Supplements

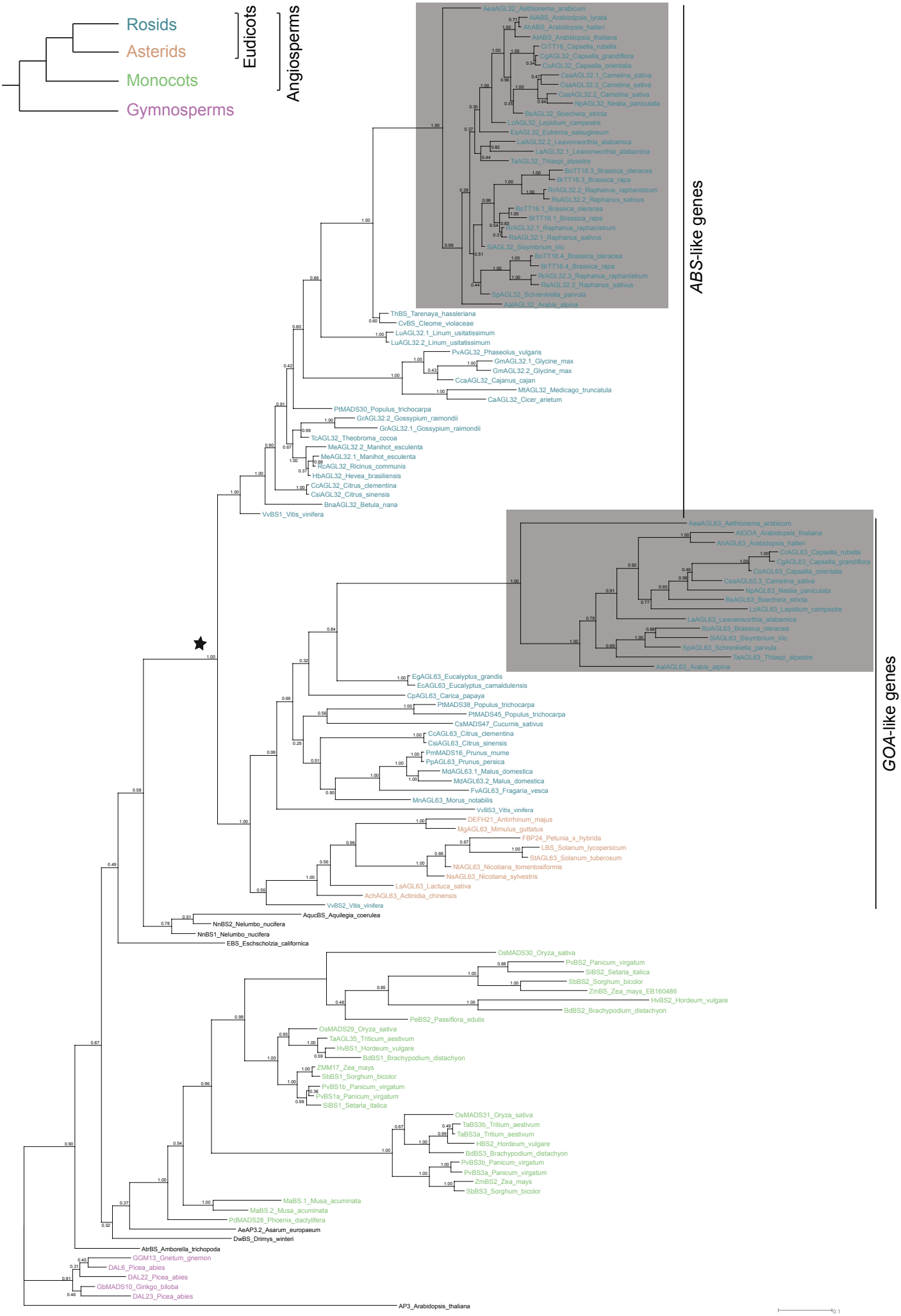
Supplemental Data Sets

Supplemental Data Set 1: List of sequences used in this study.

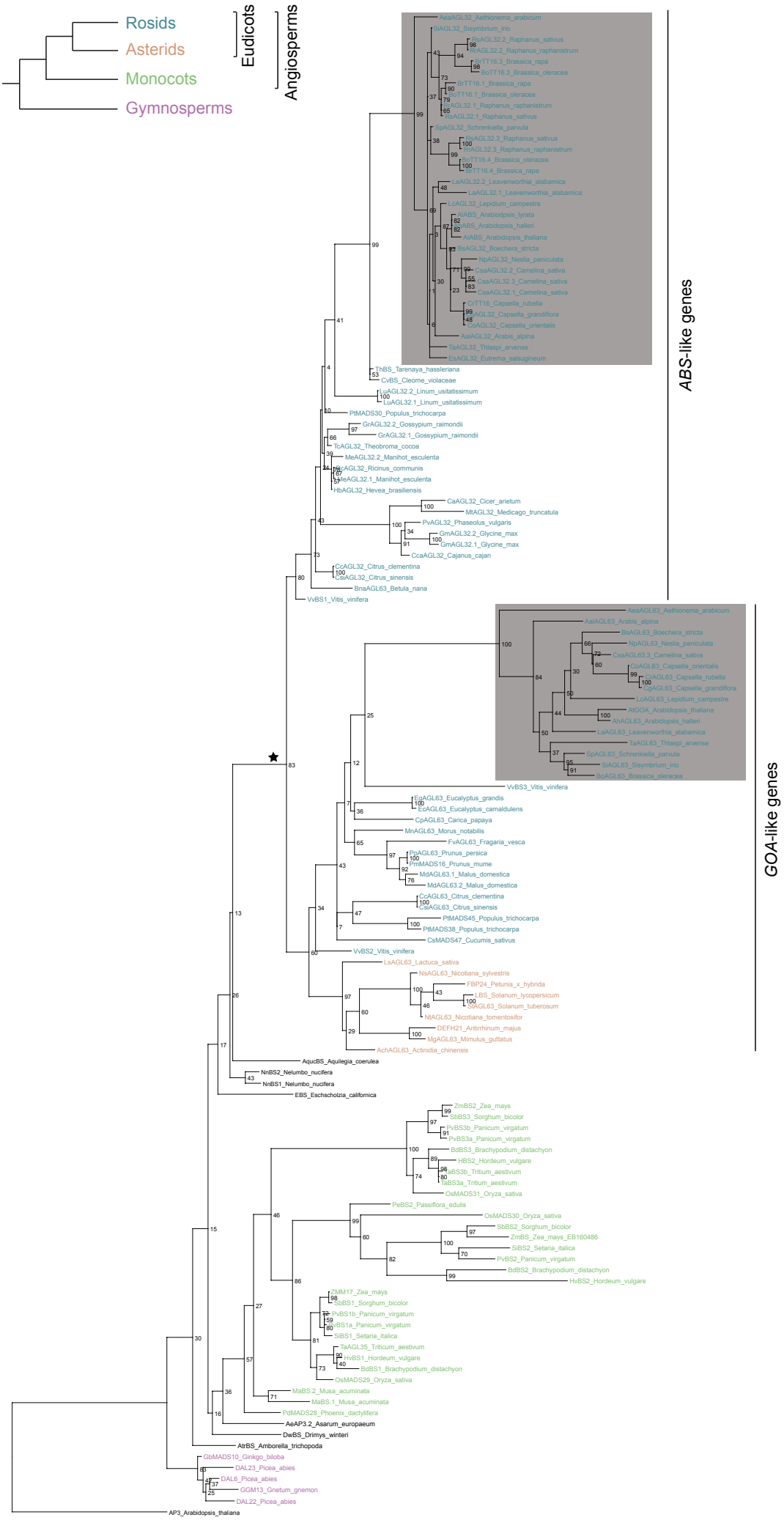
Supplemental Data Set 2: Primer Sequences for the amplification of *ABS*-like, *GOA*-like, and B_{sister} genes or cDNA.

see separate Excel file

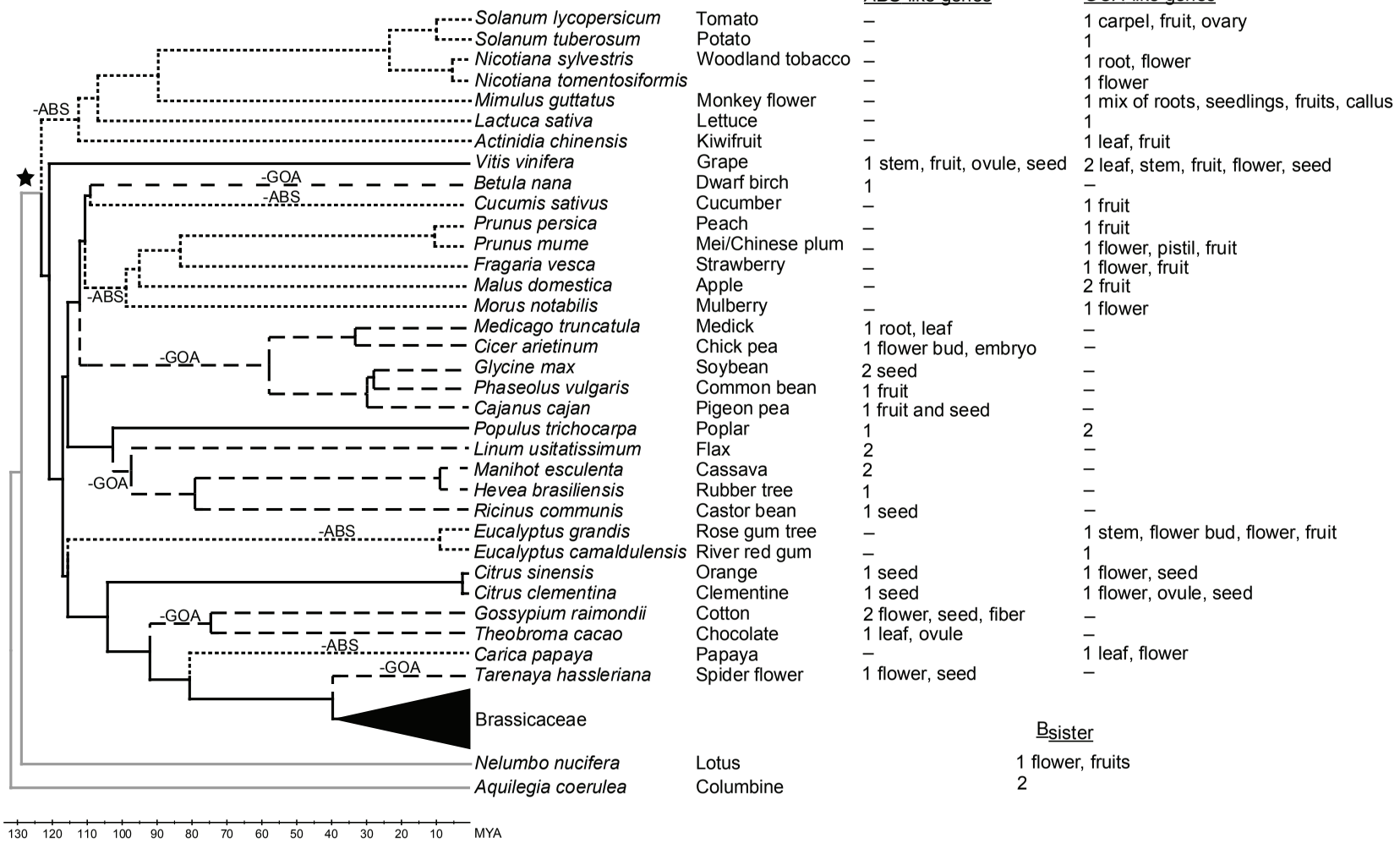
Supplemental Figures



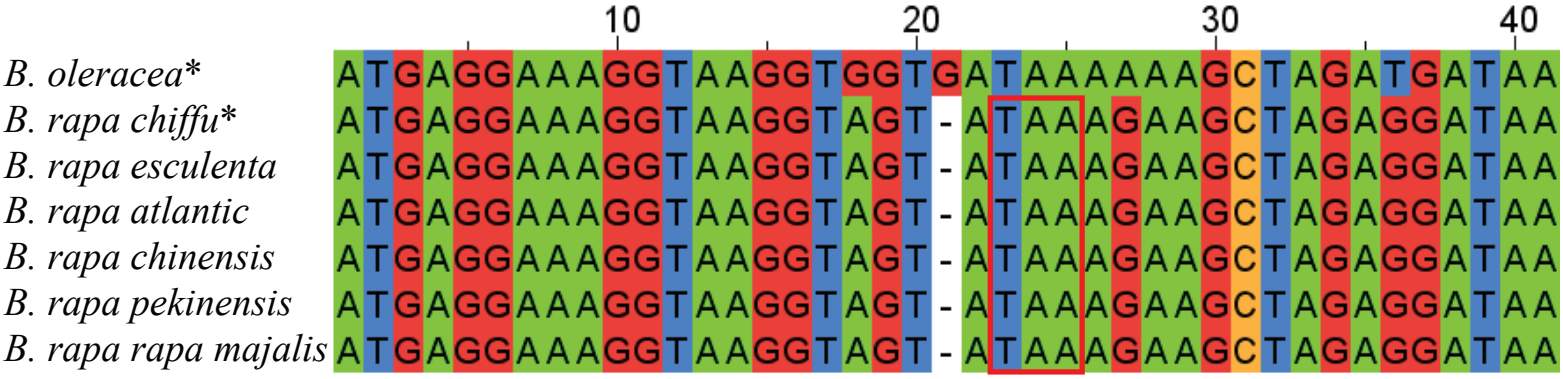
Supplemental Figure 1: Phylogeny of B_{sister} genes of seed plants as reconstructed by Bayesian inference based on protein sequences. The duplication event of the B_{sister} genes resulting in *ABS*- and *GOA*-like genes is indicated by a star. The clades of *ABS*- and *GOA*-like genes are marked on the right. The *ABS*- and *GOA*-like genes of Brassicaceae are indicated by shading. APETALA3 (AP3) was used as representative of the outgroup. Posterior probabilities of the MrBayes phylogeny are shown at the corresponding nodes. The branch lengths are drawn proportionally to the number of substitutions. Upper left corner: simplified phylogeny of seed plants with a color code for the large phylogeny; purple: gymnosperm genes, green: genes from monocots, red: genes from Asterids, blue: genes from rosids.



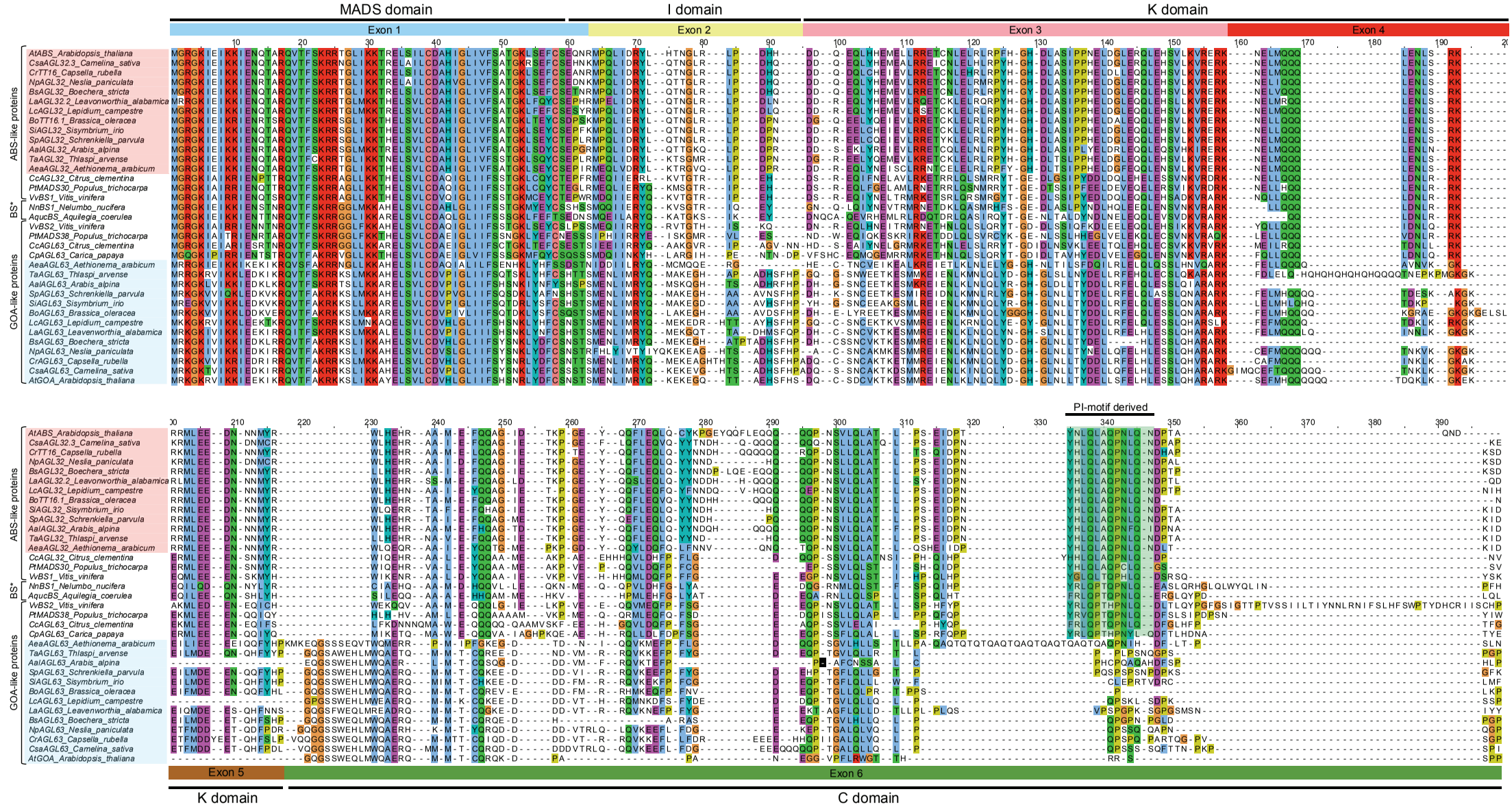
Supplemental Figure 2: Maximum Likelihood phylogeny of B_{sister} genes in seed plants reconstructed based on protein sequences. The duplication event of the B_{sister} genes resulting in *ABS*- and *GOA*-like genes is indicated by a star. The clades of *ABS*- and *GOA*-like genes are marked on the right. The *ABS*- and *GOA*-like genes of Brassicaceae are indicated by shading. *APETALA3* (*AP3*) was used as representative of the outgroup. Bootstrap values are shown at the corresponding nodes. The branch lengths are drawn proportionally to the number of substitutions.



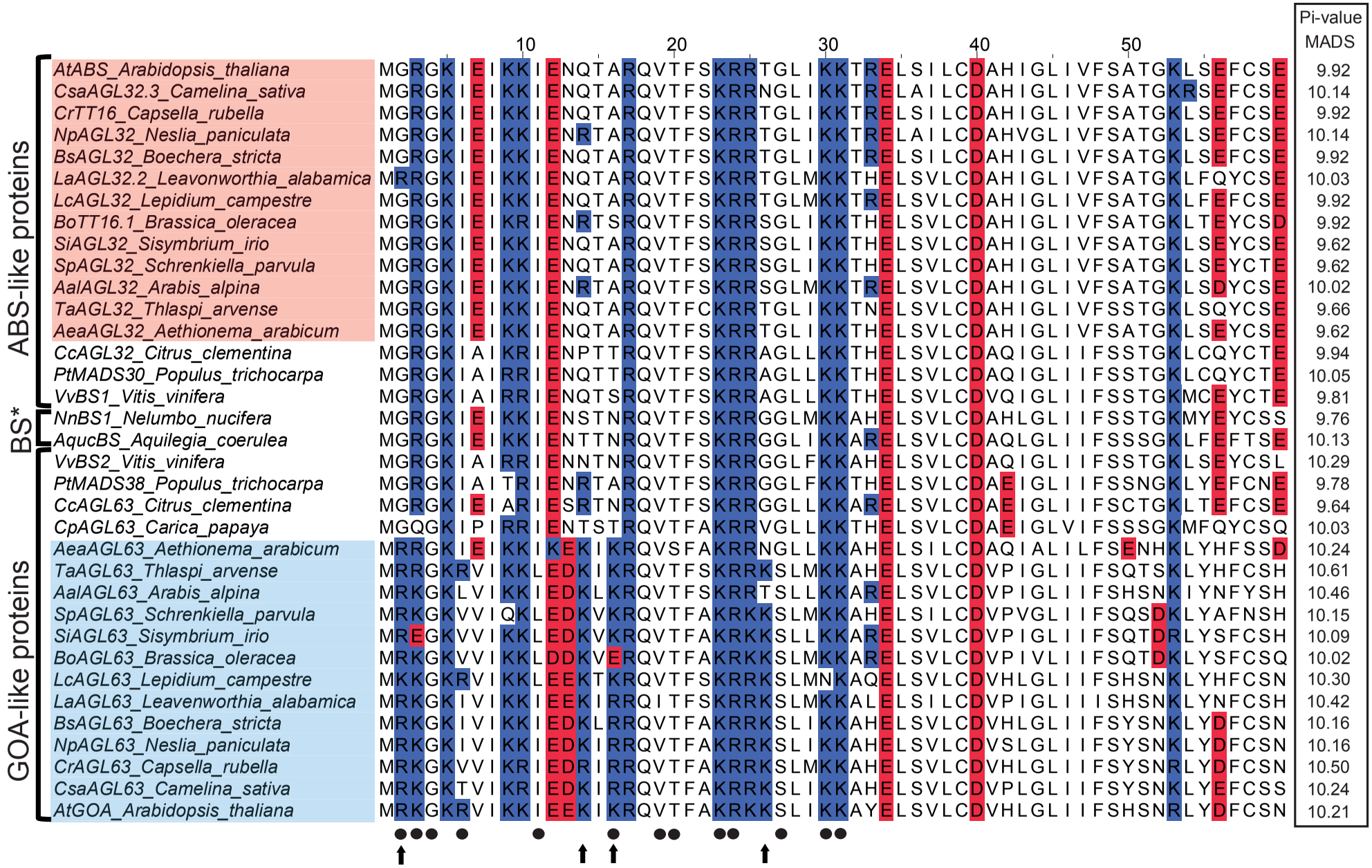
Supplemental Figure 3: Expression pattern of *ABS*- and *GOA*-like genes of different eudicots. The phylogeny was drawn according to (Magallón et al., 2015). Branches leading to lineages having both *ABS*- and *GOA*-like genes are shown in solid black lines; branches leading to lineages which do not have *GOA*-like genes are shown in broken lines; branches leading to lineages which do not have *ABS*-like genes are shown in dotted lines. Loss of *ABS*- and *GOA*-like genes is marked by –ABS and –GOA, respectively on the corresponding branches. Estimated divergence times (in MYA, Million Years Ago) are given below the phylogeny as determined by (Magallón et al., 2015). Branches leading to lineages which diverged before the duplication of *ABS*- and *GOA*-like genes are shown in gray. The stem group in which the duplication event of *ABS*- and *GOA*-like genes likely happened, is indicated by a star. For simplification, the Brassicaceae species are combined to a triangle. The numbers next to the species indicate the number of identified *ABS*- and *GOA*-like genes. The tissue of gene expression was identified by NCBI BLAST searches in the following short read archives (SRA): *Nicotiana sylvestris* (ERX248671, ERP002501) (Sierro et al., 2013); *Nicotiana tomentosiformis* (ERX248672, ERX248675) (Sierro et al., 2013); *Actinidia chinensis* (SRX219919, SRX219918)(Huang et al., 2013); *Cucumis sativus* (SRX732893, SRX7328934) (Zhao et al., 2016); *Prunus persica* (SRX173255); *Prunus mume* (SRX177763); *Fragaria vesca* (SRR1107994, SRX030874, SRX030797) (Shulaev et al., 2011; Kang et al., 2013); *Morus notabilis* (SRR1210501) (Li et al., 2014); *Medicago truncatula* (SRX2235516, SRX1998644) (Afkhami and Stinchcombe, 2016; Bond et al., 2016); *Cicer arietinum* (SRX208033, SRX125162) (Singh et al., 2013; Pradhan et al., 2014); *Glycine max* (SRR1777405) (Song et al., 2013); *Phaseolus vulgaris* (SRX963056) (Astudillo-Reyes et al., 2015); *Cajanus cajan (*SRR4341980); *Eucalyptus grandis* (SRX367258); *Theobroma cacao* (SRX278083, SRX277980) (Motamayor et al., 2013); *Carica papaya* (SRX1770718, SRX1770817) (Urasaki et al., 2012); *Tarenaya hassleriana* (SRX393170, SRX466643) (Bhide et al., 2014; Külahoglu et al., 2014) and *Aquilegia coerulea* (SRX152572, SRX152573, SRX152575). Expression data of the B_{sister} MADS-box genes from *Vitis vinifera* were taken from (Grimplet et al., 2016), of the *GOA*-like gene from *Eucalyptus grandis* was taken from (Dias et al., 2005), of the B_{sister} genes from the two Citrus species from (Hou et al., 2014). The accession numbers for the cDNA data from which the expression pattern for the B_{sister} genes of the remaining species was inferred can be found in the Supplemental Data Set 1.



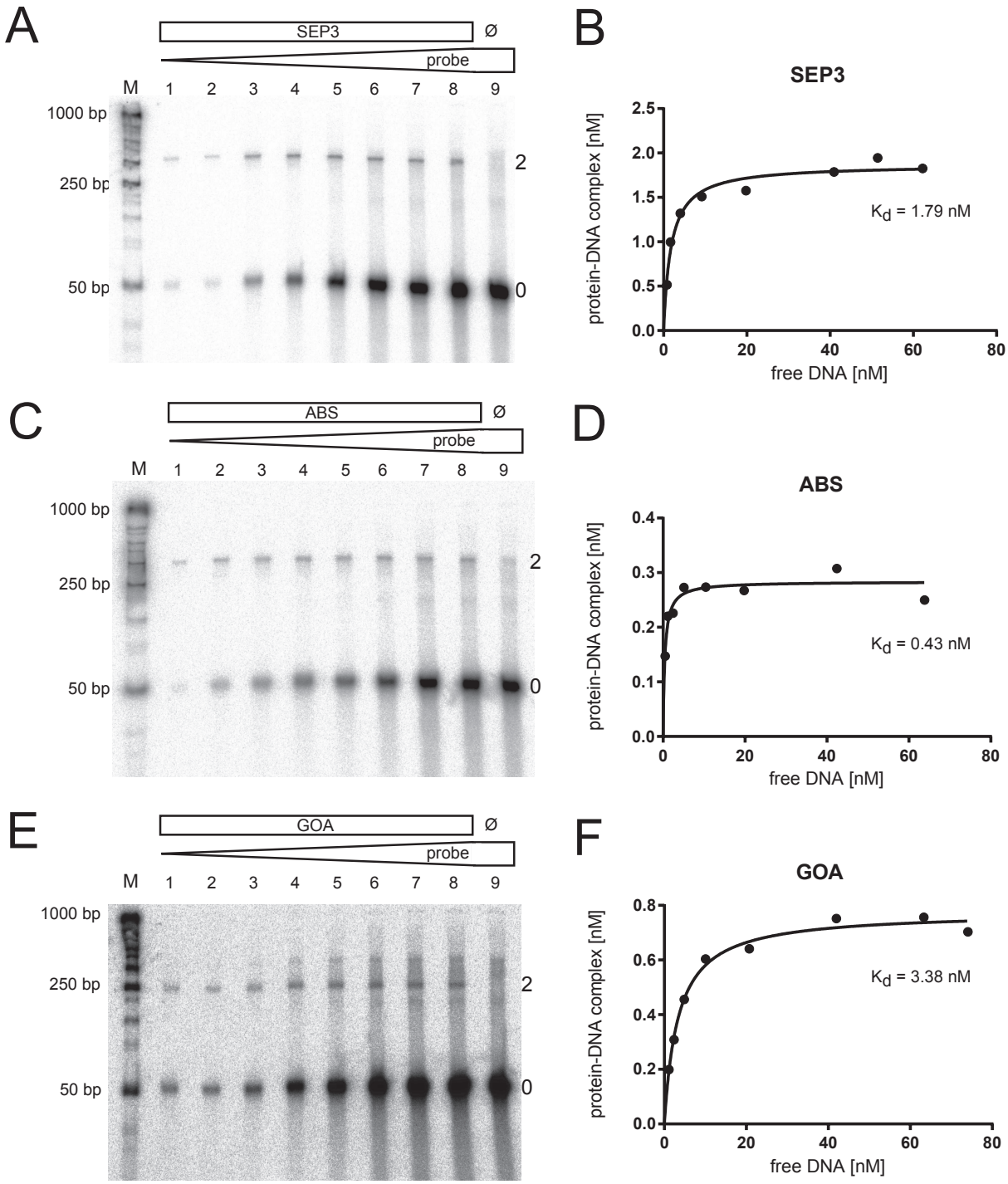
Supplemental Figure 4: Alignment of the 5' part of the MADS-box sequences of the GOA orthologs of *Brassica oleracea* and of different *Brassica rapa* subspecies. All analyzed *Brassica rapa* subspecies have a nonfunctional GOA ortholog due to a deletion at position 21 leading to a translational frameshift and a premature stop codon (indicated by a red rectangle). A star indicates the sequences identified in the published genomes.



Supplemental Figure 5: Alignment of B_{sister} proteins of basal eudicots (BS*) and ABS- and GOA-like proteins. The names of ABS- and GOA-like proteins of Brassicaceae are shaded in light red and light blue, respectively. The exons coding for the protein sequences are given in different colors above and below the alignment. The numbering of exons is according to Figure 6. The domains of the MIKC-type MADS domain proteins are indicated by black lines above and below the alignment. The derived PI motif (encoded by exon 6, highlighted in green) is present in all proteins except for GOA-like proteins from Brassicaceae. Amino acids are colored according to the ClustalW coloring scheme (Thompson et al., 1997) to highlight conservation.



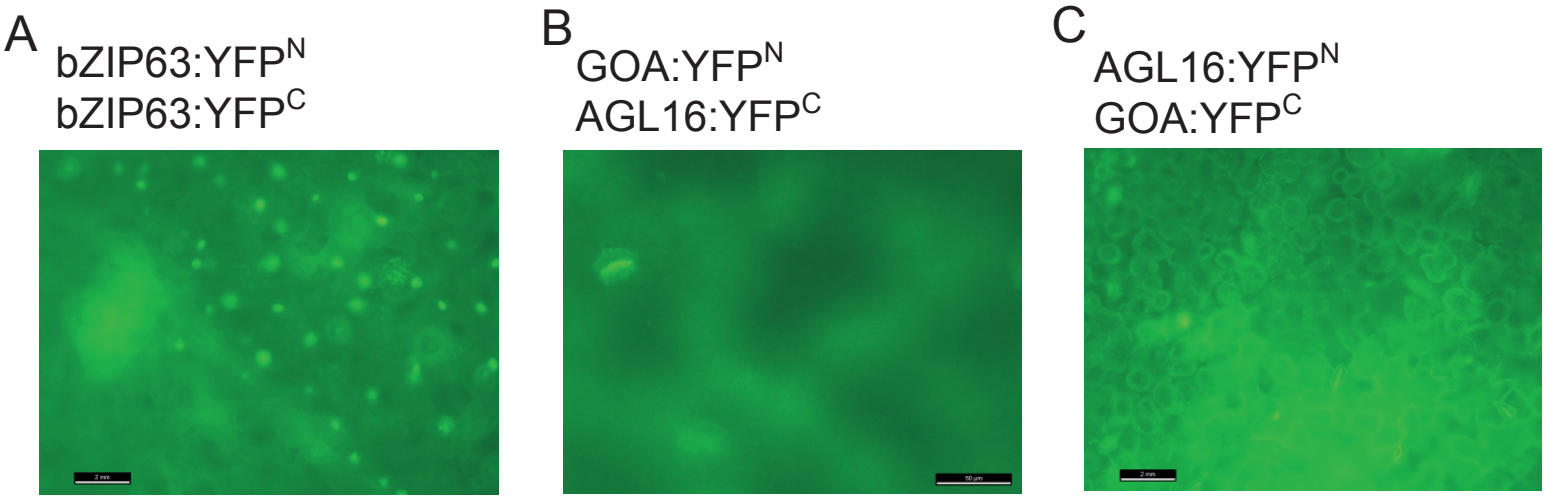
Supplemental Figure 6. Alignment of the MADS domains of B_{sister} proteins of basal eudicots (BS*) and of ABS- and GOA-like proteins. ABS- and GOA-like proteins of Brassicaceae species are shaded in light red and light blue, respectively. Positively and negatively charged amino acids are colored in blue and red, respectively. The pI values as calculated by the program ProtParam (Gasteiger et al., 2005) are given next to the sequences. Additional positively charged amino acid positions for GOA-like proteins of Brassicaceae are indicated by a black arrow. Black dots indicate residues important for DNA-binding in SRF and MEF2A (Pellegrini et al., 1995; Huang et al., 2000).



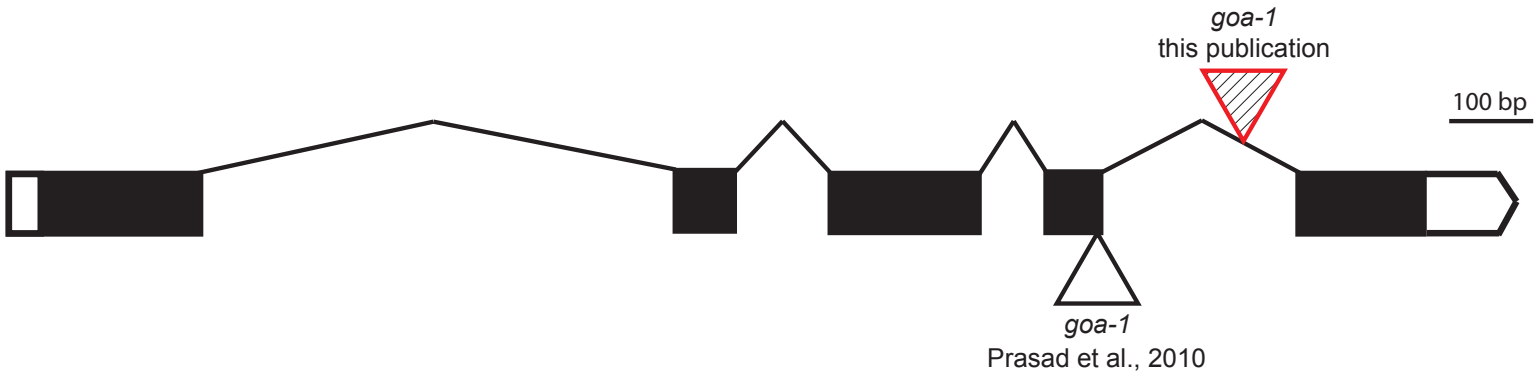
G

	SEP3	ABS	GOA
Experiment 1	1.14 nM	0.32 nM	3.38 nM
Experiment 2	1.79 nM	0.16 nM	2.26 nM
Experiment 3	2.42 nM	0.43 nM	4.87 nM
Experiment 4		0.55 nM	
Mean K_d ± SD	1.78 ± 0.52 nM	0.37 ± 0.14 nM	3.50 ± 1.07 nM

Supplemental Figure 7. Analysis of DNA binding of SEP3 (A, B), ABS (C, D) and GOA (D, E) by electrophoretic mobility shift assays (EMSA). Increasing concentrations of a DNA probe containing a single CArG-box were incubated with a constant amount of protein. The protein studied is indicated on top of each gel. Ø represents a negative control where the *in vitro* translation assay was performed using an empty pTNT vector. M indicates the DNA marker lane. The concentrations of DNA used were: 1.31, 2.62, 5.23, 10.47, 20.94, 41.87, 52.34, and 62.81 nM from lane 1 to 8. The numbers 0 and 2 adjacent to the bands represent the respective number of protein molecules bound. Next to each gel the concentration of the protein-DNA complex is shown as a function of the concentration of free DNA and the inferred dissociation constants (K_d) are given. Graphs were calculated as described in materials and methods. The chosen gels are only an example from several independently performed experiments, for which dissociation constants are shown in (G).



Supplemental Figure 8: Protein dimerization analysis by bimolecular fluorescence complementation (BiFC). bZIP63 (positive control), GOA and AGL16 were fused to the N- and C-terminal parts of YFP (YFP^N and YFP^C) as indicated above each subfigure. Fluorescence of nuclearly localized YFP visible as bright green dots is considered as indication for the interaction of the fusion proteins. Dimerization of bZIP63 **(A)** but no interaction between GOA and AGL16 **(B, C)** was observed. Chloroplast and cell wall auto-fluorescence are visible in several subfigures. Bars in **(A)** and **(C)** = 200 μm; bar in **(B)** = 50 μm.



Supplemental Figure 9: Location of the T-DNA insertion of the *goa-1* mutant line (*A. thaliana*). Using PCR and sequencing the position of the T-DNA of the *goa-1* mutant was found to be within intron 4, but not as described by Prasad et al. (2010) in exon 3.

Supplemental Tables

Supplemental Table 1: Phylogenetic Analysis by Maximum Likelihood (PAML) results and results of the Likelihood Ratio Test for selection analysis. lnL, Likelihood; np, number of parameters; ω0 and ω1, ω ratios, 2ΔlnL, twice the difference in the Likelihood values for the two models tested; Δnp, difference in the number of parameters; p, p value

PAML results for different models tested on *ABS*- and *GOA*-like genes							
	lnL	np	ω_0	ω_1	ω_2	ω_3	ω_3
one ratio	-11546.39	55	0.2833				
three ratio	-11513.43	57	0.1961	0.2159	0.4411		
five ratio	-11509.16	59	0.1983	0.2306	0.1293	0.4586	0.2202
free-ratio	-11445.00	107					

Results of the Likelihood Ratio Test for different models tested on *ABS*- and *GOA*-like genes.			
	2ΔlnL	Δnp	p
one vs. three	65.9266	2	4.83E-15
one vs. five	74.4595	4	2.59E-15
one vs. free	202.7796	52	1.11E-19

Supplemental Table 2: Pairwise amino acid identities in percent of B_{sister} proteins of basal eudicots and of ABS- and GOA-like proteins. Values were calculated using the SIAS (Sequences Identities and Similarities) server. Pairwise identities of ABS- and GOA-like proteins of Brassicaceae to each other are shaded in blue and red, respectively.

	AtABS	CsaAGL32.3	CrTT16	NpAGL32	BsAGL32	LaAGL32.2	LcAGL32	BoTT16.1	SiAGL32	SpAGL32	AalAGL32	TaAGL32	AeaAGL32	CcAGL32	PtMADS30	VvBS1	NnBS1	AqucBS	VvBS2	PtMADS38	CcAGL63	CpAGL63	AeaAGL63	TaAGL63	AalAGL63	SpAGL63	SiAGL63	BoAGL63	LcAGL63	LaAGL63	BsAGL63	NpAGL63	CrAGL63	CsaAGL63	AtGOA
AtABS	100																																	
CsaAGL32.3	84	100																																
CrTT16	84	86	100																															
NpAGL32	83	90	86	100																														
BsAGL32	89	89	90	91	100																													
LaAGL32.2	81	80	82	80	84	100																												
LcAGL32	84	85	85	84	89	86	100																											
BoTT16.1	81	82	82	81	84	83	83	100																										
SiAGL32	83	82	84	84	87	86	86	90	100																									
SpAGL32	83	82	84	83	87	86	86	88	94	100																								
AalAGL32	79	77	79	78	81	81	81	82	84	84	100																							
TaAGL32	82	79	81	82	86	84	85	85	90	90	82	100																						
AeaAGL32	75	73	75	74	78	75	79	75	80	80	75	79	100																					
CcAGL32	59	58	59	60	60	60	61	62	62	62	60	62	61	100																				
PtMADS30	58	57	59	58	59	60	61	61	61	62	58	62	62	81	100																			
VvBS1	54	53	56	55	55	55	57	57	57	56	53	56	57	76	74	100																		
NnBS1	49	47	50	49	49	49	50	52	49	49	48	50	49	58	54	61	100																	
AqucBS	46	46	48	48	47	46	47	45	46	46	45	47	46	53	50	56	63	100																
VvBS2	49	47	49	48	48	49	49	52	51	50	48	49	51	55	54	60	53	49	100															
PtMADS38	50	49	51	51	51	49	52	51	51	51	49	51	48	56	53	57	53	52	54	100														
CcAGL63	48	47	48	48	49	48	50	49	48	47	48	48	48	55	52	56	50	49	56	59	100													
CpAGL63	43	43	43	43	44	44	46	46	45	44	42	46	45	53	51	55	49	48	55	55	57	100												
AeaAGL63	34	33	33	33	33	34	33	35	35	33	35	35	32	37	33	32	30	32	32	35	36	33	100											
TaAGL63	33	32	33	32	33	34	33	33	33	32	34	33	32	35	35	35	35	33	34	32	36	35	44	100										
AalAGL63	33	33	33	32	32	32	33	32	32	31	33	32	32	31	31	32	32	31	33	30	33	32	43	62	100									
SpAGL63	32	31	33	32	31	32	32	31	32	31	33	32	30	33	32	34	32	33	32	32	35	35	49	73	64	100								
SiAGL63	31	30	32	30	30	31	31	30	30	30	33	30	30	33	32	33	31	33	32	31	35	35	46	72	65	79	100							
BoAGL63	33	33	33	32	32	32	33	32	32	32	34	32	31	32	31	33	32	33	33	32	33	35	45	69	61	77	77	100						
LcAGL63	35	34	35	34	35	34	35	34	35	33	35	34	33	34	33	35	35	33	33	35	35	34	42	64	62	67	64	63	100					
LaAGL63	32	32	32	31	32	31	32	30	32	31	33	32	30	33	32	33	32	32	33	32	34	35	46	63	60	68	63	63	64	100				
BsAGL63	34	33	33	33	32	33	33	33	34	33	33	33	33	32	31	32	32	33	32	32	33	33	44	66	64	67	66	65	65	66	100			
NpAGL63	33	32	32	32	32	32	33	32	32	32	33	33	32	33	32	33	32	30	31	32	33	32	42	59	58	62	59	60	64	61	68	100		
CrAGL63	32	31	32	31	32	32	32	33	32	31	33	32	32	32	31	32	34	32	34	33	33	33	42	62	60	62	62	61	67	64	71	75	100	
CsaAGL63	33	32	32	32	32	32	32	34	33	32	32	33	33	34	33	33	34	32	34	34	36	33	43	62	60	65	62	62	67	64	69	81	78	100
AtGOA	31	31	31	31	30	31	31	31	31	31	32	31	31	29	30	29	32	31	29	28	30	29	40	58	61	61	61	62	67	62	71	65	66	67

Supplemental Table 3: Summary of the protein interaction analysis of ABS-like proteins from Brassicales as revealed by Yeast Two-Hybrid analysis. Growth at 30°C on selective SD –LWH medium is symbolized by +, no growth by -, weak growth by +/-. * indicates protein interaction at 22°C, # indicates interaction at 30°C, n.d., not determined.

pGADT7 \ pGBKT7	ABS *A. thaliana*	AalAGL32 *A. alpina*	BhAGL32 *B. holboellii*	CrTT16 *C. rubella*	LcAGL32 *L. campestre*	ThAGL32 *T. hassleriana*
Homodimer formation	+/-	+	+	-	+	n.d.
SEP1 *A. thaliana*	+	+	+	+	+	+
SEP2 *A. thaliana*	-	+* -#	+	+* -#	-	-
SEP3 *A. thaliana*	+	+	+	+	+	+
SEP4 *A. thaliana*	+	+	+	+	+	+
AP1 *A. thaliana*	+	+	+	-	+	+

Supplemental Table 4: Summary of the protein interaction analysis of GOA-like proteins as revealed by Yeast Two-Hybrid analyses. Growth on selective SD –LWH medium is symbolized by +, no growth by -.

pGADT7 \ pGBKT7	GOA *A. thaliana*	CrAGL63 *C. rubella*	LcAGL63 *L. campestre*
Homodimer formation	-	-	-
Corresponding ABS-like protein	-	-	-
AGL16 *A. thaliana*	+	-	-

Supplemental Table 5: Silique measurements of *goa-1* and wildtype plants of stage-17b fruits. The silique length, width and seed number per silique are given as means ± SDs. *n* is the number of measured fruits. Plants were grown under standard long-day conditions. n.d., not determined.

	Experiment 1 (Jena)		Experiment 2 (Jena)		Experiment 3 (Gießen)	
Genotype	Wildtype (n=54)	*goa-1* (n=52)	Wildtype (n=50)	*goa-1* (n=60)	Wildtype (n=66)	*goa-1* (n=70)
Silique width in mm	0.83 ± 0.11	0.82 ± 0.16	0.95 ± 0.12	0.96 ± 0.13	0.82 ± 0.09	0.82 ± 0.08
Silique length in mm	12.65 ± 1.47	13.01 ± 1.88	13.83 ± 2.11	13.82 ± 1.86	13.13 ± 2.65	14.13 ± 232
Seed number/fruit	41.02 ± 11.60	44.71 ± 12.71	48.86 ± 15.05	52.75 ± 11.76	n.d	n.d

Supplemental Methods

Plant growth

Seeds were obtained from the following people: *Boechera holboellii* seeds from Jochen Kumlehn, (Gatersleben, Germany), *Tarenaya hassleriana* seeds from Eric Schranz (Wageningen, Netherlands), *Arabis alpina* seeds from Maria Albani and George Coupland (Cologne, Germany), *Eutrema salsugineum* seeds from Arnould Savoure (Paris, France), *Lepidium campestre* and *Brassica napus* L. var *biennis* seeds from Klaus Mummenhoff (Osnabrück, Germany).

Arabidopsis thaliana Col-0 (NASC ID: N1092), *Capsella rubella* Monte Gargano (MTE, NASC: N9609), *Eutrema salsugineum* (accession Shandong), *Arabis alpina* (Dorfertal) and *Tarenaya hassleriana* (Purple Queen, ES1100) plants were grown in a peat based potting soil (ED73) with perlite (H. Nitsch & Sohn GmbH & Co.KG). *Brassica rapa* subspecies (seeds ordered from www.samenhaus.de), *Brassica napus* L. var *biennis* and *Arabidopsis lyrata* ssp. *lyrata* (NASC ID: N22696) plants were grown on a mixture of seedling substrate (Kammlott GmbH), sand and vermiculite (1–3 mm) (8:1:1). All plants were grown in a greenhouse under long day conditions (16 hours light and 8 hours dark). The temperature in the greenhouse was between 21 and 25°C at daytime and between 16 to 18°C at night with a relative humidity of 60%. Seeds of *C. rubella* and *E. salsugineum* were sown and stratified in moist soil at 4°C in the dark for 3 days (*C. rubella*) and 14 days (*E. salsugineum*), respectively, then allowed to germinate and grown in the greenhouse until the rosette was about 2.5 cm in diameter. Subsequently, the plants were vernalized at 4°C under short day conditions (8 hours light, 16 hours dark) for 6-8 weeks (*C. rubella*) or 4-5 weeks (*E. salsugineum*), respectively, and then returned to the greenhouse with long day conditions for further development.

*Aethionema arabicum (*KM 2416), *Sisymbrium irio*, and *Lepidium campestre* plants were grown in a mixture of loam, sand, pumice, and compost (1:2:2:5). *Ae. arabicum* seeds were sown and stratified in moist soil at 4°C in the dark for 4 days, then germinated in a climate chamber (14 hours light, 10 hours dark) at 15°C with a relative humidity of 50%. After 3 weeks the conditions were changed to 22°C (day) and 15°C (night). *Sisymbrium irio* plants were grown with 12 hours light and 12 hours dark at 22°C throughout and *L. campestre* plants were grown at with 16 hours light at 20°C and 8 hours dark at 15°C, respectively in the greenhouse. Flowering was induced by keeping the plants at 4°C for six weeks with 8 h of light per day.

In silico identification of B_{sister} genes

The sequence data of previously identified B_{sister} genes were taken from Phytozome (Goodstein et al., 2012) or NCBI (Sayers et al., 2011)(Supplemental Table 1). The genome of *Boechera stricta* (Lee et al., 2017) and the genome of *Eutrema salsugineum* (Yang et al., 2013) were downloaded from Phytozome. The genomes of *Arabidopsis lyrata* (Hu et al., 2011), *Schrenkiella parvula* (Dassanayake et al., 2011), *Aethionema arabicum*, *Leavenworthia alabamica*, *Sisymbrium irio*, a second genome of *Eutrema salsugineum* (Haudry et al., 2013), *Capsella grandiflora* and *Capsella orientalis* (Williamson et al., 2014) were downloaded from the UCSC Genome Browser (http://grandiflora.eeb.utoronto.ca:8086/downloads.html). The genomes of *Actinidia chinensis* (Huang et al., 2013), *Arabidopsis halleri* (DOE-JGI, http://phytozome.jgi.doe.gov/), *Betula nana* (Wang et al., 2013), *Camelina sativa* (Kagale et al., 2014), *Cicer arietum* (Varshney et al., 2013), *Cajanus cajan* (Varshney et al., 2012), *Eucalyptus camaldulensis*, *Lactuca sativa* (Reyes-Chin-Wo et al., 2017), *Morus notabilis* (He et al., 2013), *Nelumbo nucifera* (Ming et al., 2013), *Nicotiana sylvestris*, *Nicotiana tomentosiformis* (Sierro et al., 2013), *Phoenix dactylifera* (Al-Dous et al., 2011), *Prunus mume* (Zhang et al., 2012), *Raphanus sativus* (Kitashiba et al., 2014) and *Thlaspi arvense* (Dorn et al., 2015) were downloaded from the NCBI genomes database (Sayers et al., 2011). The genome of *Amborella trichopoda* (Albert et al., 2013) was downloaded from the *Amborella* genome database (http://www.amborella.org/). The genome of *Raphanus raphanistrum* (Moghe et al., 2014) was downloaded from RadishDB. The genome from *Lepidium meyenii* (Zhang et al., 2016) was downloaded from (http://www.herbal-genome.cn/html/Genomes/).

BLAST (Altschul et al., 1990) was used to search these genomes with the CDS of known *ABS*- and *GOA*-like genes from the most closely related species to the one studied as query sequences. The BLAST search for *Neslia paniculata* (Ågren et al., 2014) was conducted on https://genomevolution.org/coge/CoGeBlast.pl. The scaffolds on which the highest scoring results were found were saved and shortened such that the remaining part contained 10 kb up- and downstream of the BLAST result. FGENESH (Salamov and Solovyev, 2000) (default settings, "Dicot plants (*A. thaliana*)") was used to predict coding sequences (CDS). The genomes of *Aquilegia coerulea* (*Aquilegia coerulea* Genome Sequencing Project, http://phytozome.jgi.doe.gov/), *Citrus sinensis*, *Citrus clementina* (Wu et al., 2014), *Eucalyptus grandis* (Myburg et al., 2014), *Fragaria vesca* (Shulaev et al., 2011), *Malus domestica* (Velasco et al., 2010) and *Prunus persica* (Verde et al., 2013) were searched using BLAST as described above at Phytozome. If a CDS was provided in the genomic region where the highest scoring results were found by Phytozome, this CDS was used for further analysis, otherwise a CDS

prediction was done as described above using FGENESH. The gene identified by us as *CcAGL63* in *Citrus clementina* is likely the same gene described as *CcTT16* by Garcia-Lor et al. (2012), however since it is a *GOA*-like gene, we renamed it for better understanding.

For the *GOA* ortholog from *A. lyrata (AlAGL63)*, a BLAST search in the trace archives at NCBI (Sayers et al., 2011) was performed; the identified fragments were assembled using Sequencher version 5.2 sequence analysis software (Gene Codes Corporation) into two non-overlapping contigs separated by a fragment of unknown length. The two identified contigs and the fragment in between were verified by PCR and sequencing.

The genomic sequences of the B_{sister} genes from *A. alpina* (Willing et al., 2015) had been kindly provided before publication by George Coupland. Based on these sequences we predicted the CDS as described above and verified it subsequently by PCR and sequencing.

The CDS of the *ABS*-like gene from *Tarenaya hassleriana* was taken from (Bhide et al., 2014). Using this information, primers were designed for the amplification of cDNA and genomic DNA of the B_{sister} gene. The raw reads of the genome sequence of *Cleome violaceae* were downloaded from the Joint Genome Institute (Grigoriev et al., 2011). BLAST searches were conducted using first the CDS of the *ABS*-like gene from *T. hassleriana* as query and subsequently sequences of the assembled parts of the genomic locus of the *ABS*-like gene of *C. violaceae.* All resulting sequences were assembled after each BLAST search until the genomic locus was complete.

Perl Script for visualization of the alignment of exon-intron structures

```
#! /usr/bin/perl

use strict;
use warnings;

# call main
main();

# main
sub main {

	# define variables

	my %exons;
	my %length;
	my %seq;
	my @genes;
	my $name;
	my $count;
	my $start;
	my $end;
	my $max = 0;
	my $num = 0;
	my $alignmentfile;
	my $exons;
	my $output;

	# read command line parameters

	my $length = @ARGV;

	if ($length < 5){
```

```
160                     print "Call program like this: perl exon_intron_structure_align.pl -a
161     <alignment-file> -e <exonpositions-file> -o <output-file>\n";
162                     die();
163             }
164             else {
165                     foreach (my $i=0; $i<$length; $i++){
166                             if ($ARGV[$i] eq "-a"){
167                                     $alignmentfile = $ARGV[$i+1];
168                             }
169                             if ($ARGV[$i] eq "-e"){
170                                     $exons = $ARGV[$i+1];
171                             }
172                             if ($ARGV[$i] eq "-o"){
173                                     $output = $ARGV[$i+1];
174                             }
175                     }
176             }
177
178             # read positions of exons
179             open(EXONS, $exons);
180
181             while (<EXONS>){
182                     if ($_ =~ /^(\w+)\s+(\d+)\s/){
183                             $name = $1;
184                             $length{$name} = $2;
185                             push(@genes, $name);
186                             if ($length{$name} > $max){
187                                     $max = $length{$name};
188                             }
189                             $count = 0;
190                             while ($_ =~ /(\d+)-(\d+)/g){
191                                     $exons{$name}{$count}{start} = $1;
192                                     $exons{$name}{$count}{end} = $2;
193                                     $count++;
```

```
194                         }
195                         $num++;
196                 }
197         }
198
199         close(EXONS);
200
201         # read alignment
202
203         $name = "";
204         my $sequence = "";
205         my $prevpos;
206         my $currpos;
207
208         open(ALN, $alignmentfile);
209
210         while (<ALN>){
211                 chomp($_);
212                 if ($_ =~ />(\w+)/){
213                         if ($name ne ""){
214                                 $seq{$name} = $sequence;
215                         }
216                         $name = $1;
217                         $sequence = "";
218                 }
219                 else {
220                         $sequence .= $_;
221                 }
222         }
223
224         if ($name ne ""){
225                 $seq{$name} = $sequence;
226         }
227         my $alnlength = length($sequence);
```

```
228
229             close(ALN);
230
231             # draw exon-intron-structure and alignment
232
233             my $xsize = $max + 200;
234             my $ysize = 200*$num + 150;
235             my $yposition;
236             my $len;
237             $num = 0;
238             my $previous = "";
239             my $prevypos;
240             my $currypos;
241
242             open(PS, ">$output");
243
244             print PS "%!PS-Adobe EPSF-3.0\n";
245             print PS "%%BoundingBox: 0 0 $xsize $ysize\n";
246
247             foreach my $gene (@genes){
248                     $yposition = $ysize - 100 - 200*$num;
249                     $len = $length{$gene};
250                     # print gene structure
251                     print PS "newpath\n100 $yposition moveto\n$len 0 rlineto\nclosepath\n0 0 0
252     setrgbcolor\n2 setlinewidth\nstroke\n";
253                     print PS "/Times-Roman findfont\n20 scalefont\nsetfont\nnewpath\n0
254     $yposition moveto\n($gene) show\n";
255                     foreach my $key (sort keys %{$exons{$gene}}){
256                             $start = $exons{$gene}{$key}Re{start};
257                             $end = $exons{$gene}{$key}{end};
258                             $len = $end - $start;
259                             $start += 100;
```

```
260                         print PS "newpath\n$start $yposition moveto\n0 25 rlineto\n$len 0
261 rlineto\n0 -50 rlineto\n-$len 0 rlineto\nclosepath\ngsave\n0 0 0 setrgbcolor\nfill\ngrestore\n0
262 0 0 setrgbcolor\n1 setlinewidth\nstroke\n";
263                 }
264                 # print conservation
265                 if ($previous ne ""){
266                         $prevpos = 100 + 0;
267                         $currpos = 100 + 0;
268                         $prevypos = $yposition + 200 - 30;
269                         $currypos = $yposition + 30;
270                         for (my $pos = 0; $pos < $alnlength; $pos++){
271                                 if ((substr($seq{$previous}, $pos, 1) eq substr($seq{$gene},
272 $pos, 1)) && (substr($seq{$previous}, $pos, 1) ne "-")){
273                                         print PS "newpath\n$prevpos $prevypos
274 moveto\n$currpos $currypos lineto\nclosepath\n0.7 0.7 0.7 setrgbcolor\n0.8
275 setlinewidth\nstroke\n";
276                                 }
277                                 if (substr($seq{$previous}, $pos, 1) ne "-"){
278                                         $prevpos++;
279                                 }
280                                 if (substr($seq{$gene}, $pos, 1) ne "-"){
281                                         $currpos++;
282                                 }
283                         }
284                 }
285                 $previous = $gene;
286                 $num++;
287         }
288
289         close(PS);
290 }
291
```

References

Afkhami, M.E., and Stinchcombe, J.R. (2016). Multiple mutualist effects on genomewide expression in the tripartite association between *Medicago truncatula*, nitrogen-fixing bacteria and mycorrhizal fungi. Molecular Ecology **25,** 4946-4962.

Ågren, J.A., Wang, W., Koenig, D., Neuffer, B., Weigel, D., and Wright, S.I. (2014). Mating system shifts and transposable element evolution in the plant genus Capsella. BMC genomics **15,** 602.

Al-Dous, E.K., George, B., Al-Mahmoud, M.E., Al-Jaber, M.Y., Wang, H., Salameh, Y.M., Al-Azwani, E.K., Chaluvadi, S., Pontaroli, A.C., DeBarry, J., et al. (2011). De novo genome sequencing and comparative genomics of date palm (Phoenix dactylifera). Nat Biotechnol **29,** 521-U584.

Albert, V.A., Barbazuk, W.B., Der, J.P., Leebens-Mack, J., Ma, H., Palmer, J.D., Rounsley, S., Sankoff, D., Schuster, S.C., and Soltis, D.E. (2013). The Amborella Genome and the Evolution of Flowering Plants. Science **342,** 1241089.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J Mol Biol **215,** 403-410.

Astudillo-Reyes, C., Fernandez, A.C., and Cichy, K.A. (2015). Transcriptome characterization of developing bean (*Phaseolus vulgaris* L.) pods from two genotypes with contrasting seed zinc concentrations. PLoS one **10,** e0137157.

Bhide, A., Schliesky, S., Reich, M., Weber, A.P., and Becker, A. (2014). Analysis of the floral transcriptome of *Tarenaya hassleriana* (Cleomaceae), a member of the sister group to the Brassicaceae: towards understanding the base of morphological diversity in Brassicales. BMC genomics **15,** 140.

Bond, D.M., Albert, N.W., Lee, R.H., Gillard, G.B., Brown, C.M., Hellens, R.P., and Macknight, R.C. (2016). Infiltration-RNAseq: transcriptome profiling of Agrobacterium-mediated infiltration of transcription factors to discover gene function and expression networks in plants. Plant Methods **12,** 41.

Dassanayake, M., Oh, D.H., Haas, J.S., Hernandez, A., Hong, H., Ali, S., Yun, D.J., Bressan, R.A., Zhu, J.K., Bohnert, H.J., et al. (2011). The genome of the extremophile crucifer *Thellungiella parvula*. Nat Genet **43,** 913-918.

Dias, B.F.d.O., Simões-Araújo, J.L., Russo, C.A., Margis, R., and Alves-Ferreira, M. (2005). Unravelling MADS-box gene family in *Eucalyptus* spp.: a starting point to an understanding of their developmental role in trees. Genetics and Molecular Biology **28,** 501-510.

Dorn, K.M., Fankhauser, J.D., Wyse, D.L., and Marks, M.D. (2015). A draft genome of field pennycress (*Thlaspi arvense*) provides tools for the domestication of a new winter biofuel crop. DNA Research **22,** 121-131.

Garcia-Lor, A., Garcia-Martinez, J.L., and Perez-Amador, M.A. (2012). Identification of ovule and seed genes from *Citrus clementina*. Tree Genetics & Genomes **8,** 227-235.

Gasteiger, E., Hoogland, C., Gattiker, A., Wilkins, M.R., Appel, R.D., and Bairoch, A. (2005). Protein identification and analysis tools on the ExPASy server. In The proteomics protocols handbook (Springer), pp. 571-607.

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., et al. (2012). Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res **40,** D1178-1186.

Grigoriev, I.V., Nordberg, H., Shabalov, I., Aerts, A., Cantor, M., Goodstein, D., Kuo, A., Minovitsky, S., Nikitin, R., and Ohm, R.A. (2011). The genome portal of the department of energy joint genome institute. Nucleic acids research**,** gkr947.

Grimplet, J., Martínez-Zapater, J.M., and Carmona, M.J. (2016). Structural and functional annotation of the MADS-box transcription factor family in grapevine. BMC genomics **17,** 1.

Haudry, A., Platts, A.E., Vello, E., Hoen, D.R., Leclercq, M., Williamson, R.J., Forczek, E., Joly-Lopez, Z., Steffen, J.G., Hazzouri, K.M., et al. (2013). An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nat Genet **45,** 891-898.

He, N.J., Zhang, C., Qi, X.W., Zhao, S.C., Tao, Y., Yang, G.J., Lee, T.H., Wang, X.Y., Cai, Q.L., Li, D., et al. (2013). Draft genome sequence of the mulberry tree Morus notabilis. Nature Communications **4**.

Hou, X.-J., Liu, S.-R., Khan, M.R.G., Hu, C.-G., and Zhang, J.-Z. (2014). Genome-wide identification, classification, expression profiling, and SSR marker development of the MADS-box gene family in Citrus. Plant molecular biology reporter **32,** 28-41.

Hu, T.T., Pattyn, P., Bakker, E.G., Cao, J., Cheng, J.F., Clark, R.M., Fahlgren, N., Fawcett, J.A., Grimwood, J., Gundlach, H., et al. (2011). The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat Genet **43,** 476-481.

Huang, K., Louis, J.M., Donaldson, L., Lim, F.L., Sharrocks, A.D., and Clore, G.M. (2000). Solution structure of the MEF2A-DNA complex: structural basis for the modulation of DNA bending and specificity by MADS-box transcription factors. Embo J **19,** 2615-2628.

Huang, S., Ding, J., Deng, D., Tang, W., Sun, H., Liu, D., Zhang, L., Niu, X., Zhang, X., and Meng, M. (2013). Draft genome of the kiwifruit *Actinidia chinensis*. Nature communications **4,** 2640.

Kagale, S., Koh, C., Nixon, J., Bollina, V., Clarke, W.E., Tuteja, R., Spillane, C., Robinson, S.J., Links, M.G., and Clarke, C. (2014). The emerging biofuel crop *Camelina sativa* retains a highly undifferentiated hexaploid genome structure. Nat. Commun. **5**.

Kang, C., Darwish, O., Geretz, A., Shahan, R., Alkharouf, N., and Liu, Z. (2013). Genome-scale transcriptomic insights into early-stage fruit development in woodland strawberry *Fragaria vesca*. The Plant Cell **25,** 1960-1978.

Kitashiba, H., Li, F., Hirakawa, H., Kawanabe, T., Zou, Z., Hasegawa, Y., Tonosaki, K., Shirasawa, S., Fukushima, A., and Yokoi, S. (2014). Draft sequences of the radish (*Raphanus sativus L.*) genome. DNA Research **21,** 481-490.

Külahoglu, C., Denton, A.K., Sommer, M., Maß, J., Schliesky, S., Wrobel, T.J., Berckmans, B., Gongora-Castillo, E., Buell, C.R., and Simon, R. (2014). Comparative transcriptome atlases reveal altered gene expression modules between two Cleomaceae C3 and C4 plant species. The Plant Cell **26,** 3243-3260.

Lee, C.-R., Wang, B., Mojica, J.P., Mandáková, T., Prasad, K.V.S.K., Goicoechea, J.L., Perera, N., Hellsten, U., Hundley, H.N., Johnson, J., et al. (2017). Young inversion with multiple linked QTLs under selection in a hybrid zone. Nature Ecology & Evolution **1,** 0119.

Li, T., Qi, X., Zeng, Q., Xiang, Z., and He, N. (2014). MorusDB: a resource for mulberry genomics and genome biology. Database **2014,** bau054.

Magallón, S., Gómez-Acevedo, S., Sánchez-Reyes, L.L., and Hernández-Hernández, T. (2015). A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. New Phytologist **207,** 437-453.

Ming, R., VanBuren, R., Liu, Y.L., Yang, M., Han, Y.P., Li, L.T., Zhang, Q., Kim, M.J., Schatz, M.C., Campbell, M., et al. (2013). Genome of the long-living sacred lotus (Nelumbo nucifera Gaertn.). Genome Biology **14**.

Moghe, G.D., Hufnagel, D.E., Tang, H., Xiao, Y., Dworkin, I., Town, C.D., Conner, J.K., and Shiu, S.-H. (2014). Consequences of whole-genome triplication as revealed by comparative genomic analyses of the wild radish *Raphanus raphanistrum* and three other Brassicaceae species. Plant Cell **6,** 1925–1937.

Motamayor, J.C., Mockaitis, K., Schmutz, J., Haiminen, N., Livingstone III, D., Cornejo, O., Findley, S.D., Zheng, P., Utro, F., and Royaert, S. (2013). The genome sequence

of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. Genome biology **14,** r53.

Myburg, A.A., Grattapaglia, D., Tuskan, G.A., Hellsten, U., Hayes, R.D., Grimwood, J., Jenkins, J., Lindquist, E., Tice, H., Bauer, D., et al. (2014). The genome of Eucalyptus grandis. Nature **510,** 356-+.

Pellegrini, L., Song, T., and Richmond, T.J. (1995). Structure of serum response factor core bound to DNA. Nature **376,** 490-498.

Pradhan, S., Bandhiwal, N., Shah, N., Kant, C., Gaur, R., and Bhatia, S. (2014). Global transcriptome analysis of developing chickpea (*Cicer arietinum* L.) seeds. Frontiers in Plant Science **5,** 698.

Prasad, K., Zhang, X., Tobon, E., and Ambrose, B.A. (2010). The Arabidopsis Bsister MADS-box protein, GORDITA, represses fruit growth and contributes to integument development. Plant J **62,** 203-214.

Reyes-Chin-Wo, S., Wang, Z., Yang, X., Kozik, A., Arikit, S., Song, C., Xia, L., Froenicke, L., Lavelle, D.O., Truco, M.-J., et al. (2017). Genome assembly with in vitro proximity ligation data and whole-genome triplication in lettuce. Nature Communications **8,** 14953.

Salamov, A.A., and Solovyev, V.V. (2000). Ab initio gene finding in Drosophila genomic DNA. Genome Res **10,** 516-522.

Sayers, E.W., Barrett, T., Benson, D.A., Bolton, E., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., and Federhen, S. (2011). Database resources of the national center for biotechnology information. Nucleic acids research **39,** D38-D51.

Shulaev, V., Sargent, D.J., Crowhurst, R.N., Mockler, T.C., Folkerts, O., Delcher, A.L., Jaiswal, P., Mockaitis, K., Liston, A., Mane, S.P., et al. (2011). The genome of woodland strawberry (Fragaria vesca). Nature Genetics **43,** 109-116.

Sierro, N., Battey, J.N., Ouadi, S., Bovet, L., Goepfert, S., Bakaher, N., Peitsch, M.C., and Ivanov, N.V. (2013). Reference genomes and transcriptomes of *Nicotiana sylvestris* and *Nicotiana tomentosiformis*. Genome biology **14,** R60.

Singh, V.K., Garg, R., and Jain, M. (2013). A global view of transcriptome dynamics during flower development in chickpea by deep sequencing. Plant biotechnology journal **11,** 691-701.

Song, Q.-X., Li, Q.-T., Liu, Y.-F., Zhang, F.-X., Ma, B., Zhang, W.-K., Man, W.-Q., Du, W.-G., Wang, G.-D., and Chen, S.-Y. (2013). Soybean *GmbZIP123* gene enhances

lipid content in the seeds of transgenic *Arabidopsis plants*. Journal of experimental botany **64,** 4329-4341.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res **25,** 4876-4882.

Urasaki, N., Tarora, K., Shudo, A., Ueno, H., Tamaki, M., Miyagi, N., Adaniya, S., and Matsumura, H. (2012). Digital transcriptome analysis of putative sex-determination genes in papaya (*Carica papaya*). PLoS One **7,** e40904.

Varshney, R.K., Chen, W.B., Li, Y.P., Bharti, A.K., Saxena, R.K., Schlueter, J.A., Donoghue, M.T.A., Azam, S., Fan, G.Y., Whaley, A.M., et al. (2012). Draft genome sequence of pigeonpea (Cajanus cajan), an orphan legume crop of resource-poor farmers. Nat Biotechnol **30,** 83-U128.

Varshney, R.K., Song, C., Saxena, R.K., Azam, S., Yu, S., Sharpe, A.G., Cannon, S., Baek, J., Rosen, B.D., Tar'an, B., et al. (2013). Draft genome sequence of chickpea (Cicer arietinum) provides a resource for trait improvement. Nat Biotechnol **31,** 240-246.

Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., Fontana, P., Bhatnagar, S.K., Troggio, M., Pruss, D., et al. (2010). The genome of the domesticated apple (Malus x domestica Borkh.). Nature Genetics **42,** 833-+.

Verde, I., Abbott, A.G., Scalabrin, S., Jung, S., Shu, S.Q., Marroni, F., Zhebentyayeva, T., Dettori, M.T., Grimwood, J., Cattonaro, F., et al. (2013). The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution. Nature Genetics **45,** 487-U447.

Wang, N., Thomson, M., Bodles, W.J.A., Crawford, R.M.M., Hunt, H.V., Featherstone, A.W., Pellicer, J., and Buggs, R.J.A. (2013). Genome sequence of dwarf birch (Betula nana) and cross-species RAD markers. Molecular Ecology **22,** 3098-3111.

Williamson, R.J., Josephs, E.B., Platts, A.E., Hazzouri, K.M., Haudry, A., Blanchette, M., and Wright, S.I. (2014). Evidence for widespread positive and negative selection in coding and conserved noncoding regions of *Capsella grandiflora*. Plos Genet **10,** e1004622.

Willing, E.-M., Rawat, V., Mandáková, T., Maumus, F., James, G.V., Nordström, K.J., Becker, C., Warthmann, N., Chica, C., and Szarzynska, B. (2015). Genome expansion of *Arabis alpina* linked with retrotransposition and reduced symmetric DNA methylation. Nature Plants **1**.

Wu, G.A., Prochnik, S., Jenkins, J., Salse, J., Hellsten, U., Murat, F., Perrier, X., Ruiz, M., Scalabrin, S., Terol, J., et al. (2014). Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. Nat Biotechnol **32,** 656-+.

Yang, R.L., Jarvis, D.E., Chen, H., Beilstein, M.A., Grimwood, J., Jenkins, J., Shu, S.Q., Prochnik, S., Xin, M.M., Ma, C., et al. (2013). The reference genome of the halophytic plant Eutrema salsugineum. Frontiers in Plant Science **4**.

Zhang, J., Tian, Y., Yan, L., Zhang, G., Wang, X., Zeng, Y., Zhang, J., Ma, X., Tan, Y., and Long, N. (2016). Genome of plant maca (*Lepidium meyenii*) illuminates genomic basis for high altitude adaptation in the central Andes. Molecular plant **9,** 1066–1077.

Zhang, Q.X., Chen, W.B., Sun, L.D., Zhao, F.Y., Huang, B.Q., Yang, W.R., Tao, Y., Wang, J., Yuan, Z.Q., Fan, G.Y., et al. (2012). The genome of Prunus mume. Nature Communications **3**.

Zhao, J., Li, Y., Ding, L., Yan, S., Liu, M., Jiang, L., Zhao, W., Wang, Q., Yan, L., and Liu, R. (2016). Phloem transcriptome signatures underpin the physiological differentiation of the pedicel, stalk and fruit of cucumber (*Cucumis sativus* L.). Plant and Cell Physiology **57,** 19-34.