JSE Journal of Systematics and Evolution

doi: 10.1111/jse.12204

Research Article

Using nuclear genes to reconstruct angiosperm phylogeny at the species level: A case study with Brassicaceae species

Liming Cai[1,3] and Hong Ma[1,2]*

[1]Ministry of Education Key Laboratory of Biodiversity Sciences and Ecological Engineering, Institute of Plant Biology, Institute of Biodiversity Sciences, Center for Evolutionary Biology, School of Life Sciences, Fudan University, Shanghai 200438, China

[2]Institutes of Biomedical Sciences, Fudan University, Shanghai 200032, China

[3]Present address: Harvard University Herbaria, Cambridge, MA 02138, USA

*Author for correspondence. E-mail: hongma@fudan.edu.cn. Tel.: 86-21-51630500. Fax: 86-21-51630504.

Received 14 January 2016; Accepted 15 March 2016; Article first published online xx Month 2016

Abstract Angiosperm phylogeny has been investigated extensively using organellar sequences; recent efforts using nuclear genes have also been successful in reconstructing angiosperm phylogenies at family or deeper levels. However, it is not clear whether nuclear genes are also effective in understanding relationships between species in a genus. Here we present a case study of phylogeny at generic and specific levels with nuclear genes, using Brassicaceae taxa as examples. Brassicaceae includes various crops and the model plant *Arabidopsis thaliana*. A recent study showed that nuclear genes can provide well-resolved relationships between tribes and larger lineages in Brassicaceae, but few species were included in any given genus. We present a phylogeny with multiple species in each of five genera within Brassicaceae for a total of 65 taxa, using three protein-coding nuclear genes, *MLH1*, *SMC2*, and *MCM5*, with up to approximately 10 200 base pairs (in both exons and introns). Maximum likelihood and Bayesian analyses of the separate gene regions and combined data reveal high resolution at various phylogenetic depths. The relationships between genera here were largely congruent with previous results, with further resolution at the species level. Also, we report for the first time the affinity of *Cardamine rockii* with tribe Camelineae instead of other *Cardamine* members. In addition, we report sequence divergence at three levels: across angiosperms, among Brassicaceae species, and between *Arabidopsis* ecotypes. Our results provide a robust species-level phylogeny for a number of Brassicaceae members and support an optimistic perspective on the phylogenetic utility of conserved nuclear data for relatively recent clades.

Key words: *Brassica*, Brassicaceae, *Cardamine*, *Lepidium*, *Rorippa*, nuclear gene, species-level phylogeny.

The cabbage family (Brassicaceae) is one of the most important and well-known plant groups, with various vegetables (such as cabbage and broccoli) and oilseed crop species cultivated throughout the world. In addition, a well-known member of the family is the model organism *Arabidopsis thaliana* (L.) Heynh., the studies of which have revolutionized our knowledge in almost every field of modern plant biology. Several studies have been carried out on molecular phylogeny across this family in recent years (Bailey et al., 2006; Beilstein et al., 2006, 2008; Warwick et al., 2010; Al-Shehbaz, 2012). In these studies, phylogenetic inference provided the basics for understanding the patterns of evolutionary history. Transcribed nuclear ribosomal spacer (ITS), plastid DNA (cpDNA), and ribosomal DNA (rDNA) markers are most commonly used because of the highly conserved sequences and high copy numbers, making them easily accessible without cloning (Baldwin et al., 1995; Baldwin & Markos, 1998; Álvarez & Wendel, 2003). However, the entire plastid genome is a single linkage group (Birky, 1995), whereas rDNAs are sometimes not completely identical within a genome (Buckler et al., 1997). As a result, these markers are not capable of dealing with the cases involving hybridization and polyploidization. More importantly, building phylogenies below genus level requires rapidly evolving markers that vary among closely related species. Organellar and ribosomal genes are sometimes too conserved in this context (Small et al., 1998; Sang, 2002). The limitation of traditional cpDNA and rDNA markers and the growing availability of large genomic datasets from multiple taxa have prompted us to mine the plant genomes for appropriate low-copy nuclear protein-coding genes as phylogenetic markers.

Compared to organellar genes, protein-coding nuclear genes are biparentally inherited, recording genetic information from both male and female inherence. Additionally, their third positions of codons, introns, and untranslated 3′/5′-untranslated regions have relatively high evolutionary rates, presenting better resolution among closely related species (Zimmer & Wen, 2013). Therefore, we hypothesized that better resolved relationships, especially at the species level, would be uncovered using suitable nuclear genes. At the same time, events of hybridization and polyploidization might also be uncovered. Despite these advantages, building nuclear phylogeny is still challenging in several aspects. The highly complex nuclear genome, particularly with the high frequency of polyploidy events in angiosperms (Soltis et al., 2009), makes it a serious problem to identify low-copy orthologs. For example, the ancestors of extent angiosperms (Jiao et al., 2011), of core eudicots (Blanc & Wolfe, 2004; De Bodt et al., 2005; Soltis et al., 2009), and of Brassicaceae (β and α) (Blanc & Wolfe, 2004; Schranz & Mitchell-Olds, 2006) all underwent one or more rounds of whole genome duplication. Repeated genome duplication and gene losses have made it difficult to distinguish orthologs from paralogs. In some cases, the loss of different duplicates in separate lineages can result in "hidden paralogs", yielding conflicting gene trees that exacerbate the uncertainty in phylogenetic reconstruction (Maddison, 1997). Thus it is crucial to identify genes that tend to return to single copy status quickly after genome duplication, thus behaving as ortholog among a wide range of plants.

In recent years, nuclear genes have been used to resolve relatively deep relationships among major land plant groups (Wickett et al., 2014), across angiosperms (Zhang et al., 2012; Zeng et al., 2014), and at the levels of order and family (Ding et al., 2012; Yang et al., 2015; Huang et al., 2016). Several of these studies used phylogenomics and/or phylotranscriptomics, with many gene sequences from large-scale datasets (Wickett et al., 2014; Zeng et al., 2014; Yang et al., 2015; Huang et al., 2016), illustrating the great power of the vast quantity of nuclear genes. However, for many questions concerning closely related species, it might not be necessary to use many hundreds of genes, and the need to sample a relatively large number of taxa might make large-scale datasets too costly to obtain. Also, the use of a large number of genes for many taxa will increase the need for great computation power. Thus, it is desirable to be able to quickly obtain a small number of genes from a large number of taxa. However, isolation of nuclear marker sequences is often time consuming and laborious (Small et al., 2004). Furthermore, the need for sequence variation to provide phylogenetic signals makes it difficult to develop universal primers. Therefore, identification of appropriate nuclear genes that are easy to amplify and sequence is a major objective, before one can address phylogenetic questions using this approach.

The identification of suitable nuclear genes initially involves gene comparisons across genomes and transcriptomes of a wide range of taxa, followed by lineage-specific analysis to test the usability of primers for amplification. Previous efforts on nuclear gene phylogenies within plant groups were largely focused on single copy genes, for example, *LFY* (Oh & Potter, 2005; Nie et al., 2008; Kim et al., 2010), or easily distinguished paralogs, such as *Adh1* and *Adh2* (Gaut & Clegg, 1991; Fukuda et al., 2005). A recent study showed that five nuclear genes were sufficient to resolve many deep relationships in the angiosperm phylogeny (Zhang et al., 2012). These nuclear genes were selected from four different gene families, the *MCM, SMC, MLH*, and *MSH* gene families. They were verified for their extensive single-copy and orthology across most angiosperm groups (Gozuacik et al., 2003; Lin et al., 2007; Surcel et al., 2008). However, whether these nuclear markers are appropriate for phylogenetic study at or below the genus level remains unknown.

As mentioned above, Brassicaceae provide an excellent model group for phylogenetic study because of its moderately large size of approximately 3700 species, relatively rapid evolutionary rate, and easy access for many of the members (Al-Shehbaz, 2012; Huang et al., 2016). Furthermore, the wealth of nuclear sequence information from multiple Brassicaceae genomic datasets (*Arabidopsis thaliana*, *Arabidopsis lyrata* (L.) O'Kane & Al-Shehbaz, *Capsella rubella* (Almq.) Almq., *Capsella grandiflora* Bioss., *Boechera stricta* (Graham) Al-Shehbaz, *Brassica rapa* L., and *Eutrema salsuginea* O. E. Schulz) allows preliminary screening of candidate nuclear genes (https://phytozome.jgi.doe.gov/pz/portal.html). Previous systematic research placed this family sister to Cleomaceae and divided it into 51 tribes; in addition, phylogenetic analyses support a clade (core Brassicaceae) with three main lineages (Lineages I, II, and III) and *Aethionema* as the sister to core Brassicaceae (Al-Shehbaz, 2012; Huang et al., 2016). Many of the core Brassicaceae species are grouped into Lineages I, II, and III with moderate to low support. A preliminary study (Ding et al., 2012) used three low-copy nuclear genes, including two from Zhang et al. (2012), to reconstruct phylogenetic relationships among 13 Brassicaceae species. The results were encouraging and indicated that nuclear genes performed better than plastid genes and ITS segments in untangling species-level phylogeny. Moreover, recent work with 55 large-scale datasets has yielded a highly resolved phylogeny of major lineages (Huang et al., 2016), but few of the genera included had more than one species.

In this study, to test the effectiveness of nuclear genes in resolving species-level phylogeny, a total of 35 taxa were selected from three genera (*Brassica* L., *Lepidium* L., and *Cardamine* L.). These three genera are widely distributed and have typical morphological features as to be distinguished from other cruciferous plants. Seventeen available Brassicaceae genomes and 11 transcriptomes were also included to expand our sampling among both closely related species and more distant ones. To test genes that are conserved among all plants (even other eukaryotes) to facilitate future use in other families, we sampled three representatives of eukaryote-wide gene families (Gozuacik et al., 2003; Lin et al., 2007; Surcel et al., 2008): *MLH1*, *SMC2*, and *MCM5*, and investigated their utility in resolving relationships among low-rank taxonomic hierarchies. We also examined the sequence similarities of these three genes across three levels of evolutionary distances: among divergent angiosperm species, between members of Brassicaceae, and within the same species—*Arabidopsis thaliana*. Our results illustrate the effectiveness of conserved nuclear genes in resolving species phylogeny within a genus and provide useful information for future investigation of relatively close relationships in many other groups.

Material and Methods

Taxon sampling

A total of 63 accessions from across Brassicaceae were included, together with two out-group species *Cleome serrulata* Pursh and *Populus trichocarpa* Torr. & A. Gray ex

Hook. The taxon sampling included 35 accessions of extracted DNAs (Table 1) and 28 accessions from genome and transcriptome datasets (Table 2). To test for the effectiveness of nuclear genes in resolving species within a genus, sampled taxa were largely members of three genera, including 8 from *Lepidium*, 8 from *Cardamine*, and 19 from *Brassica*. In addition, 6 species from *Rorippa* Scop., 1 species from *Nasturtium* W. T. Aiton, 1 species from *Leavenworthia* Torr., and 4 accessions from *Raphanus* L. were included as to expand our sampling variety within tribe Cardamineae and Brassiceae. *Brassica* is the type genus of Brassiceae with tremendous diversity, in part owing to domestication. *Cardamine* and *Lepidium* are also type genera of tribes Cardamineae and Lepidiumeae, respectively. All three genera are widely distributed in China and easily accessible. Sampled materials included both wild species and domestic vegetables. Species belonging to

Table 1 Scientific names and collection information of DNA materials of Brassicaceae species used to reconstruct angiosperm phylogeny with nuclear genes *MLH1*, *MCM5*, and *SMC2*

Taxon	*MLH1*	*MCM5*	*SMC2*	Voucher	Collection locality	Collection year
Brassica campestris L. var. *chinensis*	+	+	+	L. Cai 020102	Shanghai, China	2014
Brassica campestris L. var. *narinosa*	+	−	+	L. Cai 020701	Shanghai, China	2014
Brassica juncea (L.) Czern. var. *multisecta*	−	+	−	H. Ma 020302	Pennsylvania, USA	2014
Brassica oleracea L. var. *botrytis* L.	+	+	+	L. Cai 020501	Shanghai, China	2014
Brassica oleracea L. var. *caulorapa* Metzg.	+	+	+	L. Cai 020502	Shanghai, China	2014
Brassica oleracea L. var. *gemmifera* DC.	−	+	+	H. Ma 020505	Pennsylvania, USA	2014
Brassica oleracea L. var. *italic* Plenck	+	+	−	L. Cai 020504	Shanghai, China	2014
Brassica oleracea L. var. *gongylodes* L.	+	+	+	H. Ma 020508	Pennsylvania, USA	2014
Brassica oleracea L. var. *capitata* L.	−	+	+	H. Ma 020507	Pennsylvania, USA	2014
Brassica oleracea L. var. *acephala* (DC.) Metzg.	−	+	+	H. Ma 020506	Pennsylvania, USA	2014
Brassica oleraceae L. var. *alboglabra* Beiley	+	+	−	L. Cai 020601	Shanghai, China	2014
Brassica rapa L. var. *pekinensis* (Lour.) Kitam.	+	+	+	L. Cai 020402	Shanghai, China	2014
Brassica rapa L. var. *rapa*	+	−	+	H. Ma 020403	Pennsylvania, USA	2014
Brassica rapa L. var. *ruvo*	+	+	+	H. Ma 020405	Pennsylvania, USA	2014
Brassica napus L. var. *napobrassica* (L.) Hanelt	+	+	−	H. Ma 020701	Pennsylvania, USA	2014
Cardamine flexuosa With.	+	+	+	L. Cai 440208	Shanghai, China	2013
Cardamine hirsute L.	+	+	+	L. Cai 440502	Shanghai, China	2013
Cardamine lyrata Bunge	+	−	+	N. Zhang 440101	Shanghai, China	2011
Cardamine macrophylla Willd.	+	+	+	ZY 440401	China	2012
Cardamine oligosperma Nutt.	+	+	+	L. Cai 440503	California, USA	2013
Cardamine rockii O. E. Schulz	+	−	+	CGBOWS 440601	Yunnan, China	2014
Cardamine tangutorum O. E. Schulz	+	+	+	ZWJ 440301	Gansu, China	2012
Lepidium apetalum Willd.	−	−	+	H. Ma 100402	Shandong, China	2012
Lepidium cuneiforme C. Y. Wu	+	+	+	CGBOWS 101001	Sichuan, China	2014
Lepidium ferganense Korsh.	−	−	+	CGBOWS 100901	Xinjiang, China	2014
Lepidium latifolium L.	+	+	+	L. Cai 100201	Xinjiang, China	2012
Lepidium perfoliatum L.	+	+	−	L. Cai 100601	Xinjiang, China	2012
Lepidium ruderale L.	−	−	+	L. Cai 100301	Shandong, China	2012
Nasturtium officinale W. T. Aiton	+	+	+	L. Cai 570102	Shanghai, China	2014
Raphanus sativus L. var. (1)	+	+	+	H. Ma 060105	Pennsylvania, USA	2014
Raphanus sativus L. var. *longipinnatus* L. H. Bailey	+	+	+	L. Cai 060104	Shanghai, China	2014
Raphanus sativus L. var. (2)	+	+	−	L. Cai 060106	Shanghai, China	2014
Rorippa cantoniensis (Lour.) Ohwi	+	+	+	L. Cai 560201	Shanghai, China	2013
Rorippa dubia (Pers.) Hara	−	+	+	L. Cai 560501	Shanghai, China	2013
Rorippa islandica (Oeder) Borbás	+	−	−	L. Cai 560301	Xingjiang, China	2012

+, DNA sequences obtained for certain genes in that taxa; −, DNA sequences not obtained for certain genes in that taxa.

Table 2 Source information for Brassicaceae genomes/transcriptomes used in this study

Species	Data type	Source
Aethionema subulatum Bioss.	Transcriptome	Huang et al., 2016
Barbarea vulgaris W. T. Aiton	Transcriptome	Huang et al., 2016
Brassica nigra (L.) W. D. J. Koch	Transcriptome	Huang et al., 2016
Cardamine pensylvanica Muhl. ex Willd.	Transcriptome	Huang et al., 2016
Erysimum cheiranthoides L.	Transcriptome	Huang et al., 2016
Erysimum cheiri Crantz	Transcriptome	Huang et al., 2016
Lepidium campestre (L.) W. T. Aiton	Transcriptome	Huang et al., 2016
Lepidium didymum L.	Transcriptome	Huang et al., 2016
Rorippa indica (L.) Hiern	Transcriptome	Huang et al., 2016
Rorippa sylvestris (L.) Besser	Transcriptome	Huang et al., 2016
Rorippa globosa (Turcz.) Vassilcz.	Transcriptome	Huang et al., 2016
Aethionema arabicum (L.) Andrz. ex DC.	Genome	NCBI
Arabidopsis halleri (L.) O'Kane & Al-Shehbaz subsp. *gemmifera* (Matsum.) O'Kane & Al-Shehbaz	Genome	NCBI
Arabidopsis lyrata (L.) O'Kane & Al-Shehbaz	Genome	Phytozome v10.0
Arabidopsis thaliana (L.) Heynh.	Genome	Phytozome v10.0
Arabis alpina L. cultivar Pajares	Genome	NCBI
Boechera stricta (Graham) Al-Shehbaz	Genome	NCBI
Brassica napus L. cultivar ZS11	Genome	NCBI
Brassica oleracea L. var. *oleracea* cultivar TO1000	Genome	NCBI
Brassica rapa L. cultivar FPsc	Genome	Phytozome v10.0
Camelina sativa (L.) Crantz	Genome	NCBI
Capsella grandiflora Bioss.	Genome	NCBI
Capsella rubella (Almq.) Almq.	Genome	Phytozome v10.0
Eutrema salsuginea O. E. Schulz	Genome	Phytozome v10.0
Leavenworthia alabamica Rollins	Genome	NCBI
Raphanus raphanistrum L. subsp. *raphanistrum*	Genome	NCBI
Schrenkiella parvula (Schrenk) D. A. German & Al-Shehbaz	Genome	thellungiella.org
Sisymbrium irio L.	Genome	NCBI

NCBI, National Center for Biotechnology Information; v, version.

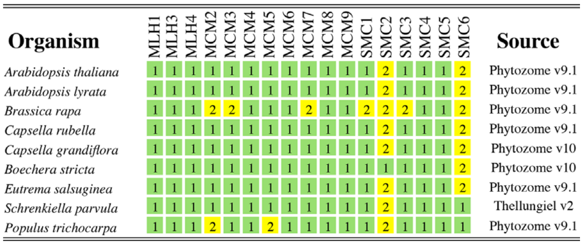
Cardamine, *Lepidium*, and *Rorippa* are mostly wild plants that are genetically more distant from other species. Cultivars of the same *Brassica* species were also sampled to represent lower-level divergence from relatively recent domestication histories. As a result, such a sampling strategy provided an effective test for the utility of the DNA markers to recover phylogenies at distinct levels. Source information for these accessions was listed in Table 1. Total genomic DNAs were extracted from leaves using the CTAB method (Stewart & Via, 1993).

Apart from our DNA sampling, we also took advantage of genome and transcriptome datasets to facilitate candidate gene screening. Transcriptome datasets used here were selected from those reported by Huang et al. (2016), including two species from *Lepidium*, one from *Cardamine* and three from *Rorippa*. Within Brassicaceae Lineage I, the clades Camelineae and Erysimeae were represented by six and two species, respectively. Likewise, within Lineage II, tribe Brassiceae was represented by eight species, including five in *Brassica*. Other tribes were represented by species with sequenced genomes: *Sisymbrium irio* L., *Eutrema salsuginea*, *Schrenkiella parvula* (Schrenk) D. A. German & Al-Shehbaz, *Thlaspi arvense* L., and *Arabis alpina* L. Two species from the basal lineage, *Aethionema subulatum* Bioss. and *Aethionema arabicum* (L.) Andrz. ex DC., were specially selected to represent the deepest genetic divergence within Brassicaceae.

Preliminary screening of candidate genes

Previously, five low-copy nuclear genes (*SMC1*, *SMC2*, *MLH1*, *MSH1*, and *MCM5*) were used to reconstruct a highly supported angiosperm phylogeny (Zhang et al., 2012). These five genes remain orthologous across a wide range of angiosperms and have conserved exon sequences, which can facilitate primer design and global alignment. Yet resolving species-level phylogeny requires rapidly evolving markers with sufficient variable sites among closely related species. Thus introns, with more variable sequence, are expected to play a key role in untangling recent and rapid radiation of species rendering their high evolutionary rate. These five genes all contain both exons and introns and encode proteins with at least 300 amino acids. They are all housekeeping genes with conserved functions. We then examined copy numbers of these five and related genes using eight Brassicaceae species with sequenced genomes from Phytozome version 10 (https://phytozome.jgi.doe.gov/pz/portal.html) (Fig. 1).

We compared genomic sequences of the five genes and calculated the following average nucleotide sequence identities: *SMC1*, 78.58%; *SMC2*, 82.20%; *MLH1*, 74.49%; *MSH1*, 65.96%; and *MCM5*, 77.21%. *MSH1* was not tested further due to its



Organism	MLH1	MLH3	MLH4	MCM2	MCM3	MCM4	MCM5	MCM6	MCM7	MCM8	MCM9	SMC1	SMC2	SMC3	SMC4	SMC5	SMC6	Source
Arabidopsis thaliana	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	2	Phytozome v9.1
Arabidopsis lyrata	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	2	Phytozome v9.1
Brassica rapa	1	1	1	2	2	1	1	1	2	1	1	2	2	2	1	1	2	Phytozome v9.1
Capsella rubella	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	2	Phytozome v9.1
Capsella grandiflora	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	2	Phytozome v10
Boechera stricta	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	Phytozome v10
Eutrema salsuginea	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	2	Phytozome v9.1
Schrenkiella parvula	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	Thellungiel v2
Populus trichocarpa	1	1	1	2	1	1	2	1	1	1	1	1	2	1	1	1	1	Phytozome v9.1

Fig. 1. Gene copy number of *MLH*, *MCM*, and *SMC* family members in selected Brassicaceae species with sequenced genomes. The gene names are provided above and copy numbers are highlighted by different colors. Data source information is provided on the right. v, version.

relatively low sequence identity and possible difficulty in polymerase chain reaction (PCR) amplification. To further test the feasibility of using the remaining four genes in resolving species-level phylogeny, we examined the genomic sequences for both conserved regions for primer design flanking variable regions for phylogenetic signals and found *SMC2*, *MHL1*, and *MCM5* to be more favorable.

Primer design, amplification, and sequencing

For PCR reactions to amplify specific regions of the *SMC2*, *MHL1*, and *MCM5* genes, we designed primers based on comparison (Fig. 2) of genomic sequences of eight Brassicaceae species. We identified primers matching regions conserved among all these genome sequences, then used the leave-one-out approach (Berger et al., 2011) for assessing site-specific congruence as implemented in RAxML 8.0.2 (Stamatakis, 2006). This test prunes a single taxon at a time from a reference tree, carrying out site-specific computations for it, followed by reinserting into the original position. A sliding window of 100 sites in the alignment is used for calculation. The mean distance between the best placements for all sliding windows is calculated, assessing the phylogenetic variability of different areas of a gene. Targeted loci were located in the regions with the lowest average node distance. Degenerate sites were used when there were different sites among sampled species.

Several primer pairs for each of the three genes (*SMC2*, *MHL1*, and *MCM5*) were tested by PCR with multiple template DNAs and the one that was chosen had most reliably yielded amplified products across all samples; the primer sequences and their properties are listed in Table 3. All three primer pairs amplified loci with conserved exons flanking variable introns (Fig. 2). *SMC2* was typically amplified in a segment that contained two partial exons, three exons and four introns, making up 1174 base pairs (bp) in *A. thaliana*. The amplified *A. thaliana MLH1* segment contained two partial exons, five exons, and six introns, with 1462 bp. The *MCM5* segment from *A. thaliana* contained two partial exons, four exons, and five introns totaling 1313 bp. We used DNA polymerase from Takara (Otsu, Shiga, Japan) for PCR reactions. This DNA polymerase possesses high proofreading activity and can reduce mismatch in DNA amplification. Cycling conditions for *SMC2* started with initial denaturation at 94.0 °C for 5 min, followed by 30 cycles of amplification, 94.0 °C for 60 s,

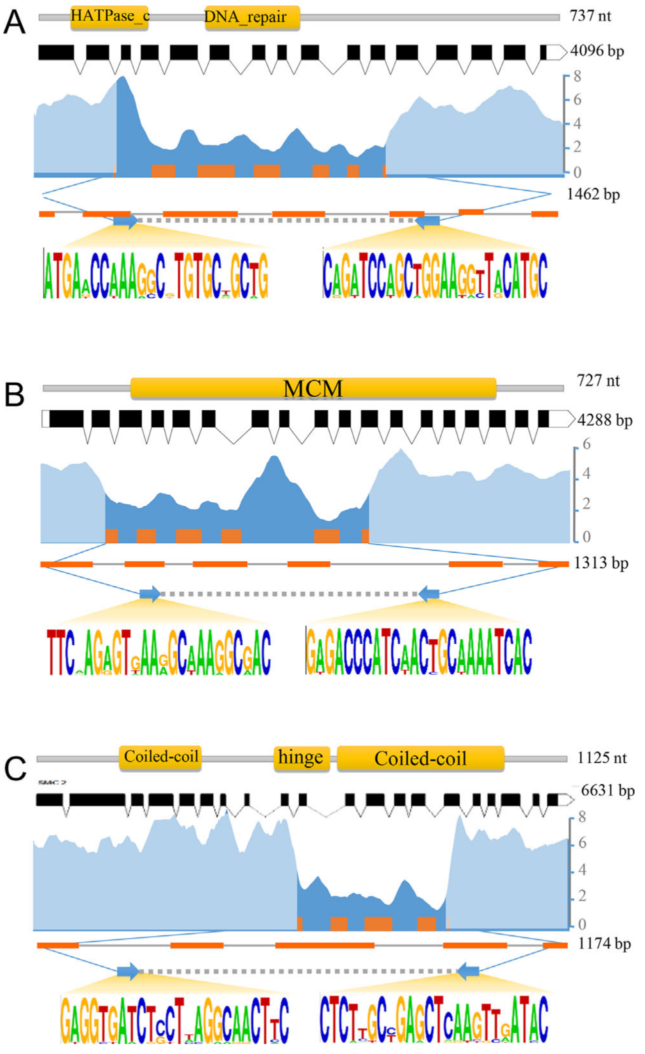


Fig. 2. Domains and nucleotide sequence conservation of *MLH1* (**A**), *MCM5* (**B**), and *SMC2* (**C**). Protein domains predicted by SMART are shown as yellow boxes at the top, with the number of translated amino acids in *Arabidopsis thaliana* shown on the right. Below the protein domain shown in black is the complete *A. thaliana* gene structure displayed by Exon-Intron Graphic Maker (http://wormweb.org/exonintron), with the number of nucleotides on the right. Below is a graph showing levels of site-specific congruence of phylogenetic signal identities among Brassicaceae species calculated by RAxML 8.0.2 (Stamatakis, 2006) with the window size of 100. Low value indicates high level of congruent phylogenetic signal. The region in dark blue represents the polymerase chain reaction (PCR)-amplified portion with the exons marked by orange, which give the most uniform phylogenetic signal. The gene structure of the PCR-amplified region is shown below with the exonic region in orange, intronic regions in gray, and the length of the *A. thaliana* PCR product shown on the right. Primers used in this study are marked as arrows, with the conserved and divergent sequences shown as generated using WebLogo (http://weblogo.berkeley.edu/logo.cgi). HATPase_c, histidine kinase-like ATPase, C-terminal domain (SMART accession number SM00387).

Table 3 Primer sequences and relative characters including length, Tm, GC%, and degeneracy

Gene	Primer	Sequence	Length, bp	Tm	GC%	Degeneracy
MLH1	MLH1_1F	5′-ATGAACCAAAGGCGTGTGCDGCTG-3′	24	66.3	54.2	3
MLH1	MLH1-2R	5′-ATTGATATCAACATGTTCVCGTGGCA-3′	25	63.9	52	1
MCM5	MCM5_1F	5′-TTCVAGAGTGAARGCAAAGGCGAC-3′	24	63.7	50	6
MCM5	MCM5_2R	5′-GTGATTTTGCAGTTGATGGGTCTC-3′	24	60.9	45.8	1
SMC2	SMC2_1F	5′-GAGGTGATCTCCTYAGGCAACTTC-3′	24	61	50	2
SMC2	SMC2_1R	5′-GTATYAACTTGAGCTCGGCAAGAG-3′	24	61.7	50	2

F, forward; R, reverse; Tm, annealing temperature; GC-content (GC%), percentage of either guanine or cytosine bases in a DNA molecule.

59.0 °C for 60 s, and 72.0 °C for 1.5 min, followed by a final extension at 72.0 °C for 5 min. Annealing temperatures for the other two genes were 60 °C for *MLH1* and 58 °C for *MCM5*. Additional cloning steps were used for some species within *Cardamine*, using vector pGEM-T and competent cell DH5 alpha. Three to five clones per amplification were sequenced on both forward and reverse strands using the same primer pair. Forward and reverse sequence strands were assembled with the ContigExpress program (http://www.contigexpress.com) and were then confirmed manually. Nucleotide sequences obtained from PCR and transcriptomes were submitted to GenBank, with accession numbers provided in Table S1.

Phylogenetic analysis

For sequence alignment and phylogenetic analysis, full-length *MLH1, MCM5*, and *SMC2* genes (with exons and introns) including 3051, 3005, and 4156 bp (for *A. thaliana*), respectively, were retrieved from genomic datasets, while the exon portion of their homologs were obtained from transcriptomic datasets. In addition, the PCR-amplified regions contained partial sequences with several exonic and intronic regions (mentioned above), whereas regions flanking the PCR-amplified loci were treated as missing data in the alignment for those taxa with only PCR-amplified sequences for the three genes. The sequences of each taxon were concatenated, forming a supermatrix with Seaview 4.4.2 (Gouy et al., 2010) and aligned using muscle 3.8.31 (Edgar, 2004) and subsequently adjusted manually. Maximum likelihood analyses were carried out using RAxML 8.0.2 (Stamatakis, 2006). Analysis was performed under the general time reversible model with the shape of the gamma distribution (GTR $+ \Gamma$) as determined by Modeltest 3.7 (Posada & Crandall, 1998). Optimal tree searches were carried out with 100 random sequence addition replicates. Branch support was assessed using 100 rapid bootstrap replicates. Maximum likelihood bootstrap proportions (BP) $\geq$70% were considered strong support (Hillis & Bull, 1993).

Bayesian analyses were implemented with MrBayes 3.1.2 (Ronquist & Huelsenbeck, 2003) under a time-free model. Posterior probability (PP) support values $\geq$0.95 were considered strong support for individual clades. MrBayes analyses were performed on the concatenated data. For consistency of results, two independent Markov chain Monte Carlo analyses were carried out for 200 000 generations to calculate PP. Prior probabilities for all trees were equal, starting trees were random, sampling every 1000 generations, and burn-in values were determined empirically from the likelihood values. The consistency of stationary-phase likelihood values and estimated parameter values was determined using Tracer 1.5 (Rambaut & Drummond, 2009). Bayesian PPs were determined by building a 50% majority-rule consensus tree from two Markov chain Monte Carlo analyses after discarding the 20% burn-in generations.

Results

Single-copy nuclear genes are excellent phylogenetic markers

As shown by Zhang et al. (2012), *MLH1*, *SMC2*, and *MCM5* are maintained as single-copy in most angiosperm species, consistent with an earlier report that genes engaged in DNA/RNA metabolisms tend to lose duplicate and remain orthologous after duplication (Blanc & Wolfe, 2004). Extensive phylogenetic studies have shown that members of *SMC*, *MCM*, and *MLH* gene families are maintained as one copy in most species (Forsburg, 2004; Lin et al., 2007; Surcel et al., 2008). These nuclear genes provide excellent markers to trace the evolution history of plants. We inspected their copy numbers in eight Brassicaceae species with fully sequenced genomes and found that most of them had only one copy for these species, except for *Brassica rapa*, which has undergone a recent whole genome duplication (Lysak et al., 2005; Yang et al., 2006). Two members from the *SMC* family (*SMC2* and *SMC6*) each have two copies across Brassicaceae. The *SMC2* duplication event occurred before the divergence of Brassicaceae (Fig. S1). There have been enough variations accumulated between the two copies, so that we could easily distinguish them, which is also seen in the application of two paralogous copies of *Adh1* and *Adh2*. In *Arabidopsis thaliana*, one copy of *SMC2* functions as subunit E in chromosome condensation complex, condensin. It is located on chromosome 5 (*AT5G62410*). Another copy functions similarly as structural maintenance of chromosome protein 2 and is mapped to chromosome 3 (*AT3G47460*). Both copies are maintained in all of the sequenced Brassicaceae genomes (Fig. 1). In our study, the homologs of *AT5G62410* were used as representative of *SMC2* for all sampled species.

The percentages of successful PCR for *SMC2*, *MCM5*, and *MLH1* were 93.7%, 87%, and 90.1%, respectively. These, along with gene sequences retrieved from public databases, include in total 53 *SMC2*, 55 *MCM5*, and 55 *MLH1* gene sequences from 63 Brassicaceae species. The lengths of each gene's initial alignment range from 3005 bp (*MCM5*) to 4156 bp (*SMC2*), with over one-third of the aligned regions covered by

Table 4 Characters of selected nuclear genes

Gene	Length, bp	Length of exon, bp	PI characters, bp (%)	Variable sites	GC%	α Parameter	Function anotation
MLH1	4096	2322	908 (39.1)	1265	41.49	0.4655	MUTL-homologue 1, DNA mismatch repair protein
MCM5	4288	2208	753 (34.1)	1066	42.69	0.4936	Minichromosome maintenance family protein, DNA replication licensing factor
SMC2	6364	3528	1333 (37.7)	1920	41.49	0.4655	Structural maintenance of chromosomes (SMC) family protein

GC-content (GC%), percentage of either guanine or cytosine bases in a DNA molecule.
Sequence lengths of selected genes, including length of exons, came from *Arabidopsis thaliana* (L.) Heynh. Parsimony-informative (PI) sites, variable sites, GC%, and gamma parameter for site rates (α Parameter) were calculated by mega. Function annotations were cited from Phytozome version 10.0 (http://phytozome.jgi.doe.gov/pz/portal.html).

the amplified loci, which contain both highly conserved and divergent sites. Gene length, length of exons, species coverage, parsimony-informative (PI) sites, variable sites, Guanine-cytosine (GC) frequencies, and α parameter (gamma parameter for site rates) are summarized in Table 4. The lengths of coding regions among different taxa are generally conserved. But lengths of introns could sometimes vary dramatically across taxa. We deleted such highly variable intronic regions in our alignment to reduce phylogenetic noise and the proportion of missing data. These highly variable regions mostly occurred in the middle of the intron with length varying from 6 bp to 25 bp in the alignment. Their corresponding positions in *A. thaliana* can be found in Table S2. The nucleotide sequences are highly conserved in exons (>92% global identities of all three genes). The conservation in terms of length, copy number, and exon sequences could be attributed to their functions in DNA/RNA metabolism and is important for primer design and sequence alignment. In the cases of our taxon sampling, the gamma parameters for site rates range from 0.466 for both *SMC2* and *MLH1*, to 0.494 for *MCM5*; thus, the three genes showed similar patterns of variability. Further analyses indicated that these genes are phylogenetically informative, with average frequency of PI sites greater than 30% (Table 4). The PI site proportion does not show significant heterogeneity between genes, but was especially high at third codon positions and in introns.

Further analyses of the nucleotide substitution model shows that $\mathrm{GTR}+\mathrm{I}+\Gamma$ is the fittest model for all three genes (Table S3). This result indicates these genes may have evolved under essentially very similar evolutionary patterns. A more detailed examination of site-specific placement bias in each gene reveals that the PCR-amplified region within each gene has the most stable phylogenetic signal (Fig. 2). Although the average node placement distance can be high in other regions of the genes, with possibly incongruent phylogenetic signals, this would not have a strong impact on our phylogenetic reconstructions in that only few genome or transcriptome data are available for those regions. For example, the PCR-amplified region from 540 to 1900 bp of the *MCM5* coding DNA sequence region (Fig. 2B) has a relatively low average node distance, whereas the intron region has higher phylogenetic divergence. Consequently, the exon regions with uniform phylogenetic signal are useful for resolving the deep nodes of early diversification. The divergent intron regions can provide crucial information to discern the subtle differences between closely related species.

Single-gene phylogenies were reconstructed for each of the three genes (Figs. S2–S4). They were largely consistent with well-established organismal relationships, suggesting that these genes were orthologous and phylogenetically informative. Although multiple sequences were occasionally obtained in one species, they always formed adjacent terminal branches in phylogenetic trees, suggesting that they resulted from recent polyploidization and should not affect phylogenetic relationships of more distantly related groups in this study.

Strongly supported phylogenies within *Cardamine, Lepidium*, and *Brassica*

Using a concatenated supermatrix of the sequences isolated from PCR amplification and retrieved from public databases, we built phylogenetic trees including 65 taxa. Our results showed that a combination of three nuclear loci resulted in a well-resolved phylogeny at various phylogenetic depths (Fig. 3). In the topology here with *Populus trichocarpa* as an outgroup, *Cleome serrulata* (Cleomaceae) was sister to a maximally supported clade of Brassicaceae; furthermore, *Aethionema* (including *Aethionema subulatum* and *Aethionema arabicum)* was the sister to all other Brassicaceae species (members of the core Brassicaceae), in agreement with previous studies (Al-Shehbaz et al., 2002; Huang et al., 2016). Within the core Brassicaceae clade, the phylogeny grouped all but one (*Arabis alpina*) of the remaining taxa into two clades. One included species from the previously defined Lineage I, and the other has those of Lineage II, again in agreement with recent results on Brassicaceae phylogeny (Al-Shehbaz et al., 2002; Huang et al., 2016). Eight species of the genus *Lepidium* formed a maximally supported group as the first divergent branch of Lineage I. Within the *Lepidium* clade, *L. campestre* (L.) W. T. Aiton and *L. perfoliatum* L. clustered as the sister to a large clade with the other *Lepidium* species (100 BP/1.0 PP). In addition, *L. didymum* L., which was formerly identified as *Coronopus didymus* (L.) Sm., is nested within *Lepidium* (100 BP/1.0 PP). This result was in agreement with a previous report on the systematics of *Lepidium* (Al-Shehbaz et al., 2002).

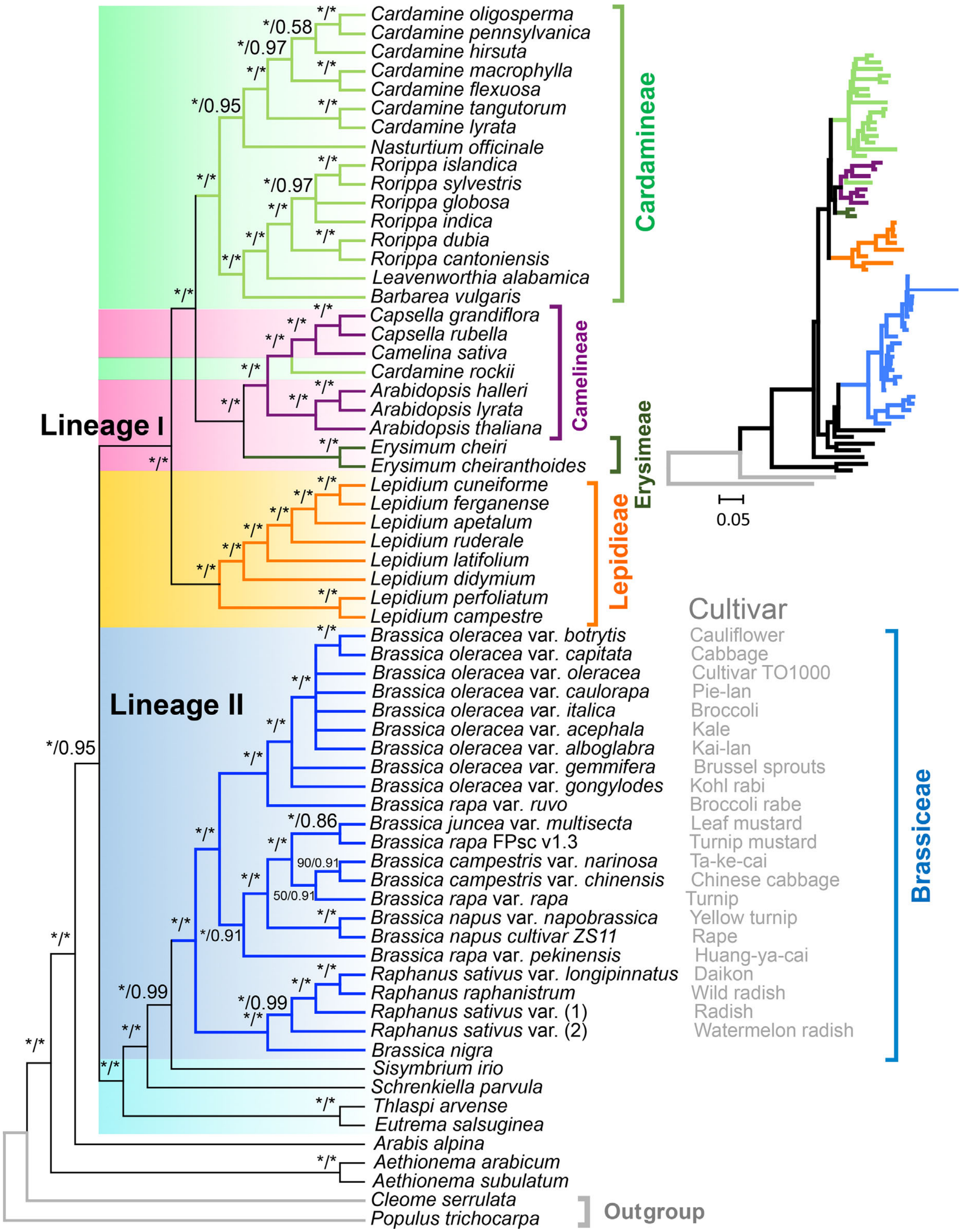
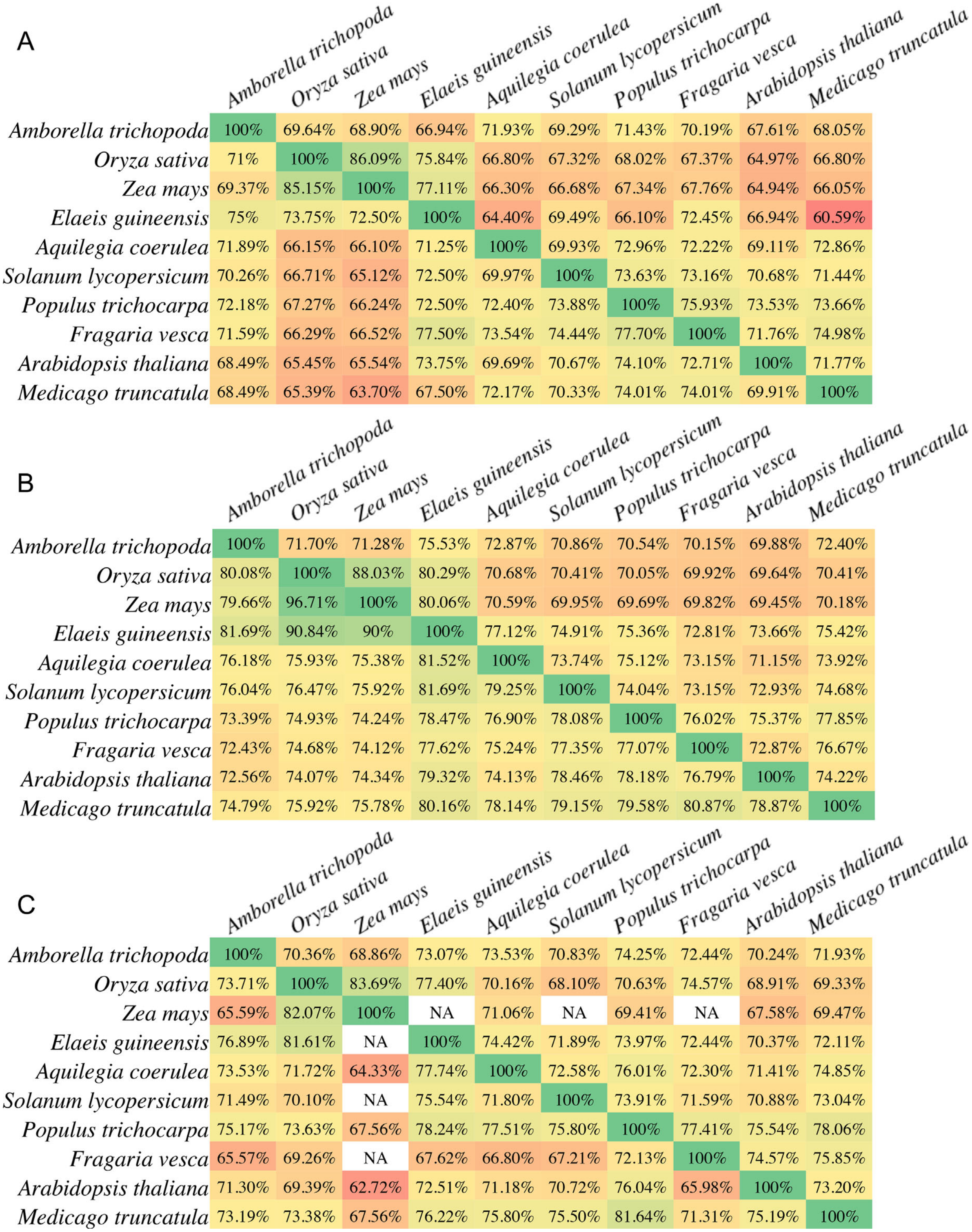


Fig. 3. Maximum likelihood majority-rule consensus tree of Brassicaceae based on the concatenated *MLH1*, *MCM5*, and *SMC2* datasets. Values above branches are maximum likelihood bootstrap values (left) and Bayesian inference posterior probabilities (right). Star indicates either a bootstrap proportion of 100 or posterior probability of 1.0. The phylogram of the reconstructed phylogeny is displayed on the upper right. Within clade Brassiceae, common names of cultivated vegetables are shown on the right.



A

	Amborella trichopoda	Oryza sativa	Zea mays	Elaeis guineensis	Aquilegia coerulea	Solanum lycopersicum	Populus trichocarpa	Fragaria vesca	Arabidopsis thaliana	Medicago truncatula
Amborella trichopoda	100%	69.64%	68.90%	66.94%	71.93%	69.29%	71.43%	70.19%	67.61%	68.05%
Oryza sativa	71%	100%	86.09%	75.84%	66.80%	67.32%	68.02%	67.37%	64.97%	66.80%
Zea mays	69.37%	85.15%	100%	77.11%	66.30%	66.68%	67.34%	67.76%	64.94%	66.05%
Elaeis guineensis	75%	73.75%	72.50%	100%	64.40%	69.49%	66.10%	72.45%	66.94%	60.59%
Aquilegia coerulea	71.89%	66.15%	66.10%	71.25%	100%	69.93%	72.96%	72.22%	69.11%	72.86%
Solanum lycopersicum	70.26%	66.71%	65.12%	72.50%	69.97%	100%	73.63%	73.16%	70.68%	71.44%
Populus trichocarpa	72.18%	67.27%	66.24%	72.50%	72.40%	73.88%	100%	75.93%	73.53%	73.66%
Fragaria vesca	71.59%	66.29%	66.52%	77.50%	73.54%	74.44%	77.70%	100%	71.76%	74.98%
Arabidopsis thaliana	68.49%	65.45%	65.54%	73.75%	69.69%	70.67%	74.10%	72.71%	100%	71.77%
Medicago truncatula	68.49%	65.39%	63.70%	67.50%	72.17%	70.33%	74.01%	74.01%	69.91%	100%

B

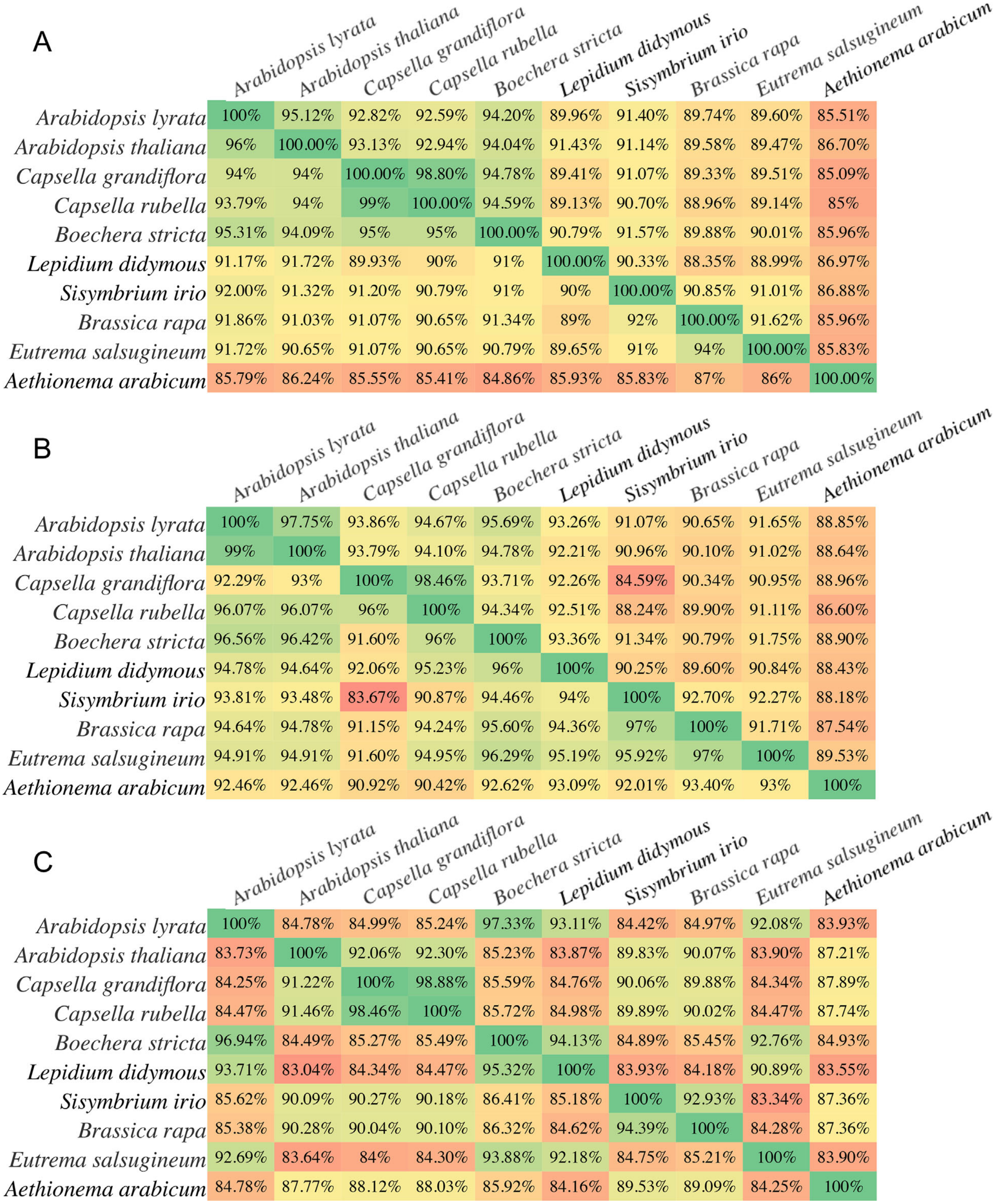
	Amborella trichopoda	Oryza sativa	Zea mays	Elaeis guineensis	Aquilegia coerulea	Solanum lycopersicum	Populus trichocarpa	Fragaria vesca	Arabidopsis thaliana	Medicago truncatula
Amborella trichopoda	100%	71.70%	71.28%	75.53%	72.87%	70.86%	70.54%	70.15%	69.88%	72.40%
Oryza sativa	80.08%	100%	88.03%	80.29%	70.68%	70.41%	70.05%	69.92%	69.64%	70.41%
Zea mays	79.66%	96.71%	100%	80.06%	70.59%	69.95%	69.69%	69.82%	69.45%	70.18%
Elaeis guineensis	81.69%	90.84%	90%	100%	77.12%	74.91%	75.36%	72.81%	73.66%	75.42%
Aquilegia coerulea	76.18%	75.93%	75.38%	81.52%	100%	73.74%	75.12%	73.15%	71.15%	73.92%
Solanum lycopersicum	76.04%	76.47%	75.92%	81.69%	79.25%	100%	74.04%	73.15%	72.93%	74.68%
Populus trichocarpa	73.39%	74.93%	74.24%	78.47%	76.90%	78.08%	100%	76.02%	75.37%	77.85%
Fragaria vesca	72.43%	74.68%	74.12%	77.62%	75.24%	77.35%	77.07%	100%	72.87%	76.67%
Arabidopsis thaliana	72.56%	74.07%	74.34%	79.32%	74.13%	78.46%	78.18%	76.79%	100%	74.22%
Medicago truncatula	74.79%	75.92%	75.78%	80.16%	78.14%	79.15%	79.58%	80.87%	78.87%	100%

C

	Amborella trichopoda	Oryza sativa	Zea mays	Elaeis guineensis	Aquilegia coerulea	Solanum lycopersicum	Populus trichocarpa	Fragaria vesca	Arabidopsis thaliana	Medicago truncatula
Amborella trichopoda	100%	70.36%	68.86%	73.07%	73.53%	70.83%	74.25%	72.44%	70.24%	71.93%
Oryza sativa	73.71%	100%	83.69%	77.40%	70.16%	68.10%	70.63%	74.57%	68.91%	69.33%
Zea mays	65.59%	82.07%	100%	NA	71.06%	NA	69.41%	NA	67.58%	69.47%
Elaeis guineensis	76.89%	81.61%	NA	100%	74.42%	71.89%	73.97%	72.44%	70.37%	72.11%
Aquilegia coerulea	73.53%	71.72%	64.33%	77.74%	100%	72.58%	76.01%	72.30%	71.41%	74.85%
Solanum lycopersicum	71.49%	70.10%	NA	75.54%	71.80%	100%	73.91%	71.59%	70.88%	73.04%
Populus trichocarpa	75.17%	73.63%	67.56%	78.24%	77.51%	75.80%	100%	77.41%	75.54%	78.06%
Fragaria vesca	65.57%	69.26%	NA	67.62%	66.80%	67.21%	72.13%	100%	74.57%	75.85%
Arabidopsis thaliana	71.30%	69.39%	62.72%	72.51%	71.18%	70.72%	76.04%	65.98%	100%	73.20%
Medicago truncatula	73.19%	73.38%	67.56%	76.22%	75.80%	75.50%	81.64%	71.31%	75.19%	100%

Fig. 4. Nucleotide and amino acid sequence identity of *MLH1* (**A**), *MCM5* (**B**), and *SMC2* (**C**) among 10 representative angiosperm species. Pairwise sequence identities are calculated using the SIAS webserver (http://imed.med.ucm.es/Tools/sias.html). Nucleotide and amino acid sequence identities are shown at upper right and bottom left, respectively. NA, data not available.

The other major branch of Lineage I was comprised of species belonging to the tribes Cardamineae, Camelineae, and Erysimeae. The Camelineae and Erysimeae members were more closely related (100 BP/1.0 PP) than they are to Cardamineae. The tribe Camelineae was represented by six species with sequenced genomes: *A. thaliana*, *A. lyrata*, *A. halleri* O'Kane & Al-Shehbaz, *C. grandiflora*, *C. rubella*, and *Camelina sativa* (L.) Crantz. In addition, *Cardamine rockii* O. E.



A

	Arabidopsis lyrata	Arabidopsis thaliana	Capsella grandiflora	Capsella rubella	Boechera stricta	Lepidium didymous	Sisymbrium irio	Brassica rapa	Eutrema salsugineum	Aethionema arabicum
Arabidopsis lyrata	100%	95.12%	92.82%	92.59%	94.20%	89.96%	91.40%	89.74%	89.60%	85.51%
Arabidopsis thaliana	96%	100.00%	93.13%	92.94%	94.04%	91.43%	91.14%	89.58%	89.47%	86.70%
Capsella grandiflora	94%	94%	100.00%	98.80%	94.78%	89.41%	91.07%	89.33%	89.51%	85.09%
Capsella rubella	93.79%	94%	99%	100.00%	94.59%	89.13%	90.70%	88.96%	89.14%	85%
Boechera stricta	95.31%	94.09%	95%	95%	100.00%	90.79%	91.57%	89.88%	90.01%	85.96%
Lepidium didymous	91.17%	91.72%	89.93%	90%	91%	100.00%	90.33%	88.35%	88.99%	86.97%
Sisymbrium irio	92.00%	91.32%	91.20%	90.79%	91%	90%	100.00%	90.85%	91.01%	86.88%
Brassica rapa	91.86%	91.03%	91.07%	90.65%	91.34%	89%	92%	100.00%	91.62%	85.96%
Eutrema salsugineum	91.72%	90.65%	91.07%	90.65%	90.79%	89.65%	91%	94%	100.00%	85.83%
Aethionema arabicum	85.79%	86.24%	85.55%	85.41%	84.86%	85.93%	85.83%	87%	86%	100.00%

B

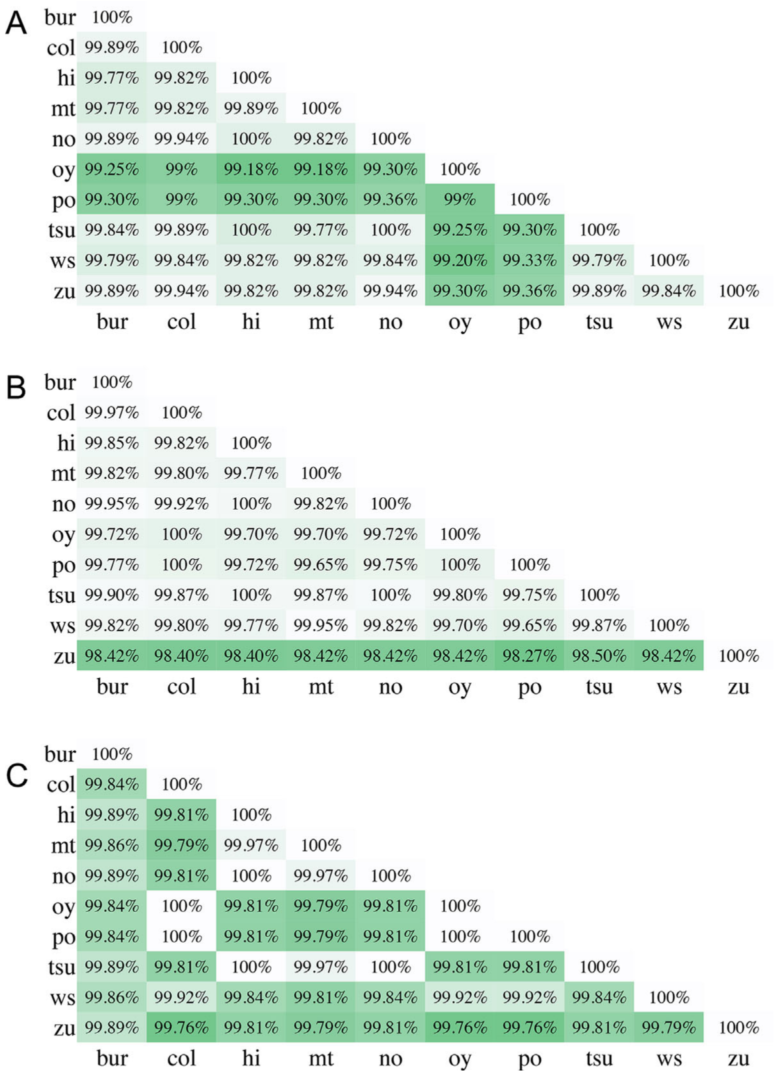
	Arabidopsis lyrata	Arabidopsis thaliana	Capsella grandiflora	Capsella rubella	Boechera stricta	Lepidium didymous	Sisymbrium irio	Brassica rapa	Eutrema salsugineum	Aethionema arabicum
Arabidopsis lyrata	100%	97.75%	93.86%	94.67%	95.69%	93.26%	91.07%	90.65%	91.65%	88.85%
Arabidopsis thaliana	99%	100%	93.79%	94.10%	94.78%	92.21%	90.96%	90.10%	91.02%	88.64%
Capsella grandiflora	92.29%	93%	100%	98.46%	93.71%	92.26%	84.59%	90.34%	90.95%	88.96%
Capsella rubella	96.07%	96.07%	96%	100%	94.34%	92.51%	88.24%	89.90%	91.11%	86.60%
Boechera stricta	96.56%	96.42%	91.60%	96%	100%	93.36%	91.34%	90.79%	91.75%	88.90%
Lepidium didymous	94.78%	94.64%	92.06%	95.23%	96%	100%	90.25%	89.60%	90.84%	88.43%
Sisymbrium irio	93.81%	93.48%	83.67%	90.87%	94.46%	94%	100%	92.70%	92.27%	88.18%
Brassica rapa	94.64%	94.78%	91.15%	94.24%	95.60%	94.36%	97%	100%	91.71%	87.54%
Eutrema salsugineum	94.91%	94.91%	91.60%	94.95%	96.29%	95.19%	95.92%	97%	100%	89.53%
Aethionema arabicum	92.46%	92.46%	90.92%	90.42%	92.62%	93.09%	92.01%	93.40%	93%	100%

C

	Arabidopsis lyrata	Arabidopsis thaliana	Capsella grandiflora	Capsella rubella	Boechera stricta	Lepidium didymous	Sisymbrium irio	Brassica rapa	Eutrema salsugineum	Aethionema arabicum
Arabidopsis lyrata	100%	84.78%	84.99%	85.24%	97.33%	93.11%	84.42%	84.97%	92.08%	83.93%
Arabidopsis thaliana	83.73%	100%	92.06%	92.30%	85.23%	83.87%	89.83%	90.07%	83.90%	87.21%
Capsella grandiflora	84.25%	91.22%	100%	98.88%	85.59%	84.76%	90.06%	89.88%	84.34%	87.89%
Capsella rubella	84.47%	91.46%	98.46%	100%	85.72%	84.98%	89.89%	90.02%	84.47%	87.74%
Boechera stricta	96.94%	84.49%	85.27%	85.49%	100%	94.13%	84.89%	85.45%	92.76%	84.93%
Lepidium didymous	93.71%	83.04%	84.34%	84.47%	95.32%	100%	83.93%	84.18%	90.89%	83.55%
Sisymbrium irio	85.62%	90.09%	90.27%	90.18%	86.41%	85.18%	100%	92.93%	83.34%	87.36%
Brassica rapa	85.38%	90.28%	90.04%	90.10%	86.32%	84.62%	94.39%	100%	84.28%	87.36%
Eutrema salsugineum	92.69%	83.64%	84%	84.30%	93.88%	92.18%	84.75%	85.21%	100%	83.90%
Aethionema arabicum	84.78%	87.77%	88.12%	88.03%	85.92%	84.16%	89.53%	89.09%	84.25%	100%

Fig. 5. Nucleotide and amino acid sequence identity of *MLH1* (**A**), *MCM5* (**B**), and *SMC2* (**C**) among 10 representative Brassicaceae species. Pairwise sequence identities are calculated using the SIAS webserver (http://imed.med.ucm.es/Tools/sias.html). Nucleotide and amino acid sequences identities are displayed at upper right and bottom left, respectively.

Schulz clustered with a clade composed of *Camelina* Crantz and *Capsella* Medik. (100 BP/1.0 PP), rather than being close to members of Cardamineae. *Erysimum cheiri* Crantz and *Erysimum cheiranthoides* L., with sequences from transcriptomics, are sisters with maximal support, representing the tribe Erysimeae. Our sampling of the tribe Cardamineae included 16 species from five genera. Cardamineae were divided into two clades (100 BP/1.0 PP),



A

	bur	col	hi	mt	no	oy	po	tsu	ws	zu
bur	100%									
col	99.89%	100%								
hi	99.77%	99.82%	100%							
mt	99.77%	99.82%	99.89%	100%						
no	99.89%	99.94%	100%	99.82%	100%					
oy	99.25%	99%	99.18%	99.18%	99.30%	100%				
po	99.30%	99%	99.30%	99.30%	99.36%	99%	100%			
tsu	99.84%	99.89%	100%	99.77%	100%	99.25%	99.30%	100%		
ws	99.79%	99.84%	99.82%	99.82%	99.84%	99.20%	99.33%	99.79%	100%	
zu	99.89%	99.94%	99.82%	99.82%	99.94%	99.30%	99.36%	99.89%	99.84%	100%

B

	bur	col	hi	mt	no	oy	po	tsu	ws	zu
bur	100%									
col	99.97%	100%								
hi	99.85%	99.82%	100%							
mt	99.82%	99.80%	99.77%	100%						
no	99.95%	99.92%	100%	99.82%	100%					
oy	99.72%	100%	99.70%	99.70%	99.72%	100%				
po	99.77%	100%	99.72%	99.65%	99.75%	100%	100%			
tsu	99.90%	99.87%	100%	99.87%	100%	99.80%	99.75%	100%		
ws	99.82%	99.80%	99.77%	99.95%	99.82%	99.70%	99.65%	99.87%	100%	
zu	98.42%	98.40%	98.40%	98.42%	98.42%	98.42%	98.27%	98.50%	98.42%	100%

C

	bur	col	hi	mt	no	oy	po	tsu	ws	zu
bur	100%									
col	99.84%	100%								
hi	99.89%	99.81%	100%							
mt	99.86%	99.79%	99.97%	100%						
no	99.89%	99.81%	100%	99.97%	100%					
oy	99.84%	100%	99.81%	99.79%	99.81%	100%				
po	99.84%	100%	99.81%	99.79%	99.81%	100%	100%			
tsu	99.89%	99.81%	100%	99.97%	100%	99.81%	99.81%	100%		
ws	99.86%	99.92%	99.84%	99.81%	99.84%	99.92%	99.92%	99.84%	100%	
zu	99.89%	99.76%	99.81%	99.79%	99.81%	99.76%	99.76%	99.81%	99.79%	100%

Fig. 6. Pairwise nucleotide sequence identity of *MLH1* (**A**), *MCM5* (**B**), and *SMC2* (**C**) among 10 representative *Arabidopsis thaliana* *ecotypes, calculated using the SIAS webserver (http://imed.med.ucm.es/Tools/sias.html). Data were collected from the 19 genomes of the A. thaliana project (http://mus.well.ox.ac.uk/19genomes/). Abbreviations for ecotype names follow Gan et al. (2011): bur, Burren; col, Columbia; hi, Hilversum; mt, Martuba; no, Nossen; oy, Oystese; po, Poppelsdorf; tsu, Tsu; ws, Wassilewskija; zu, Zurich.*

one containing genera *Barbarea* W. T. Aiton, *Leavenworthia*, and *Rorippa* (100 BP/1.0 PP), the other containing *Nasturtium* and *Cardamine* (100 BP/0.95 PP). Within *Rorippa*, six species were divided into two clades, containing two and four species, respectively. *Rorippa dubia* (Pers.) Hara and *Rorippa cantoniensis* (Lour.) Ohwi formed one of the clades (100 BP/1.0 PP). Within the other clade, *R. islandica* (Oeder) Borbás and *R. sylvestris* (L.) Besser share the closest affinity (100 BP/1.0 PP), with their relationship to the other two species unclear. Similarly, *Nasturtium officinale* W. T. Aiton was sister to a well-supported clade of seven *Cardamine* species. Relationships within genus *Cardamine* were more complicated. A basal clade was composed of two species *C. lyrata* Bunge and *C. tangutorum* O. E. Schulz (100 BP/1.0 PP). *Cardamine oligosperma* Nutt. and *C. pensylvanica* Muhl. ex Willd. were sisters with maximal support, as were *C. flexuosa* With. and *C. macrophylla* Willd. (both 100 BP/1.0 PP), and the position of *C. hirsuta* L. as sister to the clade of *C. oligosperma* and *C. pensylvanica* was moderately supported (100 BP/0.58 PP).

In Lineage II, four species belonging to the previously defined Expanded Lineage II formed successive sisters to the tribe Brassiceae, with maximal support. These species included *Sisymbrium irio*, *Thlaspi arvense*, *Eutrema salsuginea*, and *Schrenkiella parvula*. Our sampled taxa of the tribe Brassiceae included two genera, *Brassica* and *Raphanus* and they form a maximally supported clade. Within Brassiceae, all four *Raphanus sativus* L. cultivars form a well-supported clade (100 BP/1.0 PP), and most *Brassica* taxa form another maximally supported clade. Our results indicated that *Brassica nigra* (L.) W. D. J. Koch was more closely related to *Raphanus* than to other *Brassica* species (100 BP/1.0 PP), which was also consistent with previous findings that *Brassica* is not monophyletic (Yang et al., 2002; Arias & Pires, 2012). Within the large *Brassica* clade, the genetic similarity and complicated domestication histories of various cultivars brought challenge for phylogenetic reconstruction. However, the introns and third codon position of the nuclear markers provided crucial information for good resolution even within species. Four *Brassica rapa* accessions held four different positions. Cultivars belonging to *Brassica oleracea* L. composed the majority of our sampling diversity within *Brassica*, forming a maximally supported clade with internal resolutions of either high or low support values. The relationships among *B. oleracea*, *B. rapa*, *B. napus* L., *B. juncea* (L.) Czern., and *B. campestris* L. was not clear due to the limitation of sequence data and the non-monophyletic results. Among these vegetables, *B. napus* L. var. *napobrassica* (L.) Hanelt, or yellow turnip, is a cross between the cabbage and the turnip. Its placement on the tree suggested that it was more similar to cabbage (*B. oleracea*) than others. *Arabis alpina* was the basal-most core Brassicaceae species (100 BP/1.0 PP) among taxa sampled here.

Analysis of sequence similarity for three marker genes among angiosperms

The above phylogenetic analysis suggested that the *MCM5*, *MHL1*, and *SMC2* genes could be effective markers for revolving relationships between members of a family or even a genus. To further explore the sequence similarities of these genes for their potential applications in the phylogenetic studies of plants of various evolutionary diversities, we obtained their homologs in representative angiosperms, Brassicaceae species, as well as *A. thaliana* ecotypes (Gan et al., 2011) and compared their pairwise sequence identity (Figs. 4–6). Ten angiosperm species were selected to be phylogenetically representative. This included the basal-most angiosperm *Amborella trichopoda* Baill., three monocot species, and representatives of major eudicot clades. Within this angiosperm sampling, the pairwise sequence identity of *MCM5* ranged between 72.43%–96.71% for amino acid sequences and 69.45%–88.03% for nucleotide acid sequences (Fig. 4). Similar results can also be seen for *SMC2* and *MLH1*. When we compared sequence identities within the Brassicaceae family, higher pairwise identities could be obtained. Generally, the nucleotide sequence identity ranges from 83.34% to 98.88% and the amino acid sequence identity ranges from 83.04% to 99.00%. Likewise, results from the 10 *A. thaliana* accessions revealed that, although both nucleotide and amino acid sequences are highly similar among different ecotypes, we can still find SNP and indel sites within the sampled loci. Because these genes are unusually long with 3000 bp or more, even a low percentage of differences can provide dozens or more sites for comparison. As shown in Fig. 1, other members of these three gene families are also stably maintained as single copy or low copy, providing additional markers if more information is needed.

Discussion

Newly identified nuclear genes are suitable for phylogenetic reconstructions

Extensive phylogenetic analyses have been undertaken using mainly organellar or rDNA markers. Although they are easy to obtain, they are often too conserved to provide sufficient signals for resolving relatively close relationships. Nuclear genes are both numerous and rich in phylogenetic signals, but many nuclear genes have paralogs and should be used with care to avoid misleading signals. Previous work on angiosperm phylogeny revealed that nuclear markers were also highly informative for low-rank taxonomic groups (Zhu & Ge, 2005; Yuan et al., 2009; Salas-Leiva et al., 2013). These genes are primarily single-copy and conserved. The nuclear markers used here were previously described for use in a study of angiosperm-wide phylogeny (Zhang et al., 2012) and they were comparable to other markers that were screened from ~1000 low-copy putative orthologous genes by comparing genomes of representative angiosperms (Zeng et al., 2014). They were reported to be suitable for angiosperm phylogenetic reconstruction when conserved exon sequences were used. Here, to provide more divergent sequences with signals for within-genus relationships, we took advantage of both exonic and intronic regions for phylogenetic reconstruction. As illustrated by our result, the highly conserved exonic regions could facilitate the design of primers with high amplification efficiency for a wide range of organisms. First, these loci are easy to amplify by PCR reaction. Cloning steps are not necessary unless there are occasional recent gene duplication events. Second, conserved exons make it easy to align across distantly related species. Finally, together with intron sequences, the nuclear gene markers provide information at various phylogenetic depths. They are especially powerful for resolving relationships involving recent and rapid radiation.

With the development of sequencing technology, more and more genome and transcriptome data will be available. The growing genome datasets will facilitate the identification of low-copy nuclear genes as phylogenetic markers for more and more plant groups. At the same time, the PCR-based approach presented here provides a complementary means for obtaining a small number of genes from a large number of taxa, without the expense of transcriptomics and avoiding the need for great computation capability that is associated with the analysis of many genes. Furthermore, the information on the sequence similarity indicate that there are many variable sites from the comparison of different Brassicaceae species; there are even variations between different *Arabidopsis thaliana* ecotypes. This information and the phylogenetic results (see below) together provide strong evidence that these genes can serve as effective markers for investigation of relationships between species in the same genus.

Potential newly defined species relationships in Brassicaceae

One of the advantages of nuclear gene markers is their usability for phylogenies at various depths. For example, with three nuclear loci, we obtained placement of all sampled genera congruent to the latest and most comprehensive analysis (Al-Shehbaz, 2012) with strong support (100 BP/>0.95 PP). Within each genus, internal nodes are well resolved, providing important information to investigate their recent evolutionary history. At the same time, greater resolution for some of the relationships among species would probably benefit from some additional sequences, either from these three genes, or from other similarly conserved genes. In addition, very difficult relationships might also need more genes, as shown recently by the phylogenetic study of a greater number of tribes in Brassicaceae (Huang et al., 2016).

Cardamineae is a tribe containing 14 genera and 352 species. We sampled 16 species from five genera, which were grouped into two clades: one clade with *Cardamine* and *Nasturtium*, the other containing *Rorippa*, *Leavenworthia*, and *Barbarea*. Previous studies found a close relationship between *Cardamine* and *Rorippa* (Yang et al., 1999), and between *Cardamine* and *Nasturtium* (Beilstein et al., 2006), but they did not include all these genera. Our placement of the five genera within tribe Cardamineae has received the highest support so far. This is in good agreement with the initial hypothesis that *Nasturtium* is more closely related to *Cardamine* than to other genera in tribe Cardamineae (Al-Shehbaz & Price, 1998). Our results contribute to the understanding of the early divergence events within Cardamineae. Additionally, we found that one *Cardamine* species, *Cardamine rockii*, did not cluster with other *Cardamine* species, but rather it was grouped with members of Camelineae with strong support (100 BP/1.0 PP). Such phylogenetic placement has not been reported before. Thus this result suggests that this species might be misclassified and further investigation with more *Cardamine* species and Camelineae members will be needed to test this idea.

Within *Lepidium*, all relationships were strongly supported in the phylogeny here (100 BP/1.0 PP). *Lepidium perfoliatum* and *L. campestre* (100 BP/1.0 PP) formed the basal clade of *Lepidium*, consistent with previous results (Mummenhoff et al., 2001; Lee et al., 2002). Both of these studies placed the two species as the basal lineage with a larger sampling size in *Lepidium*. Three species, *L. apetalum* Willd., *L. ferganense* Korsh., and *L. cuneiforme* C. Y. Wu, form a highly supported group, allowing the placement of *L. cuneiforme*, which is endemic to China, for the first time. In addition, the placement of *L. apetalum* is similar to the results based on ITS sequences (Mummenhoff et al., 2001). The markers used here might be able to resolve the relationships in *Lepidium* when more species can be analyzed in the future.

Previous studies have shown that *Brassica* is not a monophyletic group (Yang et al., 2002; Arias & Pires, 2012). For example, a phylogenetic study with *B. rapa*, *B. nigra*, and *R. sativus* (Yang et al., 2002) indicated that *B. nigra* is sister to the clade of *B. rapa* and *R. sativus*. Human domestication of taxa within tribe Brassiceae also caused difficulty for tracing their ancient origins. In particular, multiple hybridization events between different cultivars along with whole genome duplication make it a major challenge to build phylogenetic trees across *Brassica* using nuclear genes. Our attempt revealed that the three nuclear markers used, *MLH1*, *SMC2*, and *MCM5*, are sufficiently variable and useful for detecting subtle differences between accessions within a species. Human domestication of *Brassica* is often explained by the U triangle theory (Nagaharu, 1935). The theory states the evolutionary relationships between modern vegetables and three ancestral species of *Brassica* by comparing their chromosome number. However, the long-discussed U triangle theory has not been rigorously tested by phylogenetic analysis. Our sampling of U triangle species included assumed ancestor species (*B. nigra*, *B. oleracea*, and *B. rapa*), as well as modern vegetables and oil seed crops (*B. juncea* and *B. napus*). This provides a unique opportunity to test the U triangle theory and further investigate the impact of hybridization on phylogenetic reconstruction. The problem lies in that our four accessions of species *B. rapa* hold four different places on the tree. This could be a result either from incorrect classification of vegetables, or other problems. More marker genes and more taxa are needed to resolve the proposed hybridization events. Nevertheless, the preliminary analysis here suggests that these nuclear genes contain variable sequences with phylogenetic signals that could be used to address such difficult questions.

Acknowledgements

We thank Professors Dunyan Tan, Hongzhi Kong, Guangyuan Rao, and Ji Yang for their help in plant sample collection and species identification, and Yi Hu for RNA extraction of some taxa. We are also grateful to Ning Zhang, Liping Zeng, and Qi Li for their help in computational analysis. In addition, assistance from Yue Jin, Engao Zhu, Chenjiang You, and Bailong Zhang in laboratory work was of great importance and we wish to express our gratitude. This work was supported by a grant from the National Natural Science Foundation of China (Grant No. 91531301) to H.M.

References

Al-Shehbaz IA. 2012. A generic and tribal synopsis of the Brassicaceae (Cruciferae). *Taxon* 61: 931–954.

Al-Shehbaz IA, Mummenhoff K, Appel O. 2002. *Cardaria*, *Coronopus*, and *Stroganowia* are united with *Lepidium* (Brassicaceae). *Novon* 12: 5–11.

Al-Shehbaz IA, Price RA. 1998. Delimitation of the genus *Nasturtium* (Brassicaceae). *Novon* 8: 124–126.

Álvarez I, Wendel JF. 2003. Ribosomal ITS sequences and plant phylogenetic inference. *Molecular Phylogenetics and Evolution* 29: 417–434.

Arias T, Pires JC. 2012. A fully resolved chloroplast phylogeny of the *Brassica* crops and wild relatives (Brassicaceae: Brassiceae): Novel clades and potential taxonomic implications. *Taxon* 61: 980–988.

Bailey CD, Koch MA, Mayer M, Mummenhoff K, O'Kane SL, Warwick SI, Windham MD, Al-Shehbaz IA. 2006. Toward a global phylogeny of the Brassicaceae. *Molecular Biology and Evolution* 23: 2142–2160.

Baldwin BG, Markos S. 1998. Phylogenetic utility of the External Transcribed Spacer (ETS) of 18S–26S rDNA: Congruence of ETS and ITS trees of *Calycadenia* (Compositae). *Molecular Phylogenetics and Evolution* 10: 449–463.

Baldwin BG, Sanderson MJ, Porter JM, Wojciechowski MF, Campbell CS, Donoghue MJ. 1995. The ITS region of nuclear ribosomal DNA: A valuable source of evidence on angiosperm phylogeny. *Annals of the Missouri Botanical Garden* 82: 247–277.

Beilstein MA, Al-Shehbaz IA, Kellogg EA. 2006. Brassicaceae phylogeny and trichome evolution. *American Journal of Botany* 93: 607–619.

Beilstein MA, Al-Shehbaz IA, Mathews S, Kellogg EA. 2008. Brassicaceae phylogeny inferred from phytochrome A and *ndhF* sequence data: Tribes and trichomes revisited. *American Journal of Botany* 95: 1307–1327.

Berger SA, Krompass D, Stamatakis A. 2011. Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. *Systematic Biology* 60: 291–302.

Birky CW. 1995. Uniparental inheritance of mitochondrial and chloroplast genes: Mechanisms and evolution. *Proceedings of the National Academy of Sciences USA* 92: 11 331–11 338.

Blanc G, Wolfe KH. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* 16: 1667–1678.

Buckler ES, Ippolito A, Holtsford TP. 1997. The evolution of ribosomal DNA divergent paralogues and phylogenetic implications. *Genetics* 145: 821–832.

De Bodt S, Maere S, Van de Peer Y. 2005. Genome duplication and the origin of angiosperms. *Trends in Ecology & Evolution* 20: 591–597.

Ding M, Zeng L, Zhang N, Ma H. 2012. The use of low-copy nuclear genes for reconstructing the phylogeny of low-level taxonomic hierarchies: Evidence from Brassicaceae. *Plant Diversity and Resources* 34: 211–221.

Edgar RC. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32: 1792–1797.

Forsburg SL. 2004. Eukaryotic MCM proteins: Beyond replication initiation. *Microbiology and Molecular Biology Reviews* 68: 109–131.

Fukuda T, Yokoyama J, Nakamura T, Song I-J., Ito T, Ochiai T, Kanno A, Kameya T, Maki M. 2005. Molecular phylogeny and evolution of alcohol dehydrogenase (*Adh*) genes in legumes. *BMC Plant Biology* 5: 6.

Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, Lyngsoe R, Schultheiss SJ, Osborne EJ, Sreedharan VT, Kahles A, Bohnert R, Jean G, Derwent P, Kersey P, Belfield EJ, Harberd NP, Kemen E, Toomajian C, Kover PX, Clark RM, Rätsch G, Mott R. 2011. Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature* 477: 419–423.

Gaut BS, Clegg MT. 1991. Molecular evolution of *Alcohol Dehydrogenase 1* in members of the grass family. *Proceedings of the National Academy of Sciences USA* 88: 2060–2064.

Gouy M, Guindon S, Gascuel O. 2010. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution* 27: 221–224.

Gozuacik D, Chami M, Lagorce D, Faivre J, Murakami Y, Poch O, Biermann E, Knippers R, Bréchot C, Paterlini-Bréchot P. 2003. Identification and functional characterization of a new member of the human Mcm protein family: hMcm8. *Nucleic Acids Research* 31: 570–579.

Hillis DM, Bull JJ. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology* 42: 182–192.

Huang CH, Sun R, Hu Y, Zeng L, Zhang N, Cai L, Zhang Q, Koch MA, Al-Shehbaz I, Edger PP, Pires JC. 2016. Resolution of Brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. *Molecular Biology and Evolution* 33: 394–412.

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, Soltis DE, Clifton SW, Schlarbaum SE, Schuster SC, Ma H, Leebens-Mack J, dePamphilis CW. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* 473: 97–100.

Kim C, Shin H, Chang YT, Choi HK. 2010. Speciation pathway of *Isoëtes* (Isoëtaceae) in East Asia inferred from molecular phylogenetic relationships. *American Journal of Botany* 97: 958–969.

Lee JY, Mummenhoff K, Bowman JL. 2002. Allopolyploidization and evolution of species with reduced floral structures in *Lepidium* L. (Brassicaceae). *Proceedings of the National Academy of Sciences USA* 99: 16835–16840.

Lin Z, Nei M, Ma H. 2007. The origins and early evolution of DNA mismatch repair genes – multiple horizontal gene transfers and co-evolution. *Nucleic Acids Research* 35: 7591–7603.

Lysak MA, Koch MA, Pecinka A, Schubert I. 2005. Chromosome triplication found across the tribe Brassiceae. *Genome Research* 15: 516–525.

Maddison WP. 1997. Gene trees in species trees. *Systematic Biology* 46: 523–536.

Mummenhoff K, Brüggemann H, Bowman JL. 2001. Chloroplast DNA phylogeny and biogeography of *Lepidium* (Brassicaceae). *American Journal of Botany* 88: 2051–2063.

Nagaharu U. 1935. Genome analysis in *Brassica* with special reference to the experimental formation of *B. napus* and peculiar mode of fertilization. *Japanese Journal of Botany* 7: 389–452.

Nie ZL, Wen J, Azuma H, Qiu YL, Sun H, Meng Y, Sun W-B, Zimmer EA. 2008. Phylogenetic and biogeographic complexity of Magnoliaceae in the Northern Hemisphere inferred from three nuclear data sets. *Molecular Phylogenetics and Evolution* 48: 1027–1040.

Oh SH, Potter D. 2005. Molecular phylogenetic systematics and biogeography of tribe Neillieae (Rosaceae) using DNA sequences of cpDNA, rDNA, and *LEAFY*. *American Journal of Botany* 92: 179–192.

Posada D, Crandall KA. 1998. Modeltest: Testing the model of DNA substitution. *Bioinformatics* 14: 817–818.

Rambaut A, Drummond A. 2009. Tracer version 1.5.0. [online]. Available from http://tree.bio. ed.ac.uk/software/tracer/ [accessed 1 July 2013].

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19: 1572–1574.

Sang T. 2002. Utility of low-copy nuclear gene sequences in plant phylogenetics. *Critical Reviews in Biochemistry and Molecular Biology* 37: 121–147.

Schranz ME, Mitchell-Olds T. 2006. Independent ancient polyploidy events in the sister families Brassicaceae and Cleomaceae. *Plant Cell* 18: 1152–1165.

Salas-Leiva DE, Meerow AW, Calonje M, Griffith MP, Francisco-Ortega J, Nakamura K, Stevenson DW, Lewis CE, Namoff S. 2013. Phylogeny of the cycads based on multiple single-copy nuclear genes: Congruence of concatenated parsimony, likelihood and species tree inference methods. *Annals of Botany* 112: 1263–1278.

Small RL, Cronn RC, Wendel JF. 2004. LAS Johnson Review No. 2. Use of nuclear genes for phylogeny reconstruction in plants. *Australian Systematic Botany* 17: 145–170.

Small RL, Ryburn JA, Cronn RC, Seelanan T, Wendel JF. 1998. The tortoise and the hare: Choosing between noncoding plastome and nuclear *Adh* sequences for phylogeny reconstruction in a recently diverged plant group. *American Journal of Botany* 85: 1301–1315.

Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng C, Sankoff D, Wall PK, Soltis PS. 2009. Polyploidy and angiosperm diversification. *American Journal of Botany* 96: 336–348.

Stamatakis A. 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22: 2688–2690.

Stewart C, Via LE. 1993. A rapid CTAB DNA isolation technique useful for RAPD fingerprinting and other PCR applications. *Biotechniques* 14: 748–750.

Surcel A, Zhou X, Quan L, Ma H. 2008. Long-term maintenance of stable copy number in the eukaryotic SMC family: Origin of a vertebrate meiotic SMC1 and fate of recent segmental duplicates. *Journal of Systematics and Evolution* 46: 405–423.

Warwick SI, Mummenhoff K, Sauder CA, Koch MA, Al-Shehbaz IA. 2010. Closing the gaps: Phylogenetic relationships in the Brassicaceae based on DNA sequence data of nuclear ribosomal ITS region. *Plant Systematics and Evolution* 285: 209–232.

Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, Matasci N, Ayyampalayam S, Barker MS, Burleigh JG, Gitzendanner MA, Ruhfel BR, Wafula E, Der JP, Graham SW, Mathews S, Melkonian M, Soltis DE, Soltis PS, Miles NW, Rothfels CJ, Pokorny L, Shaw AJ, DeGironimo L, Stevenson DW, Surek B, Villarreal JC, Roure B, Philippe H, dePamphilis CW, Chen T, Deyholos MK, Baucom RS, Kutchan TM, Augustin MM, Wang J, Zhang Y, Tian Z, Yan Z, Wu X, Sun X, Wong GK-S, Leebens-Mack J. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences USA* 111: E4859–E4868.

Yang TJ, Kim JS, Kwon SJ, Lim KB, Choi BS, Kim JA, Jin M, Park JY, Lim MH, Kim HI, Lim YP, Kang JJ, Hong JH, Kim CB, Bhak J, Bancroft I, Park BS. 2006. Sequence-level analysis of the diploidization process in the triplicated *FLOWERING LOCUS C* region of *Brassica rapa*. *Plant Cell* 18: 1339–1347.

Yang Y, Moore MJ, Brockington SF, Soltis DE, Wong GK-S, Carpenter EJ, Zhang Y, Chen L, Yan Z, Xie Y, Sage RF, Covshoff S, Hibberd JM, Nelson MN, Smith SA. 2015. Dissecting molecular evolution in the highly diverse plant clade Caryophyllales using transcriptome sequencing. *Molecular Biology and Evolution* 32: 2001–2014.

Yang YW, Lai KN, Tai PY, Ma DP, Li WH. 1999. Molecular phylogenetic studies of *Brassica*, *Rorippa*, *Arabidopsis* and allied genera based on the internal transcribed spacer region of 18S-25S rDNA. *Molecular Phylogenetics and Evolution* 13: 455–462.

Yang YW, Tai PY, Chen Y, Li WH. 2002. A study of the phylogeny of *Brassica rapa*, *B. nigra*, *Raphanus sativus*, and their related genera using noncoding regions of chloroplast DNA. *Molecular Phylogenetics and Evolution* 23: 268–275.

Yuan YW, Liu C, Marx HE, Olmstead RG. 2009. The pentatricopeptide repeat (PPR) gene family, a tremendous resource for plant phylogenetic studies. *New Phytologist* 182: 272–283.

Zeng L, Zhang Q, Sun R, Kong H, Zhang N, Ma H. 2014. Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. *Nature Communications* 5: 4956.

Zhang N, Zeng L, Shan H, Ma H. 2012. Highly conserved low-copy nuclear genes as effective markers for phylogenetic analyses in angiosperms. *New Phytologist* 195: 923–937.

Zhu Q, Ge S. 2005. Phylogenetic relationships among A-genome species of the genus *Oryza* revealed by intron sequences of four nuclear genes. *New Phytologist* 167: 249–265.

Zimmer EA, Wen J. 2013. Using nuclear gene data for plant phylogenetics: Progress and prospects. *Molecular Phylogenetics and Evolution* 66: 539–550.

Supplementary Material

The following supplementary material is available online for this article at http://onlinelibrary.wiley.com/doi/10.1111/jse.12204/suppinfo:

Fig. S1. Maximum likelihood tree of *SMC2* paralogs from representative Brassicaceae species, with those from *Cleome serrulata* Pursh and *Populus trichocarpa* Torr. & A. Gray ex Hook. as outgroups. Values above branches are maximum likelihood bootstrap support values. Numbers after species names represent different paralogs.

Fig. S2. Fifty percent maximum likelihood majority-rule consensus tree of Brassicaceae based on single-gene phylogeny of *MLH1*. Values above branches are maximum likelihood bootstrap proportions (BP). Only branches with >50 BP are displayed.

Fig. S3. Fifty percent maximum likelihood majority-rule consensus tree of Brassicaceae based on single-gene phylogeny of *SMC2*. Values above branches are maximum likelihood bootstrap proportions (BP). Only branches with >50 BP are displayed.

Fig. S4. Fifty percent maximum likelihood (ML) majority-rule consensus tree of Brassicaceae based on single-gene phylogeny of *MCM5*. Values above branches are ML bootstrap proportions (BP). Only branches with >50 BP are displayed.

Table S1. GenBank accession numbers of *MLH1*, *MCM5*, and *SMC2* gene sequences from extracted DNA samples and transcriptomes.

Table S2. Positions and lengths of highly variable intron regions in *Arabidopsis thaliana*. The corresponding sites in the alignment were deleted in order to reduce phylogenetic noise.

Table S3. Result of Modeltest for *MLH1*, *MCM5*, and *SMC2* showing the top three substitution models with highest probability.