Polymorphism and Divergence at Three Duplicate Genes in *Brassica nigra*

Per Sjödin · Harald Hedman · Marita Kruskopf Österberg · Susanne Gustafsson ·
Ulf Lagercrantz · Martin Lascoux

Received: 31 October 2006 / Accepted: 9 April 2008
© Springer Science+Business Media, LLC 2008

Abstract The *CONSTANS*-like gene family has been shown to evolve exceptionally fast in Brassicaceae. In the present study we analyzed sequence polymorphism and divergence of three genes from this family: *COL1* (*CONSTANS-LIKE 1*) and two copies of *CO* (*CONSTANS*), *COa* and *COb*, in *B. nigra*. There was a significant fourfold difference in overall nucleotide diversity among the three genes, with *BniCOb* having twice as much variation as *BniCOL1*, which in turn was twice as variable as *BniCOa*. The ratio of nonsynonymous-to-synonymous substitutions (dN/dS) was high for all three genes, confirming previous studies. While we did not detect evidence of selection at *BniCOa* and *BniCOb*, there was a significant excess of polymorphic synonymous mutations in a McDonald-Kreitman test comparing *COL1* in *B. nigra* and *A. thaliana*. This is apparently the result of an increase in selective constraint on *COL1* in *B. nigra* combined with a decrease in *A. thaliana*. In conclusion, a complex scenario involving

both demography and selection seems to have shaped the pattern of polymorphism at the three genes.

Keywords *Brassica nigra* · Duplicated genes · *COL1* · *COa* · *COb* · Flowering time

Introduction

The fate of duplicated genes has been the object of numerous empirical and theoretical studies since Ohno suggested that most new genes originate through duplication (Ohno 1970). Ohno's idea was that while one copy preserves the original function, the other, not being subjected to selection, is free to evolve a new function. Early theoretical studies established that most duplicated genes should become nonfunctional pseudogenes, as deleterious mutations will quickly accumulate (Haldane 1933; Fisher 1935). Yet, genomic studies showed that functional duplicated genes are very common (e.g., Lynch and Conery 2000). How, then, could two copies be maintained in the face of deleterious mutations? Most theories that try to explain the preservation of duplicated copies involve some amount of symmetry breaking. For instance, the so-called DDC model (duplication, degeneration, and complementation) proposes that duplicates are preserved through subfunctionalization whereby different subfunctions of the original gene are partitioned between the two copies (e.g., Force et al. 1999). Thus both copies are essential to perform the original function. The DDC model has recently gained support, for instance, in yeast (van Hoof 2005), and subfunctionalization may, more generally, be an important factor, especially in developmental genes, though the part of the gene that is partitioned is the regulatory region and not the coding region (Zhang 2003; Averof 2002 and

Sequence data were deposited in the GenBank database under accession nos. EU693946-EU694094.

P. Sjödin · H. Hedman · M. Kruskopf Österberg · S. Gustafsson ·
U. Lagercrantz · M. Lascoux (✉)
Department of Evolutionary Functional Genomics, Evolutionary
Biology Center, Uppsala University, Norbyvägen 18D,
752 36 Uppsala, Sweden
e-mail: Martin.Lascoux@ebc.uu.se

H. Hedman
Department of Plant Biology, Swedish University of Agricultural
Sciences, Box 7080, 750 07 Uppsala, Sweden

Present Address:

P. Sjödin
Bioinformatics Research Center, University of Aarhus, Building
1090, Hoegh Guldbergsgade 10, 8000 Aarhus C, Denmark

references therein). Other factors that can explain the retention of duplicates include concerted evolution (Arnheim 1983), acting mostly through gene conversion (Innan 2003), or selection for increased expression (Schrantz et al. 2002), or the opposite, where a gene is expressed under a very short but essential period of time. In addition, it has been argued that paralogous genes encoding multidomain proteins may be retained to mask the effects of dominant negative mutations (Gibson and Spring 1998).

CONSTANS (CO) in the model organism *Arabidopsis thaliana* is a member of a gene family with a zinc finger domain at the 5' end and a CCT (CONSTANS, CONSTANS-LIKE, TIMING OF CAB-EXPRESSION 1) domain at the 3' end (Putterill et al. 1995). The *CO* gene is an important regulatory gene in the flowering time pathway of *A. thaliana* and has been shown to promote early flowering in response to long day length (Samach et al. 2000). At least in plants, the family of *CONSTANS-LIKE* genes evolves exceptionally fast, even surpassing genes involved in resistance to parasites (Lagercrantz and Axelsson 2000). The black mustard, *Brassica nigra* ($2n = 16$), is an outcrossing annual, closely related to *Arabidopsis thaliana*. Together with other *Brassica* species, it likely descends from a hexaploid ancestor, followed by extensive rearrangements, making its genome essentially a triplicated *A. thaliana* genome (Lagercrantz 1998; Parkin et al. 2005). This implies that for every gene found in *A. thaliana*, there are potentially three copies in *B. nigra*. In a previous study, we identified a major QTL for flowering time in *Brassica nigra* located in the chromosomal region where *CO* (*BniCOa*) is situated and a minor QTL in a genomic region harboring another *CO* copy, *BniCOb* (Lagercrantz et al. 1996). However, variation in flowering time was actually associated with an indel polymorphism in an adjacent gene, *BniCOLI*, while there was practically no variation at all in *BniCOa* (Kruskopf-Österberg et al. 2002). *BniCOLI* is homologous to *AtCOLI* (*CONSTANS-LIKE 1* [Putterill et al. 1997]). *AtCOLI* is closely related to *CO* and is located about 3.5 kb upstream of *CO* in both *A. thaliana* and *B. nigra*. Unlike *AtCO*, altered expression of *AtCOLI* in transgenic plants had little effect on flowering time (Ledger et al. 2001). A later study, carried out under slightly different environmental conditions and incorporating a larger number of populations, retrieved a significant association between indel variation in *BniCOLI* and flowering time variation in some populations but not in others, suggesting that the important nucleotide variation for flowering time might not be the indel in *BniCOLI* itself but rather a polymorphism in linkage disequilibrium with it. Preliminary data indeed suggest that the intergenic space between *BniCOLI* and *BniCOa* could harbor the causal polymorphism (Krouchi et al., in press).

Importantly, the three duplicate genes, *BniCOLI*, *BniCOa*, and *BniCOb*, are not equivalent. First, they originated at different times (Shavorskaya and Lagercrantz 2006). While *BniCOa* and *BniCOb* diverged after the split between *Arabidopsis* and *Brassica*, *COLI* and *CO* resulted from a duplication event predating the split between *Arabidopsis* and *Brassica*. Second, *COLI* and *CO* are closely located on the same chromosome. Generally, the molecular evolution of duplicate loci depends on their genomic location (Thornton and Long 2005). For instance, there is a tendency for duplicates in regions of low recombination in *S. cerevisiae* to exhibit higher rates of amino acid evolution than their paralogues in regions of high recombination (Zhang and Kishino 2004). So both age and colocation of the pairs of loci need to be considered when analyzing the data. In this respect it should also be added that although it is possible that more than one copy of *COLI* exists in *B. nigra*, we assume this not to be the case, as we were unable to find additional copies. In other words, the situation is one where we compare one pair of relatively recent duplicates, *BniCOa* and *BniCOb*, with its single-copy homologue in *A. thaliana* and the previously duplicated but today single-copy gene *BniCOLI* with its single-copy homologue *AtCOLI*, keeping in mind that *CO* and *COLI* are also duplicates, though older ones.

In the present study we examine sequence polymorphism and divergence at *BniCOa*, *BniCOLI*, and *BniCOb* to look more closely at the evolutionary history of these three putative flowering time genes. More specifically we addressed three main questions: (i) Is there stronger evidence of selection on *BniCOLI* than on *BniCOa* and *BniCOb*? (ii) Has selection been acting differently on *BniCOLI* than on *AtCOLI*? And (iii) Is the level of selective constraint on *BniCOb* consistent with *BniCOb* being a pseudogene, as suggested by a separate study (Sjödin et al. 2007)?

Materials and Methods

The Species

In temperate regions, *B. nigra* was a major mustard crop before the 1950s but has been progressively replaced by *B. juncea* and is today mainly a weed. However, it is still used as a condiment in India and Ethiopia. Even if *B. nigra* is not currently the object of extensive breeding, its use as a condiment certainly implies a proportionally strong human influence. Its current natural distribution ranges from the Atlantic Ocean eastward to India and from Southern Scandinavia southward to Ethiopia. It has also been introduced in North America. Its population genetic structure and recent history are still poorly documented, but

a previous study using nuclear microsatellites delineated three main groups – European, Indian, and Ethiopian (Westman and Kresovich 1999).

Population Samples

All entries were ex situ germplasm accessions. Seed samples originating from Greece (five plants; accession number BRA 2104), and France (five plants; accession number BRA 2140) were obtained from the Institut für Pflanzengenetik und Kulturpflanzenforschung, Gatersleben, Germany, while seed samples from Ethiopia (five plants; accession number PI 273641), India (five plants; accession number PI 175073), and Turkey (five plants; accession number PI 592737) were provided by USDA, ARS, NCRPIS, Iowa State University, USA. In most cases, the samples are from natural populations, although a feral origin cannot be ruled out. We used sequences from *A. thaliana* as outgroup, accession Y10556 for *BniCOa* and *BniCOb*, and accession Y10555 for *BniCOLI*. To construct the phylogenetic trees, we also used *CONSTANS* from *Cleome spinosa* (accession number DQ415922) as an outgroup. *CONSTANS* in Cleomaceae is an ancestor to both *CO* and *COLI* in Brassicaceae (Schranz and Mitchell-Olds 2006).

Molecular Methods

Genomic DNA was prepared from leaf samples as described by Liscum and Oeller (1997). For sequencing we amplified *BniCOa* using primers CO 101 (5'-CTAACTAACCTTGCTCCTTTTG-3') and CO 102 (5'-ACTTCGGTTTCACTCATCT TTG-3'), *BniCOb* using primers CO 116 (5'-CTCTCTATGGCAAGTGGA-3') and CO 128 (5'-CCCATATGCATAACTGTGGC-3'), and, finally, *BniCOLI* using primers CO105 (5'-GTTGATGGGTCCTACAGAGAGAGAG-3') and CO106 (5'-GGAGTATTA TGAGAATGAAGGAACAAT-3'). The amplified fragment containing *BniCOa* was approximately 1700 bp long, including a 445-bp-long 5'-UTR, the first exon, the intron, and 250 bp of the second exon. The *BniCOb* fragment consisted of 195 bp of 5'-UTR, the first exon, the intron, and 300 bp of the second exon, while *BniCOLI*, which lacks the intron, contained 280 bp of 5'-UTR and 1050 bp of the exon.

Sequence Alignment

The sequences were aligned using CLUSTAL X (Thompson et al. 1997) and then edited by hand with the constraint that indels within the coding region did not disrupt the reading frame. As this procedure is somewhat subjective, we tried different alignments and the results appear robust.

Alignments between genes and between species were only meaningful for the coding region.

Intraspecific Polymorphism Analysis

Level of nucleotide polymorphism was summarized by two statistics, Watterson's (1975) θ_W and Tajima's (1983) θ_π . We used the χ^2_C and χ^2_L goodness-of-fit tests (Kreitman and Hudson 1991) to compare the level of polymorphism in the noncoding regions of *BniCOa*, *BniCOb*, and *BniCOLI*. The population recombination rate, ρ , was estimated with Hudson's (1987) method. Population differentiation was estimated with K_{ST} (Hudson et al. 1992). Two test statistics of the standard neutral model were used, Tajima's (1989) D and Fay and Wu's (2000) H . Both test statistics compare two estimates of θ ; Tajima's D measures the standardized difference between θ_π and θ_W , while Fay and Wu's H measures the difference between θ_π and θ_H . The former is most sensitive to an excess of rare variants, whereas the latter is most sensitive to an excess of high-frequency-derived variants. Both D and H are expected to be close to zero under the standard neutral model. Both tests were carried out with and without recombination, as the lack of recombination makes the tests overly conservative. The McDonald-Kreitman (1991) test was used to test the neutral mutation hypothesis. We corrected for multiple hits using the Yang and Nielsen (2000) method implemented in PAML. To calculate the intraspecific dN/dS ratio we divided the number of segregating nonsynonymous sites, weighted by the approximated number of nonsynonymous sites, by the analogously weighted number of segregating synonymous sites. All statistics, except for the Kreitman-Hudson (1991) χ^2_C and χ^2_L goodness-of-fit tests, were calculated with the DnaSP version 4.10 software (Rozas et al. 2003). The program *ms* (Hudson 2002) was used for coalescent simulations.

Indel Variation Analysis

Most of the variation in our data is due to indels. This information is generally discarded and there is no consensual view on how to analyze indel variation. However, especially in plants, for which the number of sequence differences due to indels can be as much as 55 times more common than nucleotide differences (Britten et al. 2003), and, in particular, in the present case where indel variation in one of the genes under study has been associated with flowering time (Kruskopf-Österberg et al. 2002), discarding this information seems questionable.

Indel variation can be summarized by two quantities: the number of mutational events creating indels and the

number of sites affected by this variation. The two are not equivalent, and notably the former is more sensitive to alignment parameters than the latter. We tallied the number of indel events in the coding region and compared it to different categories of segregating sites in a 2×2 contingency table analogous to the McDonald-Kreitman test. The use of the number of sites affected by indels in a similar test is problematic, as can be illustrated by the following example. First, assume that 1 in 10 indel mutations is a very long deletion and that the rest are very short deletions or insertions. We also assume that the fixed and polymorphic categories have five indels each. It may very well be the case, then, that all fixed indels are very short but one of the polymorphic indels is a very long deletion. In this case all tests comparing the number of sites affected by indels to the number of segregating sites or number of indel events in the fixed and polymorphic categories are likely to be significant, although there is no real difference in indel mutation dynamics. Furthermore, we used *CONSTANS* from *C. spinosa* as an outgroup to orient the indels as insertions or deletions where they occur in the phylogeny. The microsatellites in *BniCOa* and *BniCOLI* or indels directly adjacent to those were not included in any of the above. Moreover, only indels in the coding region were considered.

Codon Substitution Analysis

Trees relating *CO* and *CO*-like genes were constructed from the aligned coding sequences using neighbor joining as implemented in MEGA version 3.1 (Kumar et al. 2004). The dN/dS ratios along the different branches were then estimated with PAML version 3.14 (Yang 1997). In all analyses we used one sequence of *BniCOa* (allele *COa_0212_2*), the homologous gene from *Brassica napus* (*BnaCOa*; accession AY280868), *AtCO* (accession Y10556), *BniCOLI* (allele *COLI_0210_2*), the homologous gene from *B. napus* (*BnaCOLI*; accession AY280868), *AtCOLI* (accession Y10555), and, also, *CONSTANS* from *Cleome spinosa* (accession DQ415922). Figure 3 shows the phylogenetic relationship between

these genes. Before the analysis, the microsatellite region, all noncoding regions, and all codons with alignment gaps were removed. Also, *BniCOb* was excluded, as it turned out not to be of special interest and including this sequence only increased the number of codons with alignment gaps. We used a maximum likelihood ratio test to compare a model with the same dN/dS ratio on all branches to models where, based on the results from the McDonald-Kreitman test, branches of interest were free to have a different dN/dS ratio. We also used PAML to calculate pairwise dN and dS (Yang and Nielsen 2000).

Results

Within-Species Polymorphism

There was a fourfold difference in overall nucleotide diversity among the three genes, with *BniCOb* having twice as much variation as *BniCOLI*, which in turn was twice as variable as *BniCOa*. This was also true when silent and nonsynonymous sites were considered separately (Table 1) as well as when only the seemingly nondegenerate alleles (those free of frameshift mutations and/or indels) of *BniCOb* were included.

To test whether the difference among loci could simply be the results of mutation and random genetic drift, we used Kreitman and Hudson (1991) χ^2_C and χ^2_L tests. χ^2_C and χ^2_L values were 15.05 and 30.87, respectively. The χ^2_L statistic is approximately χ^2 distributed assuming no linkage between genes but total linkage within genes, while the χ^2_C statistic approximately follows a χ^2 distribution when either all genes and all sites are completely linked or all sites (and genes) are completely independent (Kreitman and Hudson 1991). Since we compared three regions, the relevant number of degrees of freedom is 2, making both tests significant assuming a χ^2 distribution. However, the assumptions underlying the χ^2 approximation are clearly not met and we therefore also used coalescent simulations to test the hypothesis that difference among the three genes is simply due to mutation and random genetic drift.

Table 1 Overall polymorphism in the three genes

	<i>N</i>	<i>L</i>	All			Synonymous			Nonsynonymous			Synonymous & noncoding		
			<i>S</i>	θ_π	θ_W	<i>S</i>	θ_π	θ_W	<i>S</i>	θ_π	θ_W	<i>S</i>	θ_π	θ_W
<i>COLI</i>	50	1333	60	0.013	0.011	30	0.034	0.029	19	0.007	0.006	42	0.022	0.020
<i>COa</i>	50	1682	42	0.006	0.006	13	0.015	0.013	16	0.004	0.004	26	0.008	0.007
<i>COb</i>	48	1309	102	0.025	0.026	36	0.062	0.046	50	0.015	0.018	46	0.052	0.040

Note: *n* is the number of sequences, *L* is the length of the sequences, *S* is the number of segregating sites, θ_π is the nucleotide diversity, and θ_W is the Watterson (1975) estimate. Only sites without ambiguities were taken into account. Complex mutations were not considered. Hence $NS+(S+NC)$ does not sum up to the number of segregating sites

Table 2 Simulated values of the χ^2_C and χ^2_L test statistics under different values of the recombination parameter (ρ) and the two simulation procedures

ρ	Three genes		One gene	
	χ^2_C	χ^2_L	χ^2_C	χ^2_L
0	5.80	13.25	2.75	5.93
0.5	5.83	13.42	3.07	6.57
1.5	5.56	12.40	3.50	7.53
2.5	5.58	12.50	3.64	7.86
3.75	5.29	11.94	3.77	8.26
5	5.24	11.73	3.93	8.53
7.5	4.97	11.12	3.97	8.68
10	4.82	10.72	4.04	8.74
20	4.30	9.53	3.81	8.23
30	4.12	9.00	3.76	8.20
40	3.86	8.44	3.50	7.62
50	3.76	8.41	3.50	7.62
60	3.68	8.08	3.51	7.66
70	3.65	8.02	3.41	7.35

Note: “Three genes” considers three separate genes with the same lengths as the genes under study. “One gene” assumes a single locus with the same length as the sum of the three loci under study. The mutation parameter (θ) was chosen to obtain that the expected number of segregating sites was 55. Hence, in the three-gene simulation procedure, the sum of the three θ values was equal to the single θ value in the one-gene case. Likewise, for each ρ value, this was the sum of the three ρ values in the three-gene case. The observed test values were $\chi^2_C = 12.97$ and $\chi^2_L = 30.91$

Following Kreitman and Hudson (1991), we used *ms* (Hudson 2002) to generate samples under different recombination rates and calculated the χ^2_C and χ^2_L statistics for each sample. Apart from varying the recombination rate, the simulations were done under two different models. The first set of simulations considered three separate genes with the same lengths as our genes (242, 639, and 282 bp for the noncoding regions of *COLI*, *COa*, and *Cob*, respectively). The second set of simulations considered only one contiguous gene with a length of 1163 bp, i.e., the sum of the lengths of the three genes. For both sets of simulations and for each value of the recombination parameter, we simulated 10,000 samples of size 100. The θ value for all simulations was set to a value such that the expected total number of segregating sites was the same as in our data (55). The results of these simulations are reported in Table 2. It is clear that recombination is not the cause of the significance of our test values.

The estimated population recombination rate, $\hat{\rho}$, using Hudson’s (1987) method was 0.0129, 0.0080, and 0.0098 per site for *BniCOB*, *BniCOLI*, and *BniCOa*, respectively (Table 3). Population structure influences the estimate of ρ , and Hudson’s method overestimates it, but this should not be a matter of concern here, as we are

Table 3 Differentiation among populations for the different genes: K_{ST} (Hudson et al. 1992)

	India	Turkey	Ethiopia	Greece
<i>COLI</i>				
Turkey	0.203***			
Ethiopia	0.069	0.435***		
Greece	0.389***	0.317***	0.550***	
France	0.497***	0.396***	0.559***	0.230**
<i>COa</i>				
Turkey	0.298**			
Ethiopia	0.049	0.581***		
Greece	0.239***	0.089*	0.459***	
France	0.494***	0.538***	0.844***	0.419***
<i>Cob</i>				
Turkey	0.245**			
Ethiopia	−0.011	0.334**		
Greece	0.197**	0.206**	0.218***	
France	0.237***	0.153*	0.300**	0.318***

Note: Significance levels: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

primarily interested in the relative values. We found no evidence of gene conversion except for two short tracts (2 and 4 bp) within *BniCOB*. Populations were highly and significantly differentiated at the three loci, with overall K_{ST} values of 0.31, 0.49, and 0.51 for *BniCOB*, *BniCOLI*, and *BniCOa*, respectively. The lower K_{ST} value obtained for *BniCOB* could, in part, simply reflect the fact that this gene was the most variable one. *BniCOa* was fixed in Ethiopia. France, Greece, and Turkey were generally the most variable populations, and Ethiopia the least variable one (Table 4).

The standard neutral model could not be rejected when no recombination was assumed. When recombination was assumed, only Fay and Wu’s *H* was significant for *BniCOB* (Table 5). The sliding window option for Tajima’s *D* in DnaSP (window length, 100 bp; step size, 25 bp) showed an interesting trend. There is a general tendency for Tajima’s *D* to increase from 5’ to 3’ in *BniCOLI* but decrease in the same direction in *BniCOa* (Fig. 1). Finally, the intraspecific dN/dS ratio was lower for *BniCOLI*, dN/dS = 0.19, than for *BniCOa* and *BniCOB*, with values of 0.37 and 0.38, respectively, suggesting a higher selective constraint on the former than on the latter two.

McDonald-Kreitman Tests and Sizes of Indels

We carried out a standard McDonald-Kreitman test to compare polymorphism and divergence at nonsynonymous and synonymous mutations, and an analogous test to compare polymorphism and divergence for the number of indels and different categories of segregating sites. The

Table 4 Polymorphism within populations

	<i>n</i>	No. of haplotypes	<i>S</i>	θ_{π}	θ_W	Tajima's <i>D</i>
<i>BniCOa</i>						
France	10	4	6	0.00113	0.00128	−0.4959
Greece	10	7	28	0.00598	0.00612	−0.0902
Turkey	10	6	16	0.00405	0.00345	0.8092
India	10	4	21	0.00358	0.00458	−1.0168
Ethiopia	10	1	0	0	0	ND
<i>BniCOb</i>						
France	8	8	42	0.0168	0.0147	0.6206
Greece	10	6	61	0.0223	0.0197	0.6216
Turkey	10	10	48	0.0186	0.0158	0.8809
India	10	9	44	0.0154	0.0148	0.1199
Ethiopia	10	6	38	0.0122	0.0130	−0.2453
<i>BniCOLI</i>						
France	10	4	28	0.010	0.008	1.37
Greece	10	5	29	0.009	0.008	0.56
Turkey	10	6	26	0.008	0.007	0.20
India	10	5	18	0.004	0.005	−0.59
Ethiopia	10	2	2	0.000	0.001	−1.40

Note: *n*, number of sequences; θ_{π} , nucleotide diversity; θ_W , Watterson estimate of the scaled mutation rate

Table 5 Tests of the standard neutral model (*D*, *D**, and *H*)

	<i>BniCOLI</i>	<i>BniCOa</i>	<i>BniCOb</i>
Tajima's <i>D</i>	0.518	0.17	−0.33
Fu & Li's <i>D</i> *	1.296	−0.69	−0.32
Fay & Wu's <i>H</i>	−4.003	−7.03	−15.99*

Note: The population recombination rate used when calculating the *p* values is the one estimated for each gene. * *p* < 0.05. None of the tests was significant at *p* = 0.05 when no recombination was assumed

only significant McDonald-Kreitman test is the one comparing *BniCOLI*-*AtCOLI* (Table 6). By contrasting the three contingency tables, there appears to be an excess of polymorphic synonymous mutations in *BniCOLI*. None of the indel tests were significant (data not shown).

Bias in Deletions

Using *C. spinosa* as an outgroup, we were able to classify almost all indels in the coding regions as either insertions or deletions. Figure 2 shows their distribution on the phylogeny, where the number to the left of a branch represents deletions and the number to the right insertions. The majority of indels are deletions (26 of 33) and most of the indels occur on *B. nigra* branches, but no clear trend could be detected between *CO* and *COLI*. The species effect can be explained by the fact that *A. thaliana* is a selfer and thus might have a smaller effective population size than

Fig. 1 A sliding window representation of how Tajima's *D* varies along *BniCOLI* and *BniCOa*. The window length was 100 bp, and the step size 25 bp

B. nigra. It could also be a result of the triplicated nature of genes in *B. nigra*.

Codon Substitution Analysis

The motivation of this analysis was to try to understand the significant result of the McDonald-Kreitman test for *COLI*. In this test, the dN/dS ratio for polymorphic sites was much lower than the dN/dS ratio for fixed sites. However, it is not possible to tell from the McDonald-Kreitman test alone, compared to a background of selective constraint in this gene family, whether (i) there has been a recent increase in selective constraint in *B. nigra*, (ii) *COLI* went through a period of relaxed selective constraint in *Brassica* during the period when there was more than one copy of this gene, or (iii) *COLI* in *A. thaliana* is experiencing a relaxation of constraint. Thus, we are interested in the dN/dS ratios of (i) the external branch from *B. nigra* *COLI* to the *B. nigra*/*B. napus* split ('B1' in Fig. 3), (ii) the internal *COLI* branch from the *Arabidopsis*/*Brassica* split to the *B. nigra*/*B. napus* split ('B2' in Fig. 3), and (iii) the external branch from the *Arabidopsis*/*Brassica* split to *A. thaliana* ('B3' in Fig. 3). If the first hypothesis is true, we would then expect that a model with a lower dN/dS ratio on 'B1' is significantly more likely than a model where all branches have the same dN/dS ratio. Likewise, a model with a higher dN/dS ratio on 'B2' or 'B3' being significantly better than a one-ratio model will lend support to the second and third hypotheses, respectively. Note specifically that we are interested in one-sided model testing, but if a two-sided test is significant, so is the one-sided one. We used PAML to compare the one-ratio model to the three models presented above. Both the model with a different dN/dS ratio on the *COLI* branch to *A. thaliana* and the model with a different dN/dS ratio on the *COLI* branch to *B. nigra* were significantly better—in a two-sided likelihood ratio test—than the one-ratio model ($2\Delta\ln L = 5.41$, *df* = 1, and $2\Delta\ln L =$

Table 6 McDonald-Kreitman tests of the three genes excluding the microsatellite region

	<i>BniCOL1/AtCOL1</i>		<i>BniCOa/AtCO</i>		<i>BniCOB/AtCO</i>	
	Fixed	Polymorphic	Fixed	Polymorphic	Fixed	Polymorphic
Nonsynonymous	83	16	98	14	116	52
Synonymous	74	30	78	12	86	48
χ^2	4.66		0.03		0.80	

Note: The fixed segregating sites were corrected for multiple hits using the implementation of the Yang and Nielsen (2000) method in PAML. The number of intraspecific sequences used was 50 for *COL1* and *COa* but only 33 for *COB*, as some of the *COB* sequences contained a very long deletion in the beginning of the coding region and these were not used

Fig. 2 Phylogenetic tree with the number of insertion and deletion events on each branch. The number to the right of a branch represent insertions; the number to the left, deletions. Note that the ‘polymorphic’ part of the tree is a schematic representation and that branch lengths are not scaled

4.15, $df = 1$, respectively). The estimated dN/dS ratios on these branches were also in line with the expectations (higher and lower than the background, respectively). A model where both branches are allowed to be different from the background is also significantly better than the one-ratio model ($2\Delta\ln L = 8.71$, $df = 2$). Figure 3 shows the estimated dN/dS ratios from this model. The model with a different dN/dS ratio on ‘B2’ was not significantly better than the one-ratio model ($2\Delta\ln L = 0.02$, $df = 1$). The high dN/dS ratio on ‘B3’ can be explained either by a generally lower selective constraint or by selection. We used the branch-site model procedure described by Yang et al. (2005) to search for evidence of selection. Yang et al. (2005) group the codons into three classes—neutral, selectively constrained, and positively selected codons. The likelihood ratio test then compares a model where the dN/dS ratio of the codons inferred to positively selected class is set to 1 on a branch of interest and a model where the dN/dS ratio of these codons is allowed to be higher than 1. We performed this test on ‘B3’ but it was not significant ($2\Delta\ln L = 0.06$, $df = 1$). Finally, in order to be able to better interpret different values of the ratio dN/dS, we also calculated dN and dS separately for different pairs of loci (Table 7).

Fig. 3 Phylogenetic tree used in the PAML analysis. In the three-ratio model, all the branches have an estimated dN/dS of 0.29 except where indicated

Table 7 Divergence between genes as measured by dN and dS obtained with PAML

Comparison	dN	dS
<i>AtCO</i> vs <i>AtCOL1</i>	0.1868	0.3528
<i>BniCOa</i> vs <i>BniCOL1</i>	0.2261	0.7157
<i>BniCOB</i> vs <i>BniCOL1</i>	0.2436	0.7027
<i>AtCO</i> vs <i>AtCOL2</i>	0.1902	0.6192
<i>AtCOL1</i> vs <i>AtCOL2</i>	0.1733	0.7500

Discussion

The aim of the present study was to investigate the evolution of three genes, *COa*, *COB*, and *COL1*, in *Brassica nigra*, one of which, *COL1*, is a strong candidate for

control of flowering time in this species (Kruskopf-Österberg et al. 2002; Krouchi et al., in press), and another, *Cob*, a putative pseudogene (Sjödin et al. 2007). We found evidence of selection at *COLI* and patterns and levels of polymorphisms indeed differed among the three genes. Level of selective constraint on *BniCOB* was also consistent with *BniCOB* being a pseudogene. These results are discussed in turn below.

Evidence of Selection on *COLI*

Analysis of molecular variation at the three duplicated loci delivered a complex and contrasted picture of their evolution. Coalescent-scale-based tests such as Tajima's test showed no clear evidence of recent selection on any of the three genes. However, the Kreitman-Hudson test indicated that the difference in levels of nucleotide variation among the three loci could not be fully explained by a mutation-random genetic drift model. Also, in contrast to the coalescent-scale tests, the McDonald-Kreitman test, which uses both polymorphism data and divergence data, detected evidence of selection on *BniCOLI*. In order to interpret the latter we analyzed the ratio of nonsynonymous and synonymous substitution rates (dN/dS) and found a significantly stronger selective constraint on *BniCOLI* than on the other genes, as well as a relaxed selective constraint on *AtCOLI*.

Let us first note that the significance of the McDonald-Kreitman for *BniCOLI* is in contrast to the results of Lagercrantz et al. (2002). The difference between the two studies is due to a sharp increase in the number of synonymous polymorphisms, other categories being almost equal. The total number of segregating sites was 60 in the present study and 47 in Lagercrantz et al. (2002). This could be due to the difference in sampling between the two studies, with the inclusion in the latter of two populations known *a priori* to differ from others, especially the Turkish population that flowered extremely late. The increase in synonymous polymorphisms relative to nonsynonymous polymorphisms could then simply be a consequence of enhanced population subdivision. However, the K_{ST} values for the two types of sites were almost equal (K_{ST} 0.45 for nonsynonymous, K_{ST} 0.47 for synonymous).

So if the significance of the McDonald-Kreitman test for *BniCOLI* is due to a deficit in polymorphic nonsynonymous changes relative to polymorphic synonymous changes, when did selection take place? The analysis of the dN/dS ratio indicates that the significant MacDonal-Kreitman reflects selection occurring after the divergence among *Brassica* species. This increase in constraint is consistent with the involvement of *COLI* in the control of flowering time in *Brassica nigra* (Kruskopf-Österberg 2002; Krouchi et al., in press), a trait likely associated with

fitness. In this respect, the relatively higher values of the coalescent-scale tests of *COLI*—compared to *COa* and *Cob*—are noteworthy, as this would be expected if selection acted to maintain variation in the former but not in the latter genes. Similarly, the apparent relaxed selective constraint observed in *AtCOLI* and the lower constraint in *B. napus* fit with available data. As already stated, altered expression of *AtCOLI* in transgenic plants did not affect flowering time (Ledger et al. 2001). In other *Brassica* species such as *B. oleracea*, *B. rapa*, and *B. napus*, QTL studies have suggested genes other than *COLI* to be the most important for flowering time variation in these species (Schranz et al. 2002; Kim et al. 2006; Okazaki et al. 2007; Lou et al. 2007). Of course, this does not rule out the involvement of *AtCOLI* in other, yet undetected, functions, and the possibility that the lack of evidence on the involvement of *COLI* in the control of flowering time is simply reflecting a lack of power of the QTL studies carried out in other *Brassica* species.

Additional evidence that *BniCOLI* may be under the influence of changing selection is the lack of variation in *BniCOLI* and *BniCOa* in Ethiopia. This is suggestive of either of two things: a bottleneck or a selective sweep of the chromosomal region harboring *BniCOa* and *BniCOLI*. There is an overall decrease in variation in Ethiopia both in *BniCOB* and at microsatellite loci, but minor compared to the reduction in *BniCOa* and *BniCOLI* (Alström-Rapaport et al., in preparation). Consequently, the Ethiopian population has probably gone through a bottleneck. However, a recent study with genetic markers in the intergenic region between *BniCOLI* and *BniCOa* as well as markers outside these showed no variation in the intergenic region in Ethiopia, while some variation was present in a marker upstream of *BniCOLI* (Krouchi et al., in press). This almost-complete depletion of variation in the *BniCOa-BniCOLI* region compared to other regions suggests that, beside a bottleneck, a selective sweep might also have affected this region.

In short, an attractive explanation to our results is that the function or importance of *COLI* in the control of flowering time has changed in the lineages leading to *A. thaliana* and *B. nigra*. The *CONSTANS-LIKE* gene family evolves exceptionally rapidly (Lagercrantz and Axelsson 2000). In light of our results, this fast evolutionary rate could be due to frequent changes of, if not function, at least relative importance, of genes within the family. Such functional turnover has been suggested by some authors (Zuckermandl 2001; Hughes 2005) to be common and may be especially relevant when studying species that have gone through one or several whole-genome duplications, as is the case for *Brassica* and *Arabidopsis*.

Indel Variation Indicates That *COB* Is a Pseudogene

In contrast to the McDonald-Kreitman test, for which the only significant value was obtained for *BniCOLI*, none of the tests comparing the number of indels to the numbers of nonsynonymous and synonymous and all segregating sites (in the coding region) were significant. This was a bit surprising, as one would expect indels to behave similarly to nonsynonymous mutations and, thus, the test to be relevant in the *COLI* case. A more conspicuous feature of the indel variation in our data set is the very long polymorphic deletion in *BniCOB*, which had no counterpart (in length) at the interspecific level. Although this could be an artifact if long deletions are very rare, it is difficult to believe that this would have occurred in a functional gene without being strongly selected against, especially as the deletion removed a major part of the highly conserved zinc finger region. A simple explanation is that *BniCOB* is a recent pseudogene, or at least that a large fraction of its alleles are nonfunctional, and that consequently selection against recurrent long deletions in nonfunctional genes is either weak or absent. Other results that also support the fact that *BniCOB* is a pseudogene, or at least on the way to becoming one, are its nucleotide polymorphism, which is twice as high as that of *BniCOLI* and four times as high as that of *BniCOA*, and the presence at high frequencies of several apparently disabling mutations, including stop codons, frameshifts, and the wrong intron splice signal (for more details see Sjödin et al. 2007).

Conclusion

We explain the significant excess of polymorphic synonymous mutations in a McDonald-Kreitman test comparing *COLI* in *B. nigra* and *A. thaliana* as a result of a recent increase in *COLI* in *B. nigra* paralleled by a recent decrease in *A. thaliana*. The *CONSTANS-LIKE* gene family evolves exceptionally fast (Lagercrantz and Axelsson 2000), and in light of our results, this fast evolutionary rate could be due to frequent changes of, if not function, at least relative importance, of genes within the family. Frequent turnover, regarding which genes are involved in a particular pathway, has been suggested by some authors (Zuckerandl 2001; Hughes 2005) to be common. If this is true, the genes in the *CONSTANS-LIKE* family make exceptionally good candidates to study this process, as their duplicated nature is likely to create a network of redundancy and pleiotropy.

To summarize, the data lead us to believe that *BniCOB* is a recent pseudogene, while there is a recent increase in selective constraint in *BniCOLI*. Whether or not these two events are correlated is not evident from our data. Studying

the homologous genes in other Brassiceae species (that are affected by the same triplication as *B. nigra* [Lysak et al. 2005]) may be revealing in this respect.

Acknowledgments We thank Martin Kreitman and an anonymous referee for extremely useful comments on an early draft. Support was provided by the Swedish Research Council (VR), the Swedish Research Council for Environment, Agricultural Sciences, and Spatial Planning (FORMAS), the Erik Philip Sörensen Foundation, and the Nilsson-Ehle Foundation.

References

- Arnheim N (1983) Concerted evolution of multigene families. In: Nei M, Koehn RK (eds) Evolution of genes and proteins. Sinauer, Sunderland, MA, pp 38–61
- Averof M (2002) Arthropod Hox genes: insights on the evolutionary forces that shape gene functions. *Curr Opin Genet Dev* 12:386–392
- Britten RI, Rowen L, Williams J, Cameron RA (2003) Majority of divergence between closely related DNA samples is due to indels. *Proc Natl Acad Sci USA* 100:4661–4665
- Eyre-Walker A (2006) The genomic rate of adaptive evolution *Trends Ecol Evol* 21:569–575
- Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. *Genetics* 155:1405–1413
- Fisher RA (1935) The sheltering of lethals. *Am Nat* 69:446–455
- Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151:1531–1545
- Gibson TJ, Spring J (1998) Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins. *Trends Genet* 14:46–49
- Haldane JBS (1933) The part played by recurrent mutation in evolution. *Am Nat* 67:5–19
- Hudson RR (1987) Estimating the recombination parameter of a finite population model without selection. *Genet Res Camb* 50:245–250
- Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18:337–338
- Hudson RR, Boos D, Kaplan NL (1992) A statistical test for detecting geographic subdivision. *Mol Biol Evol* 9:138–151
- Hughes AL (2005) Gene duplication and the origin of novel proteins. *Proc Natl Acad Sci USA* 102:8791–8792
- Innan H (2003) The coalescent and infinite-site model of a small multigene family. *Genetics* 163:803–810
- Kim JS, Chung TY, King GJ, Jin M, Yang TJ, Jin YM, Kim HI, Park BS (2006) A sequence-tagged linkage map of *Brassica rapa*. *Genetics* 174:29–39
- Krouchi F, Gustavsson S, Sjödin P, Kruskopf-Österberg M, Lagercrantz U, Lascoux M (2008) Association between *COL1* and flowering time in *Brassica nigra*: replication, validation and genotypic disequilibrium. *Int J Plant Sci* (in press)
- Kruskopf-Österberg M, Shavorskaya O, Lascoux M, Lagercrantz U (2002) Naturally occurring indel variation in the *B. nigra COL1* gene is associated with variation in flowering time. *Genetics* 161:299–306
- Kumar S, Tamura K, Nei M (2004) MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief Bioinform* 5:150–163
- 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, Putterill J, Coupland G, Lydiat D (1996) Comparative mapping in *Arabidopsis* and *Brassica*, congruence of genes controlling flowering time. *Plant J* 9:13–20
- 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
- Ledger S, Strayer C, Ashton F, Kay SA, Putterill J (2001) Analysis of the function of two circadian-regulated *CONSTANS-LIKE* genes. *Plant J* 26:15–22
- Lou P, Zhao J, Kim JS, Shen S, Del Carpio DP, Song X, Jin M, Vreugdenhil D, Wang X, Koornneef M, Bonnema G (2007) Quantitative trait loci for flowering time and morphological traits in multiple populations of *Brassica rapa*. *J Exp Bot* 58:4005–4016
- Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290:1151–1155
- Lysak M, Koch M, Pecinka A, Schubert I (2005) Chromosome triplication found across the tribe *Brassicaceae*. *Genome Res* 15:516–525
- McDonald JH, Kreitman M (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* 351:652–654
- Nielsen R (2005) Molecular signatures of natural selection. *Annu Rev Genet* 39:197–218
- Ohno S (1970) *Evolution by gene duplication*. Springer Verlag, Berlin
- Okazaki K, Sakamoto K, Kikuchi R, Saito A, Togashi E, Kuginuki Y, Matsumoto S, Hirai M (2007) Mapping and characterization of *FLC* homologs and QTL analysis of flowering time in *Brassica oleracea*. *Theor Appl Genet* 114:595–608
- Parkin IAP, Gulden SM, Sharpe AG, Lukens L, Trick M, Osborn TC, Lydiat DJ (2005) Segmental structure of the *Brassica napus* genome based on comparative analysis with *Arabidopsis thaliana*. *Genetics* 171:765–781
- 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
- Putterill J, Ledger S, Lee K, Robson F, Murphy G, Coupland G (1997) The flowering time gene *CONSTANS* and homologue *CONSTANS LIKE 1* (Accession no. Y10555 and Y10556) exist as a tandem repeat on chromosome 5 of *Arabidopsis*. *Plant Physiol* 114:396
- Rozas J, Sánchez-DelBarra JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 29:2496–2497
- Samach A, Onouchi H, Gold SE, Ditta GS, Schwarz-Sommer Z, Yanofsky M, Coupland G (2000) Distinct roles of *CONSTANS* target genes in reproductive development of *Arabidopsis*. *Science* 288:1613–1616
- Schranz ME, Mitchell-Olds T (2006) Independent ancient polyploidy events in the sister families *Brassicaceae* and *Cleomaceae*. *Plant Cell* 18:1152–1165
- Schranz ME, Quijada P, Sung SB, Lukens L, Amasino R, Osborn TC (2002) Characterization and effects of the replicated flowering time gene *FLC* in *Brassica rapa*. *Genetics* 162:1457–1468
- Shavorskaya O, Lagercrantz U (2006) Sequence divergence at the putative flowering time locus *COL1* in *Brassicaceae*. *Mol Phylogenet Evol* 39:846–854
- Sjödin P, Shavorskaya O, Hedman H, Finet C, Lascoux M, Lagercrantz U (2006) Recent degeneration of an old duplicated flowering time gene in *B. nigra*. *Heredity* 98:375–384
- Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105:437–460
- Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595
- Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* 15:4876–4882
- Thornton K, Long M (2005) Excess of amino acid substitutions relative to polymorphism between X-linked duplications in *Drosophila melanogaster*. *Mol Biol Evol* 22:273–284
- Van Hoof A (2005) Conserved functions of yeast genes support the Duplication, Degeneration and Complementation model of gene duplication. *Genetics* 171:1455–1461
- Watterson GA (1975) On the number of segregating sites in genetical models without recombination. *Theor Pop Biol* 7:256–276
- Westman AL, Kresovich S (1999) Simple sequence repeat (SSR)-based marker variation in *Brassica nigra* genebank accessions and weed populations. *Euphytica* 109:85–92
- Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl BioSci* 13:555–556
- Yang Z, Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol* 17:32–43
- Yang Z, Wong WSW, Nielsen R (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol* 22:1107–1118
- Zhang Z (2003) Evolution by gene duplication: an update. *Trends Ecol Evol* 18:292–298
- Zhang Z, Kishino H (2004) Genomic background predicts the fate of duplicated genes: evidence from the yeast genome. *Genetics* 166:1995–1999
- Zuckerandl E (2001) Intrinsically driven changes in gene interaction complexity. I. Growth of regulatory complexes and increase in number of genes. *J Mol Evol* 53:539–554