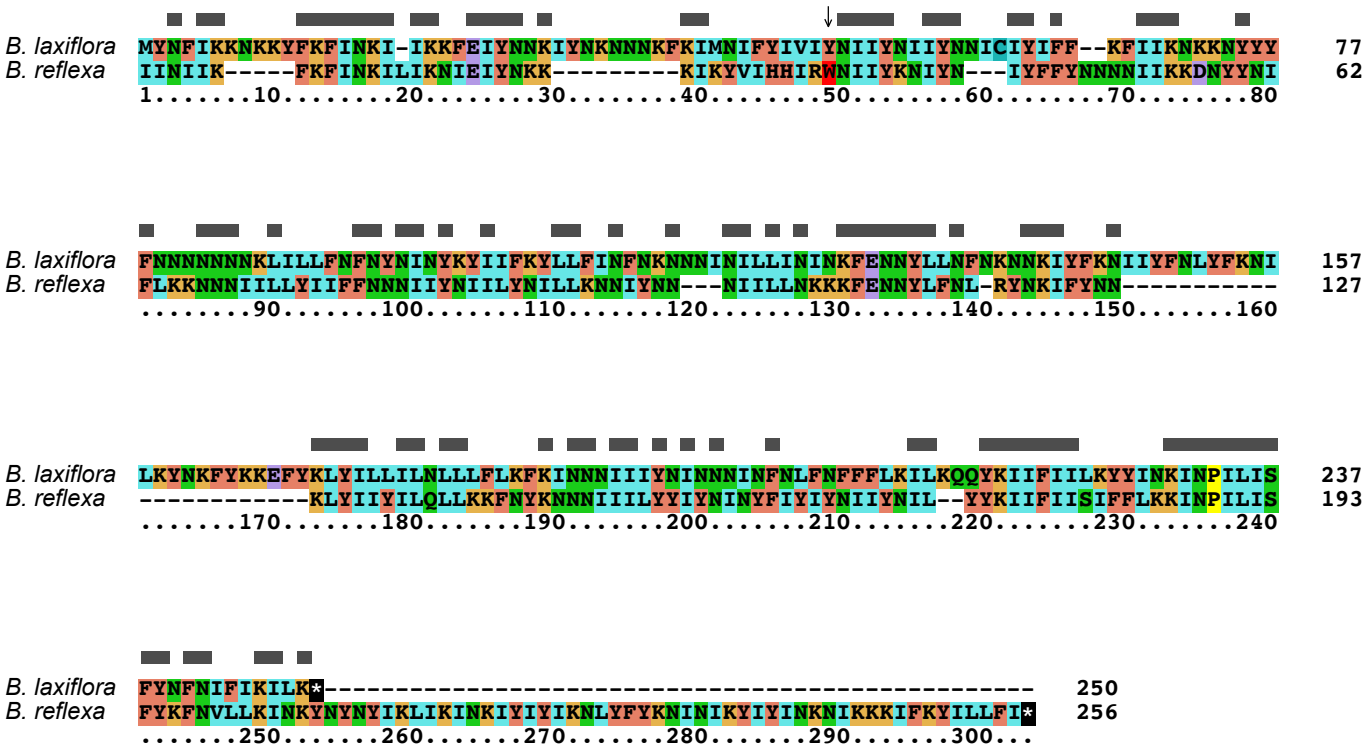


Figure S13. Pairwise alignment of YCF2 in *Balanophora*. Identities are marked by squares. The arrow marks the internal TAG codon present in *B. reflexa* and inferred to encode W. Stop codons are shaded in black (TAA).

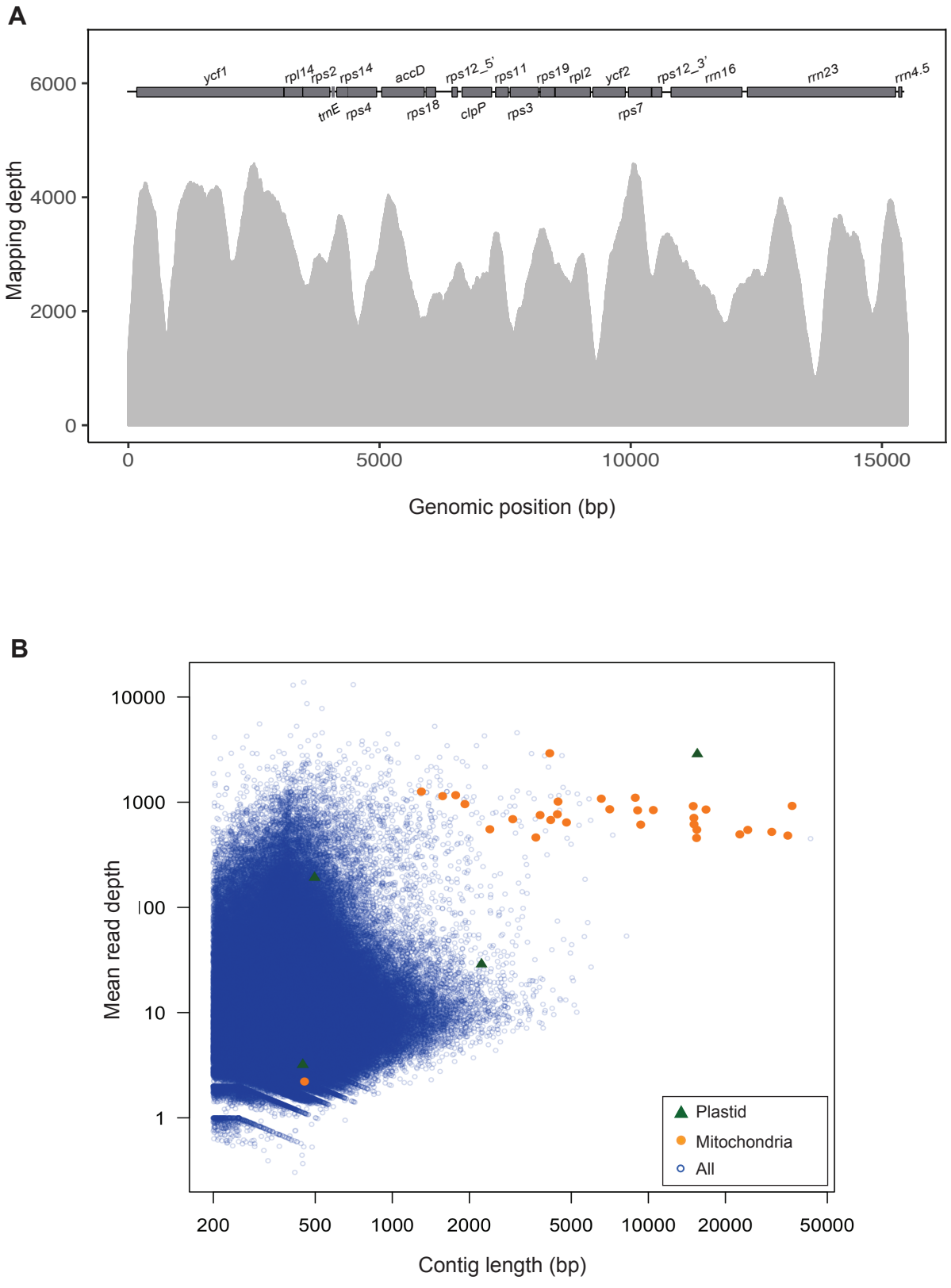


Figure S14. Read depth of the *B. laxiflora* assembly. (*A*) Coverage depth in the Velvet assembly of the raw Illumina reads based on a mapping quality of Q60. (*B*) Scatter plot of contig lengths for the CLC assembly versus mean read depth. The short plastid fragment is likely to be a nuclear fragment based on its low sequence coverage (see Materials and Methods).