Available online at www.sciencedirect.com
SCIENCE DIRECT
Molecular Phylogenetics and Evolution 39 (2006) 846–854
MOLECULAR PHYLOGENETICS AND EVOLUTION
www.elsevier.com/locate/ympev

Sequence divergence at the putative flowering time locus *COL1* in Brassicaceae

Oksana Shavorskaya [a], Ulf Lagercrantz [a,b,*]

[a] *Department of Plant Biology and Forest Genetics, Swedish University of Agricultural Sciences, Box 7080, S-750 07 Uppsala, Sweden*
[b] *Department of Evolutionary Functional Genomics, Evolutionary Biology Centre, Uppsala University, SE-752 36 Uppsala, Sweden*

Received 5 October 2005; revised 29 December 2005; accepted 9 January 2006
Available online 17 February 2006

Abstract

An insertion/deletion polymorphism (*Ind2*) in the *Brassica nigra CONSTANS LIKE 1* (*Bni COL1*) gene was previously found to be associated with variation in flowering time. In the present study we examine the inter-specific divergence of *COL1* in the family Brassicaceae. Analysis of codon substitution models did not reveal evidence of positive Darwinian selection, but comparisons of the *COL1* gene in different species revealed a surprising number of indels. A total of 24 indels were found in the 650 bp of the middle variable region of the gene. This high number of indels could reflect a lack of constraint on length of this region of the protein, or the effect of positive selection. The number of indels was close to that expected in non-coding DNA, but the indels were longer in *COL1* than those observed in non-coding regions. Reconstruction of indel evolution indicated that most indels resulted from deletions rather than insertions. The *Ind2* indel that has shown association with flowering time in *Brassica nigra* exhibited a remarkable distribution in the Brassicaceae family, indicating that the polymorphism may have persisted more than ten million years. Considering presumed historic populations sizes of Brassicaceae species, such a long persistence time seems unlikely for a neutral polymorphism.

Keywords: Brassica; *COL1*; Flowering time; Molecular evolution; Indel

1. Introduction

Recent studies indicate that genes encoding plant transcription factors appear to diverge more rapidly among species than most other classes of genes (Arabidopsis Genome Initiative, 2000; van der Hoeven et al., 2002), suggesting that changes in gene regulation may have been an important force in plant evolution. It has also been noted that several transcription factors are often divided into domains that are characterized by slowly and rapidly evolving sequences (Henry and Damerval, 1997; Lagercrantz and Axelsson, 2000; Lukens and Doebley, 2001; Purugganan and Wessler, 1994; Purugganan et al., 1995; Rausher et al., 1999; Tucker and Lundrigan, 1993; Whitfield et al., 1993). However, it has not been clarified if the changes in the rapidly evolving domains are due mainly to relaxed constraint, or enhanced adaptive evolution.

The molecular evolution of genes in the CONSTANS LIKE (*COL*) gene family is characterised by high and heterogeneous rates of evolution in different regions. The *Brassica nigra CONSTANS LIKE 1* (Bni *COL1*) gene was identified as a candidate gene for the control of natural variation in flowering time (Kruskopf Österberg et al., 2002). Specifically, an insertion/deletion polymorphism (*Ind2*) in the coding region of Bni *COL1* was associated with flowering time in several populations. Analysis of nucleotide polymorphism at Bni *COL1* failed to detect evidence for selection (Lagercrantz et al., 2002). However, analysis of nucleotide variation showed that the S and L alleles were significantly diverged ($K_{ST} = 0.17$) and formed the two major clades of Bni *COL1* alleles. The divergence between the S and L alleles was most pronounced around the indel and the 3′ end of

* Corresponding author.Fax: +46 18 471 64 24.
E-mail address: Ulf.Lagercrantz@ebc.uu.se (U. Lagercrantz).

1055-7903/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2006.01.013

the gene, suggesting that the Ind2 polymorphism could be old.

Positive Darwinian selection promoting non-synonymous nucleotide substitutions has now been reported for a considerable number of genes (Yang, 2002). However, selection could potentially also act on insertion/deletion mutations. One such case was recently reported for a primate sperm ion channel gene (CATSPER1; Podlaha and Zhang, 2003). The association between allelic variation at *Ind2* and flowering time, and the apparently unusual distribution of *Ind2* in the Brassicaceae family prompted us to perform a more detailed study of the evolution of *COL1* in the Brassicaceae family. Thus, we studied the molecular evolution of both nucleotide substitutions and indels, with particular emphasis on *Ind2*.

2. Materials and methods

2.1. Plant materials and molecular methods

The species of Brassicaceae studied are given in Table 1. Genomic DNA was prepared from leaf tissue as described by Liscum and Oeller (1997). Genomic fragments of *B. nigra COL1* were amplified using primers CO 54 (GAGAGTA ACTGGGCACAAGC) + CO55 (TCTTCTTCTTCTCTC TGTATCTCA) for *A. lyrata* and *Brassica vulgaris*, CO 124 (CCATCAAACAACTACAACATCTGCG) + CO 125 (TTGAACTTAGGCAAGGTGAGTTACG) for *Arabidopsis thaliana*, and CO 16 (TAGCCTTTCTCTCC ATTG) + CO 39 (CTGCCGAGCTGATTCTGC) for all other species.

Amplified products were treated with ExoSAP-IT (USB) and both strands were sequenced directly on an ABI 377 sequencer. PCR products from heterozygous individuals were also cloned, and at least two clones of both alleles sequenced.

2.2. Phylogenetic and evolutionary analyses

DNA sequences were aligned using CLUSTAL W (Thompson et al., 1994) and refined manually, considering also deduced amino acid sequences (see Supplementary Material). Phylogenetic relationships were inferred using maximum-parsimony (MP) and Neighbor-joining (NJ) based on the Kimura two-parameter model or a maximum-likelihood model derived using Modeltest version 3.06 (Posada and Crandall, 1998). The analyses were performed in PAUP version 4.0.0b10 (Swofford, 2002). The MP searches used heuristic search strategies, with starting trees obtained via random stepwise addition and tree-bisection-reconnection branch swapping. Support for nodes was assessed using the bootstrap method with 1000 replicates for NJ and 500 for MP. Evolution of *COL* indels was reconstructed using parsimony in MacClade (Maddison and Maddison, 1992). GapCoder (Young and Healy, 2003) was used to code indels, and simple sequence repeats were excluded.

The Macintosh version of PAML (Yang, 1997) was used to compare models of codon sequence evolution in *COL1*. To estimate whether the amino acid substitutions are under positive selection, we conducted a likelihood analysis with different codon substitution models. These models make different assumptions about the $\omega = d_N/d_S$ ratio where d_S and d_N are the number of synonymous and non-synonymous nucleotide substitutions per site between each pair of sequences, respectively. The codon substitution models included the one ratio model (M0; one $\omega = d_N/d_S$ ratio for all sites and branches in the tree), a free ratio model where ω is allowed to vary over branches in three tree, and various models where ω is allowed to vary over sites (Yang et al., 2000). The site models included a neutral model (M1; with two classes of sites with $\omega = 0$ and 1), and a selection model (M2; which adds a third class with ω estimated from the data), a discrete model (M3; with

Table 1
Taxa sampled

Taxon	Accession/origin (Number of sequenced alleles)	Number of sequences/haplotypes		
		COL1	*CO*	*COL2*
Arabidopsis thaliana	CS1594(1), CS1550(1), CS1382(1), CS1246(1), CS1136(1), CS1106(1), CS1096(1), CS1094(1), CS1092(1), CS1066(1), CS1062(1), CS1564(1)	12/2	1/1	1/1
Arabidopsis lyrata	Sweden (1)	1/1		
Barbarea vulgaris	France (3), Netherlands (3)	6/1		
Brassica elongata	Uzbekistan (1)	1/1		
Brassica rapa	var.*pekinesis* (1), var.*chinensis* (1), var.*oleifera* (6)	8/8		1/1
Sinapis alba	Finland (1), Germany (1), Korea (1)	3/3		
Raphanus sativus	var. *sativus* (3); var. *niger* (1)	4/3		
Brassica oleracea	var. *acephala* (1), var. *alboglabra* (1)	2/2		
Brassica fructiculosa	Spain (1)	1/1		
Brassica nigra	Rapid cycling (1), Italy (1)	2/2	1/1	
Brassica juncea	Rapid cycling (1), China (1)	2/2		
Brassica napus			2/2[a]	
Total		42/26	1/1	1/1

[a] Derived from orthologous loci in the *B. oleracea* and *B. rapa* genome components of *B. napus* (Robert et al., 1998).

two classes with ω_0 and ω_1 estimated). In addition, we tried a beta model (M7; in which a beta distribution is assumed for ω) and $\beta + \omega$ (M8; where an extra class, with ω estimated, is added to the β model). The *COL1* NJ tree was used in all analyses. The transition:transversion ratio and the nucleotide frequencies at each codon position were estimated from the data, and codons in indels were removed (cleandata = 1). Contingency test for independence of mutational categories (the McDonald–Kreitman test; McDonald and Kreitman, 1991), was conducted using Fisher's exact test. The Hudson–Kreitman–Aguade (HKA) multilocus tests (Hudson et al., 1987) were conducted using the HKA program available from J. Hey.

3. Results

PCR primers complementary to the *COL1* coding sequence were used to amplify partial coding sequences from a range of Brassicaceae species (Table 1). Species were mainly sampled from the tribe Brassiceae, but with species from Arabidopsis and Barbarea as outgroups. The region amplified (ranging from 636 to 916 bp) contained around 80% of the coding region. Attempts to amplify the whole coding sequence failed for several species, probably due to the rapid evolution of regions outside the two conserved domains of *COL1* (Lagercrantz and Axelsson, 2000). We obtained *COL1* sequence from a total of 42 plants, representing 27 different haplotypes (Table 1). The phylogenetic relationships between the 27 haplotypes were inferred by neighbor-joining (NJ) and maximum parsimony (MP) analyses, using indels as separate characters (coded as 0 or 1). NJ and MP yielded similar trees, with identical major clustering. Exclusion of indels in the phylogenetic reconstructions did not change the topology of the trees, but bootstrap support values were generally slightly lower. The NJ tree (Fig. 1), shows one clade with relatively strong support that includes *Brassica oleracea* and *Brassica rapa*, and a sister clade with lower support containing *B. nigra*, *Sinapus alba* and *Rhapanus sativus*. This tree shows good agreement with data from other nuclear genes (Warwick and Sauder, 2005; Yang et al., 1998, 1999a,b). The *CO* and *COL2* genes were included in the analysis as outgroups to root the *COL1* tree.

Species in the Brassicaceae contain replicated genomes (Lagercrantz, 1998; Lagercrantz et al., 1996), indicating that multiple copies of *COL1* might be present in some species. Cloning in *B. nigra* has only identified a single *COL1* gene, and PCR amplifications in the present study did not reveal multiple *COL1* genes in any species. These data, and the close correspondence between the *COL1* phylogeny and phylogenies based on other genes, suggest that the *COL1* sequences in the present study are derived from orthologous genes. Two *COL1* copies were expected in *Brassica juncea* as it contains one *B. rapa* and one *B. nigra* genome (Axelsson et al., 2000), however only one from the *B. nigra* genome was amplified.

3.1. Evolution of insertions and deletions

A single intron is present in the *A. thaliana COL1* gene (Putterill et al., 1995). A corresponding intron is also present in the closely related *CO* gene both in *A. thaliana* and

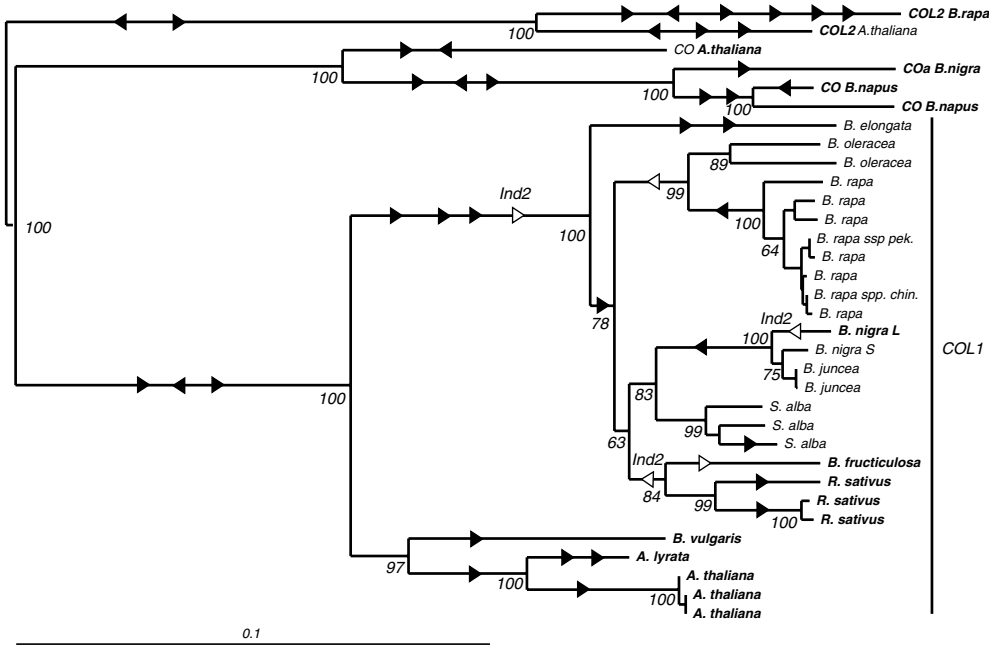

Fig. 1. Phylogenetic relationships of *CONSTANS LIKE 1* genes from Brassicaceae. The tree was constructed using neighbor-joining with Kimura 2 parameter distance, and indels coded as separate characters (coded as 0 or 1). Bootstrap support values are given for nodes with values above 60%. Species with the long allele at *Ind2* are shown in bold. Triangles on branches indicate inferred indel substitutions, a triangle pointing right indicate a deletion, and one pointing left an insertion. Open triangles denote indels that exhibit homoplasy or polymorphism. The *CO* and *COL2* genes were included as outgroups for *COL1*.

B. nigra. A homologous putative intron sequence is present in *COL1* from *Arabidopsis lyrata* and *B. vulgaris*, but absent in all other *COL1* genes sequenced. Sequencing of RT-PCR products of the *B. nigra COL1* gene showed that the intron is absent from this gene. From these data we conclude that the intron in *COL1* was lost in the lineage leading to Brassica, Raphanus, and Sinapis.

The general structure of the *COL1* gene is similar to other genes in the family (Griffiths et al., 2003; Lagercrantz and Axelsson, 2000) with two conserved domains (a pair of B-boxes and a CCT domain) flanking a variable region. Comparing *COL1* sequences from Brassicaceae species revealed that the variable region comprised a surprisingly high number of indels (Fig. 2). A total of 24 indels was found in the 650 bp of the variable *COL1* region sequenced in the present study. The indels ranged from three to 51 nucleotides—all were in multiples of three, thus preserving the correct reading frame. Two microsatellite-like trinucleotide repeats were also present in the coding region, and due to difficulties in alignment these were counted as a single indel each when calculating the total number of indels. The multitude of indels and rapid rate of nucleotide substitutions introduce some ambiguity in the alignment of *COL1* sequences. However, alternative alignments resulted in a similar number of indels, and did not change our conclusions.

Alignment of *COL1* sequences with those of the closely related genes *CO* and *COL2*, showed that indels have occurred frequently also in the evolution of these genes. To trace the evolution of the indels, the positions of the insertion/deletion events on the tree were inferred using parsimony. Non-microsatellite indels with unambiguous positions are shown on the branches of the tree in Fig. 1. Most indel events were unique, only two out of 41 indels with unambiguous position showed homoplasy. One was a deletion shared between *B. oleracea*, *B. rapa*, and *Brassica fructiculosa*, the other was *Ind2* that was polymorphic in *B. nigra* (see below). A majority of the inferred indel events were deletions. If we omit indels that are inferred on basal branches (as the position of the root of the tree is unknown) 15 deletions and 2 insertions are inferred in *COL1* ($\chi^2_1 = 9.9; P = 0.0016$), and for all three genes, 28 deletions and 7 insertions are inferred ($\chi^2_1 = 12.6; P = 0.0004$).

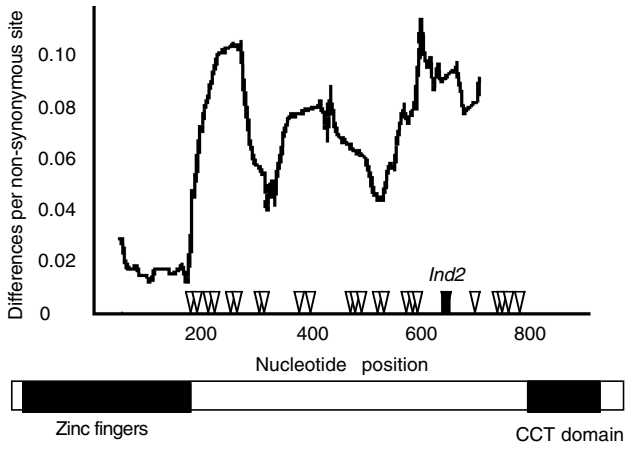

Fig. 2. Distribution of non-synonymous divergence along the *COL1* gene. The divergence was estimated as the average proportion of non-synonymous differences between genes applying Jukes and Cantor (1969) correction, and a sliding window of 50 sites. Triangles indicate indels.

3.2. Distribution of the Ind2 polymorphism

Two alleles differing by 18 nucleotides were identified at *Ind2*. The frequency of the short alleles ranged from 13 to 100% in different populations (Kruskopf Österberg et al., 2002). The variation at *Ind2* was correlated to flowering time in several populations. Individuals homozygous for short alleles flowered earlier than those carrying one or two copies of the long allele. In the sample of 42 *COL1* genes sequenced in the present study, the extra18 nucleotides representing the "long allele" at *Ind2* was present in 24 of those. The "long allele" was also present in the closely related *CONSTANS* gene as well as in *COL2* (Fig. 3). Thus, *Ind2* must have arisen through a deletion in *COL1*. All 12 *A. thaliana* sequences, all six *B. vulgaris* sequences, and the *A. lyrata* sequence contained the long allele, while a majority of genes from Brassica spp. and *S. alba* lacked the insertion. Consequently, a deletion most likely arose after the divergence of the lineage leading to Arabidopsis separated from the one leading to Brassica, Sinapis, and Raphanus. However, the insertion is also present in *R. sativus* and *B. fructiculosa* and as a polymorphism in *B. nigra*. Assuming that the deletion only occurred once and that the inferred phylogeny is correct, this implies that *Ind2* has been polymorphic since before the split of the lineage leading to *Brassica elongata* and the ones leading to Raphanus, Sinapis, *B. nigra* and *B. oleracea/rapa* (see discussion).

3.3. Tests for adaptive molecular evolution

It could be hypothesised that the stark divergence in the variable region could be an effect of positive selection resulting in novel protein function. To test for evidence of variation in the rate of non-synonymous to synonymous substitutions in *COL1* we used the codon substitution models developed by Yang and colleagues (Goldman and Yang, 1994; Yang and Nielsen, 2002). The estimated number of synonymous and non-synonymous substitutions in the *COL1* tree were 248 and 254, respectively. We compared the original model that assumes a single ω for all lineages and sites with models that account for variation among lineages, sites, and both (Yang and Nielsen, 2002). The model allowing ω to vary over branches did not result in a significantly better fit than the one-ratio model ($2\Delta l = 55.6$, $df = 51$, $P > 0.05$; Table 2). Ten out of 52 ω estimates were above one, but those were all on short branches. These results indicate no rate variation over lineages, and there is no evidence of adaptive evolution. Similarly, site-specific analysis detected variation

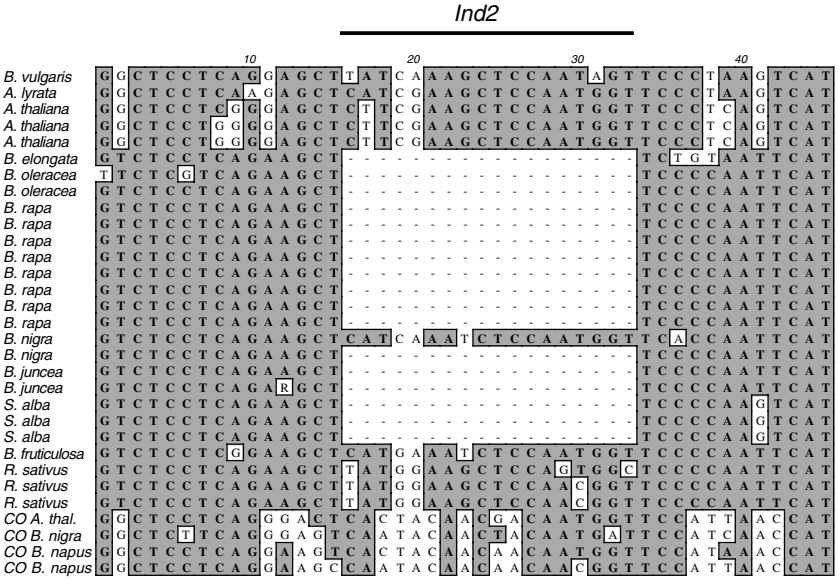

Fig. 3. Alignment of nucleotide sequences around the *Ind2* indel from *COL1* and *CO* genes. Grey boxed areas indicate nucleotides that are conserved in a majority of the genes.

Table 2
Parameter estimates for different codon substitution models

Model	*P*	*l*	Parameter estimates
M0: one-ratio	1	−3440.4	$\omega = 0.38$
Mfree: 1 ratio/branch	52	−3412.6	
M1: neutral	1	−3422.3	$p_0 = 0.43$, $p_1 = 0.57$
M2: selection	3	−3411.5	$p_0 = 0.00$, $p_1 = 0.34$, $p_2 = 0.65$, $\omega_2 = 0.12$
M3: discrete ($K = 2$)	3		$p_0 = 0.39$, $p_1 = 0.61$, $\omega_0 = 0.067$, $\omega_1 = 0.67$
M3:discrete ($K = 3$)	5	−3411.5	$p_0 = 0.65$, $p_1 = 0.02$, $p_2 = 0.32$, $\omega_0 = 0.12$, $\omega_1 = 0.98$, $\omega_2 = 0.98$
M7: β	2	−3413.0	$p = 0.33$, $q = 0.49$
M8: $\beta+\omega$	4	−3412.7	$p_0 = 0.99$, $p = 0.34$, $q = 0.50$, $p_1 = 0.01$, $\omega = 17.6$

See text for details.

among sites, but none of the tests suggested any sites affected by positive selection (Table 2). Model M1 (neutral) assumes two site classes with $\omega_0 = 0$ and $\omega_1 = 1$ with estimated proportions p_0 and p_1. This model can be compared with M2 (selection) that adds a third class with ω_2 estimated. $2\Delta l = 21.6$ and $P < 0.001$ with $df = 2$. M2 fits the data significantly better, but the estimate of ω_2 is less than one (0.12). Similarly, M3 fits the data better than M0 ($2\Delta l = 57.8$, $P < 0.001$, $df = 4$), but again none of the estimated ω parameters were larger than one (max 0.98). Model M7 that assumes a beta distribution for ω over sites fits the data comparatively well, and adding an additional site class with ω estimated from the data (model M8) did not improve the fit.

To allow further tests of non-neutral evolution of *COL1*, we sequenced *A. thaliana COL1* alleles from 12 ecotypes. The data revealed a low degree of polymorphism with only three segregating sites in 1446 nt and $\pi = 0.0008$. This is in the lower range of what has been observed in *A. thaliana* (Olsen et al., 2002; Yoshida et al., 2003). However, comparing intra-specific polymorphism with inter-specific divergence to *A. lyrata* in a multilocus HKA test (using four control genes: *AP1*, *AP3*, *CAL*, *PI*; Olsen et al., 2002) did not suggest a deviation from neutral expectations for *COL1* ($P = 0.55$). Similarly, comparing replacement to silent substitutions in the MacDonald–Kreitman test did not reveal any effect of natural selection on *COL1*. Finally, previous studies on a sample of 41 complete sequences of *B. nigra COL1* failed to detect departure form the standard neutral model or evidence of selection (Lagercrantz et al., 2002). In conclusion, we found no evidence for positive selection on nucleotide variation at *COL1*.

4. Discussion

Comparison of codon substitution models revealed no heterogenetity of ω over lineages, but showed that the rates of evolution were heterogeneous over sites in the COL1 protein. None of the tested models indicated sites with ω above one, but a large fraction of sites showed a high ω. This suggests that purifying selection is relaxed in particular in the variable middle region of *COL1*, and that positive selection has not played a major role in shaping the

diversity in the *COL1* gene. The results do not exclude that a small number of codons are subjected to divergent selection, but go undetected due to limited power. Several small regions of conservation have been identified in the variable middle region of gene in the *CONSTANS LIKE* family (Griffiths et al., 2003), indicating that at least parts of the variable region have important functions.

4.1. Indel evolution

The rapid non-synonymous substitution rate in the variable region of *COL1* is accompanied by an unusually high frequency of indels in the same region. On average, more than one indel every 30 bp was found in this coding region. Reconstruction of indel evolution in *COL* genes indicated that most of the indels have arisen from deletions rather than insertions. A more rapid accumulation of small deletions than small insertions is also found in mammalian genomes, both in coding (Taylor et al., 2004) and non-coding DNA (Ophir and Graur, 1997; Waterston et al., 2002; Zhang and Gerstein, 2003). However, in mouse and rat, the effect is diminished by selection in coding regions, indicating that insertions are actually more tolerable in protein-coding sequence than deletions (Taylor et al., 2004). If this holds true also for plants, we would expect a large excess of deletion over insertion mutations in plant DNA. Several studies have shown that plant genomes can increase rapidly in size primarily due to polyploidisation (Otto and Whitton, 2000; Wendel, 2000), and massive accumulation of retroelements (Bennetzen, 2000; Yoshida et al., 2003). These observations have led to the suggestion that genome size evolution may be largely unidirectional in plants (Bennetzen and Kellogg, 1997). However, recent data indicate that repeated increase and decrease in genome size may be common in plant evolution, and that genome size contraction may is some cases exceed that of expansion (Wendel et al., 2002). Very little is currently known of the dynamics of different genomic components. The presents study indicates that one type of event resulting in genome contraction in plants could be small deletions, occurring even in coding regions of functional genes.

Data on indel evolution in plants is still limited. Boivin et al. (2004) estimated indel frequencies when comparing a set of random, short *Capsella rubella* sequences with the corresponding sequences in *A. thaliana*. In exons, approximately 1 indel per 500 bp was estimated, and the corresponding estimates from intergenic regions and intron were around 1 indel per 40 bp. If we assume that indels in non-coding sequences are neutral, these data can be use to calculate the neutral rate of indel substitutions. It can then be used to test if the rate of indel substitutions in the *COL1* variable region is higher than neutral expectations.

The Capsella and Arabidopsis lineages diverged around 12 MYA (Koch et al., 2001). Using this divergence time yielded a neutral indel substitution rate of 1.0×10^{-9} per bp per year for the Arabidopsis–Capsella data. In comparison with coding regions it is more appropriate to calculate the substitution rate of 3n indels, as those could potentially be neutral in coding regions. Using only 3n indels (92 3n indels in 19.8 kb or 0.0046 indels per bp, R. Schmidt pers. com.) gave an estimated indel rate of 1.94×10^{-10} per bp per year. Both Arabidopsis–Capsella estimates are several times higher than the corresponding rates estimated from both human–chimpanzee and human–baboon data (Podlaha and Zhang, 2003). A high indel substitution rate in plant is also supported by data from *A. thaliana* ecotypes (Britten et al., 2003). Interestingly, in a majority of the genes that have been identified so far in Arabidopsis accounting for natural variation for various traits, this variation was due to indels (Koornneef et al., 2004).

Is the number of 3n indels in *COL1* significantly higher that the expected number, indicating selection for indel substitutions? To calculate the expected number of 3n indels in non-coding sequences between Arabidopsis and Brassica, we calculated the relative divergence times for Arabidopsis–Brassica versus Arabidopsis–Capsella. Using four different genes (18S-ITS, Yang et al., 1999b; nad4., Yang et al., 1999a; matK and Chs, Koch et al., 2001), we obtain a ratio of 1.59 ± 0.05 (mean $\pm$ SE). Using this ratio and the estimated 3n indel frequency from the Arabidopsis–Capsella comparison, the expected number of 3n indels in the *COL1* variable region between Arabidopsis–Brassica is $0.0046 * 1.59 * 650 = 4.8$. The average observed number of 3n indels between these lineages was 7.7, which is not significantly higher that the expected number (Poisson test, $P < 0.12$). It should be noted that we excluded simple sequence repeats from the calculation of indels in *COL1*, which may have biased the *COL1* estimate downwards as compared to the data from non-coding DNA.

Still, the size of indels in *COL1* is generally longer than those found in non-coding DNA comparing Arabidopsis and Capsella (Fig. 4). The two distributions are significantly different ($\chi^2 = 19.2$, df $= 4$, $P = 9.8 \times 10^{-4}$, indels $\geqslant 15$ bp were combined for the test), indicating that substitution of longer indels might be selected for in *COL1*. Due to the limited amount of indel data from non-coding regions, these results should be viewed as preliminary.

4.2. The Ind2 indel polymorphism is probably old

The distribution of Ind2 alleles in extant species (Fig. 1) suggests that *Ind2* might have been polymorphic for a long time. If we assume initially that *Ind2* arose by a single deletion event and the topology of the tree is correct, *Ind2* must have been polymorphic already when *B. elongata* separated from the Brassica/Raphanus/Sinapis lineage. We know that the "long allele" of *Ind2* is present in *B. nigra* and *R. sativus*. However, we have not sequenced enough alleles from each species to determine if that allele is absent in species such as *B. oleracea*, *B. rapa*, and *S. alba*. Still, it is enough to deduce that *Ind2* is either fixed for the "short allele" or segregating for the two alleles in these species.

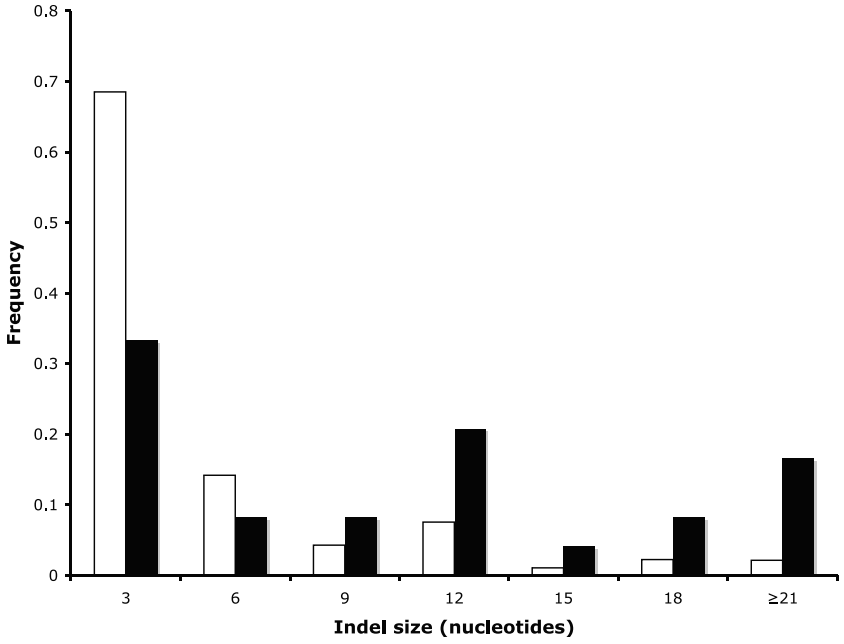

Fig. 4. Size distribution of 3n indels. White bars show 3n indels of non-coding genomic data from a comparison of Arabdopsis and Capsella (R. Schmidt pers. com.), and black bars represent 3n indel data from *COL1* sequences from in the present study.

In either case, the polymorphism must have been present already before the split of *B. elongata* from lineage leading to *B. nigra*/Raphanus/Sinapis and *B. oleracea*/*B. rapa*. Using a divergence time between the tribe Brassica and *A. thaliana* of 19 Myr (Koch et al., 2001), yields a substitution rate of 8×10^{-9} synonymous substitutions per site per year for *COL1*, which is similar to estimates of other plant nuclear genes ($5–30 \times 10^{-9}$; Gaut et al., 1996). This in turn results in an estimated divergence time between the *B. oleracea*/*B. rapa* lineage and the *B. nigra*/Raphanus/Sinapis lineage of 11 Myr. Thus, the polymorphism could be more than 10 Myr old.

To hold this conclusion requires that (i) the deletion occurred only once in the evolution of *COL1*, and (ii) that the *Ind2* polymorphism in *B. nigra* is not the result of a recent hybridisation between *B. nigra* and a species without the deletion such as *R. sativus*. Several independent deletions of the same 18 bp in different alleles seem a very unlikely event. Deletion hotspots comprising a few bases have been identified in humans and mouse (Kikkawa et al., 2003, and references therein). Two main mechanisms have been proposed, slippage of DNA polymerase and slipped mispairing (Jouanguy et al., 1969). Interactions between specific sequences (e.g., direct repeats, inverted repeats, and palindromes) are thought to cause such deletions. We see no indications of a deletion hotspot at *Ind2*. A recent hybridisation between *B. nigra* and a species with a long allele at *Ind2* is not supported by the sequence data either. The long *B. nigra Ind2* allele clusters phylogenetically with the short *B. nigra* and *B. juncea* alleles (*B. juncea* contains one *B. nigra* genome, Axelsson et al., 2000), while the *R. sativus* and *B. fructiculosa* with long *Ind2* alleles cluster together. Furthermore, detailed examination of nucleotide variation around *Ind2* does not support introgression of a long allele from one species into the short allele of another (or vice versa). Nucleotide variation within 20 bp upstream and downstream of *Ind2* supports the phylogenetic clustering of the alleles seen in Fig. 1.

An incorrect topology in the phylogenetic reconstruction could also affect the conclusion on the age of the *Ind2* polymorphism. However, all topologies suggests an old polymorphism or require several deletion events to explain the patterns of *Ind2* in extant species. In addition, the topology in the present study in supported by data from other nuclear genes (5S rRNA spacer and the 18S–25S rDNA ITS, Warwick and Sauder, 2005; Yang et al., 1998, 1999a).

Old putatively neutral polymorphisms in the form of trans-species polymorphisms have been reported in cichlid fish (Nagl et al., 1998). Several polymorphisms were estimated to be between 12,000 and 2 Myr old, and it was suggested that the long persistence time was a consequence of a large population size. If our assumption that *Ind2* arose only once is correct, it has been present for an even longer time (estimated to more than 8 Myr). A neutral mutation needs on average $4N_e$ generations to achieve fixation. Although the fixation time exhibits a large variance, a persistence of a neutral polymorphism for 8 Myr in *B. nigra* does not seem likely. Although past population size of *Brassica* species are largely unknown, estimates of nucleotide variation are relatively low (Lagercrantz et al., 2002; P. Sjödin, U. Lagercrantz, and M. Lascoux, unpubl. res.), compared to nucleotide variation in other plant species suggesting a small long term effective population size.

Acknowledgments

We thank the Swedish Research Council and the Swedish Research Council for Environment, Agricultural Sciences, and Spatial Planning (FORMAS) for support.

References

ARABIDOPSIS GENOME INITIATIVE, 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408: 796–815.

Axelsson, T., Bowman, C.M., Sharpe, A.G., Lydiate, D.J., Lagercrantz, U., 2000. Amphidiploid *Brassica juncea* contains conserved progenitor genomes. Genome 43, 679–688.

Bennetzen, J.L., 2000. Transposable element contributions to plant gene and genome evolution. Plant Mol. Biol. 42, 251–269.

Bennetzen, J.L., Kellogg, E.A., 1997. Do plants have a one-way ticket to genomic obesity? Plant Cell 9, 1509–1514.

Boivin, K., Acarkan, A., Mbulu, R.S., Clarenz, O., Schmidt, R., 2004. The Arabidopsis genome sequence as a tool for genome analysis in Brassicaceae. A comparison of the Arabidopsis and Capsella rubella genomes. Plant Physiol. 135, 735–744.

Britten, R.J., Rowen, L., Williams, J., Cameron, R.A., 2003. Majority of divergence between closely related DNA samples is due to indels. Proc. Natl. Acad. Sci. USA 100, 4661–4665.

Gaut, B.S., Morton, B.R., Mccaig, B.C., Clegg, M.T., 1996. Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL. Proc. Natl. Acad. Sci. USA 93, 10274–10279.

Goldman, N., Yang, Z.H., 1994. Codon-based model of nucleotide substitution for protein-coding DNA-sequences. Mol. Biol. Evol. 11, 725–736.

Griffiths, S., Dunford, R.P., Coupland, G., Laurie, D.A., 2003. The evolution of CONSTANS-like gene families in barley, rice, and Arabidopsis. Plant Physiol. 131, 1855–1867.

Henry, A.M., Damerval, C., 1997. High rates of polymorphism and recombination at the Opaque-2 locus in cultivated maize. Mol. Gen. Genet. 256, 147–157.

Hudson, R.R., Kreitman, M., Aguade, M., 1987. A test of neutral molecular evolution based on nucleotide data. Genetics 116, 153–159.

Jouanguy, E., Lamhamedi-cherradi, S., Lammas, D., Dorman, S.E., Jukes, M.C., Cantor, T.C., 1969. Evolution of protein molecules. In: Munro, H.N. (Ed.), Mammalian Protein Metabolism. Academic Press, New York, pp. 21–132.

Kikkawa, Y., Oyama, A., Ishii, R., Miura, I., Amano, T., et al., 2003. A small deletion hotspot in the type II keratin gene mK6irs1/Krt2-6g on mouse chromosome 15, a candidate for causing the wavy hair of the caracul (Ca) mutation. Genetics 165, 721–733.

Koch, M., Haubold, B., Mitchell-olds, T., 2001. Molecular systematics of the Brassicaceae: evidence from coding plastidic *matK* and nuclear *Chs* sequences. Am. J. Bot. 88, 534–544.

Koornneef, M., Alonso-blanco, C., Vreugdenhil, D., 2004. Naturally occurring genetic variation in *Arabidopsis thaliana*. Annu. Rev. Plant Biol. 55, 141–172.

Kruskopf Österberg, M.K., shavorskaya, O., Lascoux, M., Lagercrantz, U., 2002. Naturally occurring indel variation in the *Brassica nigra COL1* gene is associated with variation in flowering time. Genetics 161, 299–306.

Lagercrantz, U., 1998. Comparative mapping between *Arabidopsis thaliana* and *Brassica nigra* indicates that Brassica genomes have evolved through extensive genome replication accompanied by chromosome fusions and frequent rearrangements. Genetics 150, 1217–1228.

Lagercrantz, U., Axelsson, T., 2000. Rapid evolution of the family of CONSTANS LIKE genes in plants. Mol. Biol. Evol. 17, 1499–1507.

Lagercrantz, U., Kruskopf Österberg, M., Lascoux, M., 2002. Sequence variation and haplotype structure at the putative flowering-time locus COL1 of *Brassica nigra*. Mol. Biol. Evol. 19, 1474–1482.

Lagercrantz, U., Putterill, J., Coupland, G., Lydiate, D., 1996. Comparative mapping in Arabidopsis and Brassica, fine scale genome collinearity and congruence of genes controlling flowering time. Plant J. 9, 13–20.

Liscum, M., and Oeller, P., 1997. AFLP: not only for fingerprinting, but for positional cloning. <http://www-ciwdpb.stanford.edu/publications/methods/aflp.html/>.

Lukens, L., Doebley, J., 2001. Molecular evolution of the teosinte branched gene among maize and related grasses. Mol. Biol. Evol. 18, 627–638.

Maddison, W.P., Maddison, D.R., 1992. MacClade: Analysis of Phylogeny and Character Evolution. Sinauer, Sunderland MA.

McDonald, J.H., Kreitman, M., 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. Nature 351, 652–654.

Nagl, S., Tichy, H., Mayer, W.E., Takahata, N., Klein, J., 1998. Persistence of neutral polymorphisms in Lake Victoria cichlid fish. Proc. Natl. Acad. Sci. USA 95, 14238–14243.

Olsen, K.M., Womak, A., Garrett, A.R., Suddith, J.I., Purugganan, M.D., 2002. Contrasting evolutionary forces in the *Arabidopsis thaliana* floral developmental pathway. Genetics 160, 1641–1650.

Ophir, R., Graur, D., 1997. Patterns and rates of indel evolution in processed pseudogenes from humans and murids. Gene 205, 191–202.

Otto, S.P., Whitton, J., 2000. Polyploid incidence and evolution. Annu. Rev. Genet. 34, 401–437.

Podlaha, O., ZHANG, J., 2003. Positive selection on protein-length in the evolution of a primate sperm ion channel. Proc. Natl. Acad. Sci. USA 100, 12241–12246.

Posada, D., Crandall, K.A., 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14, 817–818.

Purugganan, M.D., Wessler, S.R., 1994. Molecular evolution of the plant R regulatory gene family. Genetics 138, 849–854.

Purugganan, M.D., Rounsley, S.D., Schmidt, R.J., Yanofsky, M.F., 1995. Molecular evolution of flower development: diversification of the plant MADS-box regulatory gene family. Genetics 140, 354–356.

Putterill, J., Robson, F., Lee, K., Simon, R., Coupland, G., 1995. The CONSTANS gene of Arabidopsis promotes flowering and encodes a protein showing similarities to zinc finger transcription factors. Cell 80, 847–857.

Rausher, M.D., Miller, R.E., Tiffin, P., 1999. Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway. Mol. Biol. Evol. 16, 266–274.

Robert, L.S., Robson, F., Sharpe, A., Lydiate, D., Coupland, G., 1998. Conserved structure and function of the Arabidopsis flowering time gene CONSTANS in Brassica napus. Plant Mol. Biol. 37, 763–772.

Swofford, D., 2002. *PAUP*. Phylogenetic analysis using parsimony (* and other methods*). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Taylor, M.S., Ponting, C.P., Copley, R.R., 2004. Occurrence and consequences of coding sequence insertions and deletions in Mammalian genomes. Genome Res. 14, 555–566.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties, and weight matrix choice. Nucleic Acids Res. 22, 4673–4680.

Tucker, P.K., Lundrigan, B.L., 1993. Rapid evolution of the sex determining locus in Old World mice and rats. Nature 364, 715–717.

van der Hoeven, R., Ronning, C., Giovanni, C., Martin, J., Tanksley, S., 2002. Deductions about the number, organization, and evolution of genes in the tomato genome based on analysis of a large expressed sequence tag collection and selective genomic sequencing. Plant Cell 14, 1441–1456.

Warwick, S.I., Sauder, C.A., 2005. Phylogeny of tribe Brassiceae (Brassicaceae) based on chloroplast restriction site polymorphisms and nuclear ribosomal internal transcribed spacer and chloroplast trnL intron sequences. Can. J. Bot. 83, 467–483.

Waterston, R.H., Lindblad-toh, K., BIRNEY, E., Rogers, J., Abril, J.F., et al., 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520–562.

Wendel, J.F., 2000. Genome evolution in polyploids. Plant Mol. Biol. 42, 225–249.

Wendel, J.F., Cronn, R.C., Johnston, J.S., Price, H.J., 2002. Feast and famine in plant genomes. Genetica 115, 37–47.

Whitfield, L.S., Lovell-badge, R., Goodfellow, P.N., 1993. Rapid sequence evolution of the mammalian sex-determining gene SRY. Nature 364, 713–715.

Yang, Y.W., Lai, K.N., Tai, P.Y., Li, W.H., 1999a. Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between Brassica and other angiosperm lineages. J. Mol. Evol. 48, 597–604.

Yang, Y.W., Lai, K.N., Tai, P.Y., MA, D.P., Li, W.H., 1999b. Molecular phylogenetic studies of Brassica, Rorippa, Arabidopsis, and Allied genera based on the internal transcribed spacer region of 18S–25S rDNA. Mol. Phylogenet. Evol. 13, 455–462.

Yang, Y.W., Tseng, P.F., TAI, P.Y., Chang, C.J., 1998. Phylogenetic position of Raphanus in relation to Brassica species based on 5S rRNA spacer sequence data. Bot. Bull. Acad. Sin. 39, 153–160.

Yang, Z., 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13, 555–556.

Yang, Z., 2002. Inference of selection from multiple species alignments. Curr. Opin. Genet. Dev. 12, 688–694.

Yang, Z., Nielsen, R., 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol. Biol. Evol. 19, 908–917.

Yang, Z., Nielsen, R., Goldman, N., Pedersen, A.M., 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155, 431–449.

Yoshida, K., Kamiya, T., Kawabe, A., Miyashita, N.T., 2003. DNA polymorphism at the *ACAULIS5* locus of the wild plant *Arabidopsis thaliana*. Genes Genet. Syst. 78, 11–21.

Young, N.D., Healy, J., 2003. GapCoder automates the use of indel characters in phylogenetic analysis. BMC Bioinformatics 4, 6.

Zhang, Z., Gerstein, M., 2003. Patterns of nucleotide substitution, insertion, and deletion in the human genome inferred from pseudogenes. Nucleic Acids Res. 31, 5338–5348.