Arabidopsis and Brassica Comparative Genomics: Sequence, Structure and Gene Content in the *ABI1-Rps2-Ck1* Chromosomal Segment and Related Regions

C. F. Quiros,* F. Grellet,† J. Sadowski,‡ T. Suzuki,§ G. Li* and T. Wroblewski*

**Department of Vegetable Crops, University of California, Davis, California 95616,* †*UMR CNRS 5545, Université de Perpignan, 66860 Perpignan Cedex, France,* ‡*Institute of Plant Genetics, Polish Academy of Sciences, 60-479 Poznan, Poland and* §*Institute of Agriculture and Forestry, University of Tsukuba, Tsukuba, Ibaraki 305, Japan*

Manuscript received May 8, 2000
Accepted for publication July 1, 2000

ABSTRACT

The region corresponding to the *ABI1-Rps2-Ck1* segment on chromosome 4 of *Arabidopsis thaliana* was sequenced in *Brassica oleracea*. Similar to *A. thaliana*, the *B. oleracea* homolog *BoRps2* is present in single copy. The *B. oleracea* orthologous segment was located on chromosome 4 and can be distinguished by the presence of an *N*-myristoyl transferase coding gene (*N-myr*) between the *Rps2* and *Ck1* (*BoCk1a*) genes. The *N-myr* homologs in Arabidopsis are on chromosomes 2 and 5. Additional homologs for *Ck1* are located on these two chromosomes. A second *Ck1* homolog found on *B. oleracea* (*BoCk1b*) chromosome 7 served to define another orthologous segment located in Arabidopsis chromosome 1. The two segments displayed identical gene content and order in both species, namely *BoCK1b*, a gene encoding a hypothetical protein (*BohypothA*) and transcription factor *eiF4A*. High levels of sequence identity were observed for the coding sequences of all genes examined. Although in general larger spacers were found in Brassica than in *A. thaliana*, this was not always the case. Promoters were poorly conserved, except for several sequence stretches of a few nucleotides. Comparative sequencing revealed microsyntenic changes resulting from chromosomal structural rearrangements, which are often undetectable by genetic mapping.

SINGLE mutations, transpositions, and chromosomal rearrangements, together with environmental selection pressure, over time have created the architecture of genomes and their functional integrity. Many of these changes can be readily detected by comparative microsynteny studies, which are based on actual DNA sequence comparisons. Because coding gene sequences are conserved not only among species or genera, but also across family boundaries, it has been suggested that comparative mapping and gene-specific sequencing will be useful for studying chromosome structure and gene organization among related species. Furthermore, such information might help in determining the ancestral relatives of cultivated species as well as their evolution. However, until recently most comparative studies were coarse and based on genetic maps and various kinds of markers.

Arabidopsis thaliana serves as a model for comparative microsynteny studies with Brassica species, considering that species containing genomes approaching the Brassica ancestral genome are unknown and probably extinct. Other great advantages of *A. thaliana* are its small genome size and the extensive sequence information readily available from this species. It must be kept in mind, however, that the crucifers *A. thaliana* and brassicas are classified taxonomically in different tribes (Arabidae and Brassiceae, respectively; Price *et al.* 1994), making a direct ancestral relationship between these two genera unlikely, as suggested by Gale and Devos (1998) and Lagercrantz (1998). Furthermore, on the basis of mtDNA data the lineages leading to these tribes are estimated to have split 14.5 to 20.4 million years ago (Yang *et al.* 1999). Earlier comparative mapping of Arabidopsis and Brassica species disclosed islands of conservation for the chromosome segments tested (Kowalski *et al.* 1994; Sadowski *et al.* 1994, 1996; Lagercrantz and Lydiate 1996; Conner *et al.* 1998). Furthermore, it has been estimated by restriction fragment length polymorphism mapping that only 20% of a chromosome maintains colinearity between *A. thaliana* and *Brassica oleracea*, but 90% of chromosomal regions of <5 cM remain conserved (Paterson *et al.* 1996). Some of these segments conserving gene repertoire and order may represent orthologous segments between these species. Comprehensive comparative microsynteny studies between species of these two genera will allow the assignment and alignment of these segments across species in the near future.

We selected for this study the *A. thaliana ABI1-Rps2-Ck1* segment on chromosome 4, which, according to previous comparative genetic mapping, is structurally conserved in Brassica species (Sadowski and Quiros 1998). *Rps2* is a single-copy locus in both *A. thaliana* (Mindrinos *et al.* 1994) and *B. oleracea* (Wroblewski *et al.* 2000), which facilitates assignment of orthology in

Corresponding author: Carlos F. Quiros, Department of Vegetable Crops, University of California, Davis, CA 95616.
E-mail: cfquiros@ucdavis.edu

both genomes. Genomic sequences corresponding to these two homologous chromosomal segments in *A. thaliana* and *B. oleracea* were analyzed and used to determine the structural changes distinguishing these two crucifers.

MATERIALS AND METHODS

Library construction in *B. oleracea* and screening: We used *B. oleracea* variety Purple Cauliflower (B0265) to construct a cosmid library. DNA was extracted from 6- to 8-week-old plants grown for 2 days in darkness before nuclei isolation (Kianian and Quiros 1992). The isolated DNA was cloned in the SuperCos1 (Stratagene, La Jolla, CA) cosmid vector according to the manufacturer. The library was screened according to Sambrook *et al.* (1989). Arabidopsis cDNA clones for the genes *ABI1*, *Rps2*, and *Ck1* obtained from the Arabidopsis stock center were used as probes, which were labeled with [α-^{32}P]dCTP using the Multiprime labeling system (Amersham Pharmacia Biotech, Piscataway, NJ). After selection of the colonies, cosmids were isolated by alkaline lysis (Sambrook *et al.* 1989).

Physical mapping: Cosmid clones were digested with *Not*I for incomplete vector trimming. A total of 300 ng of cosmid DNA was partially digested with 0.2–1.0 units of *Eco*RI or *Hin*dIII enzyme for 45 min in a volume of 20 μl. Samples were fractionated by electrophoresis in a 0.5% agarose gel at 1.5 V/cm for 40 hr. The gel was alkaline-blotted onto a ZetaProbe membrane (Bio-Rad, Hercules, CA). The membranes were hybridized with 5′-labeled 21-mer T3 and T7 oligonucleotides corresponding to the sequences flanking the inserts in the vector. The hybridization was conducted overnight at 50° according to Bio-Rad protocol with SDS as a blocking agent. Oligonucleotides used as probes were labeled with [γ-^{32}P]dATP by T4 polynucleotide kinase (New England Biolabs, Beverly, MA). Restriction maps for the cosmids were assembled manually.

Sequencing: For sequencing, *Hin*dIII fragments of the digested cosmids were purified from agarose gels and subcloned into plasmid pUC19. After alkaline lysis (Sambrook *et al.* 1989) the clones were purified by PEG/NaCl precipitation. Sequencing was performed on an automated sequencer, ABI PRISM 377. To complete the 5′ end sequences of two caseine kinase homologs (*BoCk1a* and *BoCk1b*), which were truncated in the cosmid clones, we isolated two BAC clones from a library constructed from a doubled-haploid broccoli line derived from the variety Early Big (G. Li and C. F. Quiros, unpublished results). Primers were based on exon 5 of *BoCk1a*, intron 2 of *BoCk1b*, and a few hundred nucleotides upstream of the ATG codons of both genes.

Data analysis: DNA Strider (Marck 1988) was used to visualize open reading frames (ORF), restriction sites, and repeated sequences. Homology searches were carried out against nonredundant public libraries on the National Center for Biotechnology Information server using the BLAST program (Altschul *et al.* 1990). Predicted coding regions were analyzed using the programs Genemark (Borodovsky and McIninch 1993), Genscan (Burge and Karlin 1998), NetPlantGene (Herbsgaard *et al.* 1996), and Genefinder (Zhang 1997). The PLACE database (Higo *et al.* 1998) was used to analyze the putative gene promoter regions.

RESULTS

Clone identification and sequencing for the *ABI1-Rps2-CK1 B. oleracea* segment: We screened a *B. oleracea* cosmid library with an *A. thaliana Rps2* probe and obtained four clones, ranging in insert size from 32.1 to 42.0 kb, which contained homologs to this gene (*AtRps2*). Three of the clones also hybridized with a gene downstream of *Rps2*, Col-0 casein kinase-like protein (*AtCk1*), and the fourth clone hybridized to the protein phosphatase *ABI1* gene (abscisic acid insensitive, *AtABI1*) upstream of *Rps2*. None of the *B. oleracea* cosmid clones carried all three genes. Restriction maps were constructed for each clone by digestion with *Hin*dIII and *Eco*RI. Two of the three Brassica clones containing the *Rps2* (*BoRps2*) and *Ck1* (*BoCk1a*) homologs were selected for sequencing on the basis of their restriction profiles. The insert of the first clone (Figure 1, clone 1) was 39.6 kb in length, whereas in the second one (Figure 1, clone 2) it was 38.3 kb. Comparison of their restriction maps matched profiles expected for the segment carrying *AtRps2* (Mindrinos *et al.* 1994). We sequenced two contiguous *Hin*dIII fragments for both clones 1 and 2 and proved that they were identical. On the other hand, the adjacent sequences in a third *Hin*dIII fragment downstream of *BoRps2*, containing the rest of the spacer and *BoCK1a*, were different in the two clones due to a rearrangement. This difference, however, did not affect the sequences targeted in the analyses of the clones (Figure 1).

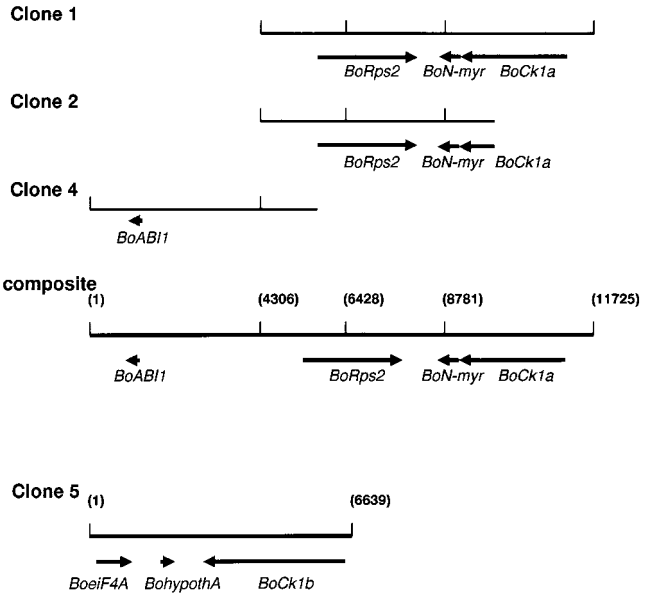

Figure 1.—Schematic representation of the *Hin*dIII segments selected for sequencing in the inserts of four cosmid clones. The composite diagram depicted was constructed by overlapping common sequences from clones 1, 2, and 4. The direction of transcription is indicated by arrows above the name of each gene.

A 4306-bp *Hin*dIII segment of clone 4, hybridizing to both *ABI1* and *Rps2*, was also sequenced (Figure 1, clone 4). After contig assembly of the *B. oleracea ABI1-Rps2-Ck1* segment (AF180355), we found structural differences between the *B. oleracea* and Arabidopsis counterparts.

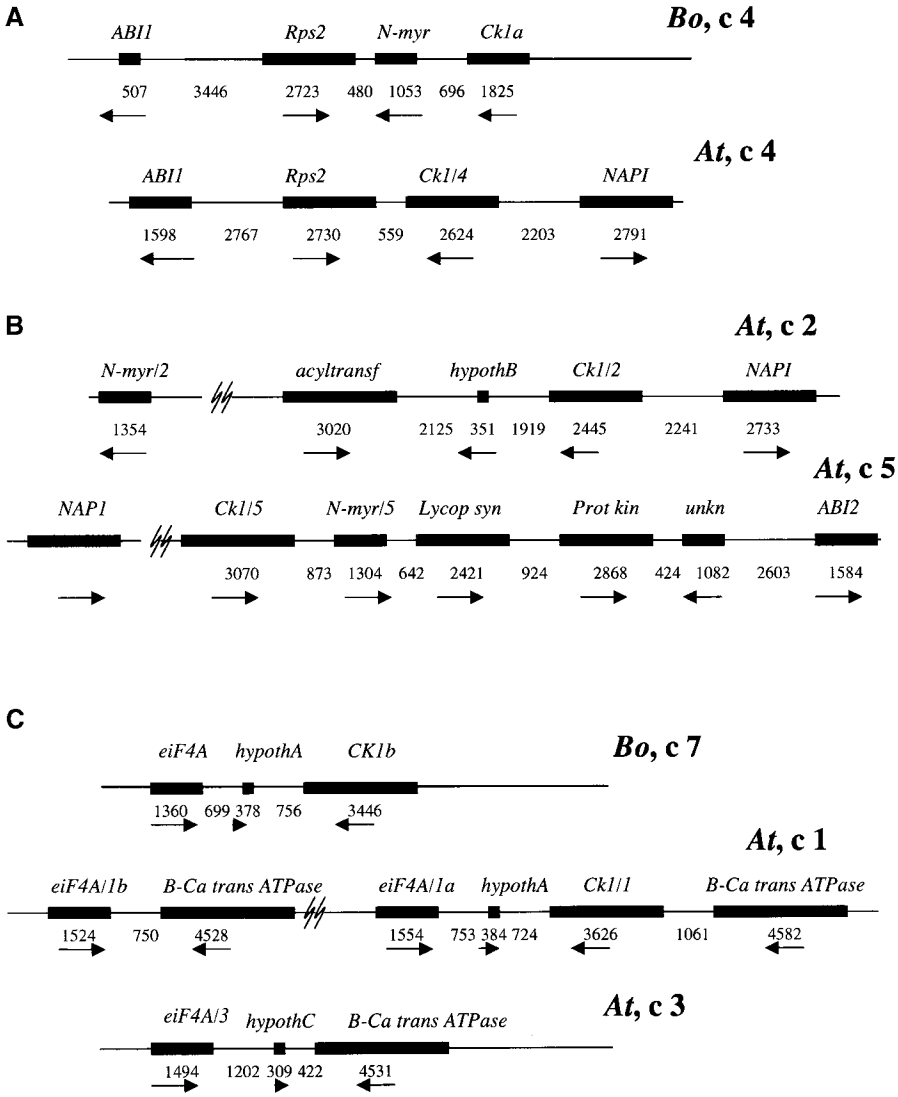

FIGURE 2.—Schematic representation of the homologous chromosome segments of *B. oleracea* and *A. thaliana* relevant to this study. Open reading frames (ORF) are depicted by boxes. Sizes of ORFs and spacers in base pairs are shown under each of these elements. Arrows show the direction of transcription. (A) *B. oleracea* chromosome 4 segment carrying *BoRps2*, *BoN-myr*, and *BoCk1a* and its Arabidopsis ortholog, also on chromosome 4. (B) *A. thaliana* segments on chromosomes 2 and 5 displaying homology to Brassica and Arabidopsis chromosomes 4. (C) *B. oleracea* chromosome 7 segment carrying *BoeiF4A BohypothA* and *BoCk1b* and its Arabidopsis ortholog on chromosome 1. This segment is partially duplicated on the same chromosome and on chromosome 3.

Microsyntenic changes for the *ABI1-Rps2-CK1* segment distinguishing *A. thaliana* from *B. oleracea*: Sequencing of the corresponding *A. thaliana* chromosome 4 segment *ABI1-Rps2-Ck1* in the *B. oleracea* clones 1 and 2 revealed a major change in gene content. An *N-myristoyl transferase* (*N-myr*) gene was found between *BoRps2* and *BoCk1a* (Figure 2A). This segment has been mapped on chromosome 4 of *B. oleracea* (J. SADOWSKI, D. BABULA and M. KACZMAREK, unpublished results). A homology search in GenBank disclosed two *N-myr* homologs in *A. thaliana*, one on chromosome 2 (*AtN-myr/2*, BAC F6E13) and another one on chromosome 5 (*AtN-myr/5*, P1 MHM17). The segment carrying this gene on chromosome 2 was annotated, showing that this chromosome contains homologs for two contiguous genes on chromosome 4, *Ck1* (*AtCk1/2*), and *NAP1* (*AtNAP1/2*). Unlike the *B. oleracea* arrangement, *AtN-myr/2* and *AtCk1/2* are not contiguous, but are a few megabases away from each other (Figure 2B). Although the *A. thaliana* clone P1 MHM17 on chromosome 5 carrying the homolog *AtN-myr/5* is not annotated, we were able to construct a physical map for *AtN-myr/5* and its flanking genes on the basis of similarity to various reported sequences. Similar to the arrangement in *B. oleracea*, in this *A. thaliana* segment a *Ck1* homolog (*AtCk1/5*) was contiguous to *AtN-myr/5*. Additionally, an *AtABI1* homolog (*AtABI2*) was on the same segment, separated from *AtN-myr/5* by three other genes (Figure 2B). In the corresponding *B. oleracea* segment, *BoN-myr* is in between *BoRps2* and *BoCk1a* instead (Figure 2A). In all segments, the orientation of the *Ck1* and *N-myr* was the same. The only exception was the presence of a *NAP1* homolog ~36 kb upstream of *AtCk1/5*. Both are on the same strand, which is contrary to the arrangement observed when these genes are found to be contiguous in other chromosomes in both Arabidopsis and Brassica (Figure 2, A and C).

Structural characteristics and similarity of specific homologs in both species: Overall, the various homologs in the compared *B. oleracea* chromosome 4 and *A. thaliana* chromosome 4 segments were highly similar, except for spacers and introns, as shown in Figure 3A.

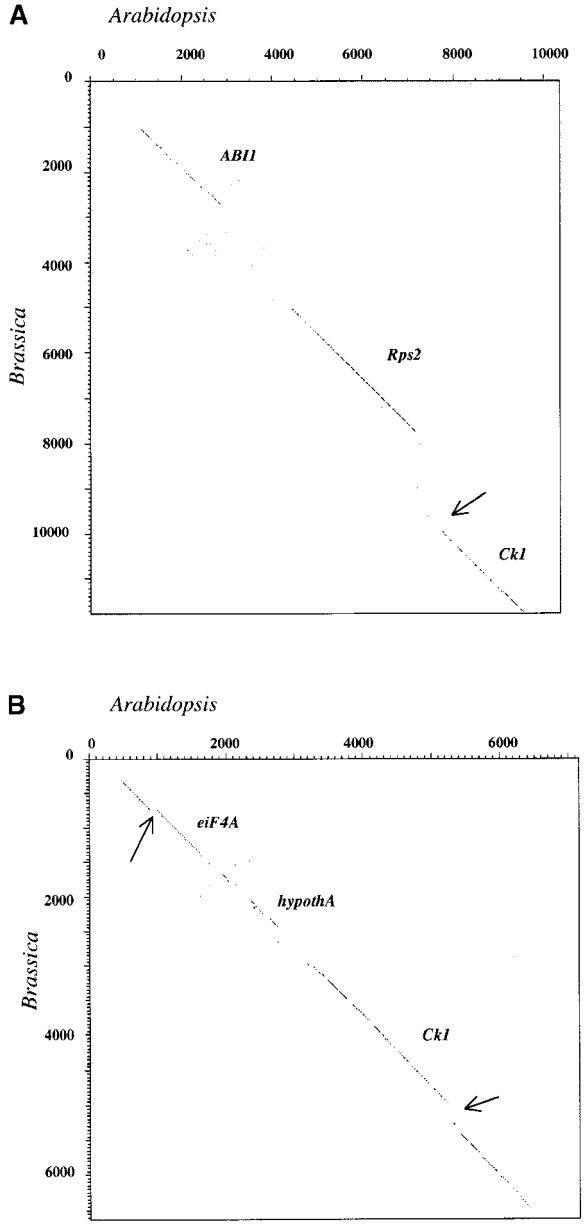

FIGURE 3.—Dot plots for Arabidopsis and Brassica sequences. (A) Brassica sequence AF180355 against Arabidopsis BAC F20B18 aligning orthologous *ABI1-Ck1* segments. The diagonal in the top left corner represents *ABI1* followed by *Rps2* and then by *Ck1*. The first gap represents the spacer between the first two genes. The second gap represents the absence of *N-myr* in the *A. thaliana* sequence and spacers. The trail of dots (arrow) represents the *N-myr* promoter and the *Ck1* untranslated trailer with homology to *Rps2*. (B) Brassica sequence AF1800356 against Arabidopsis BAC F28P22 aligning orthologous *eiF4A-Ck1* segments. The broken line in *eiF4A* (arrow) is due to the lack of intron 3 in the Brassica sequence. The short lines above represent short intergenic spacer sequences showing similarity. The larger breaks on *Ck1* correspond to the largest introns (intron 2, bottom arrow; intron 6, top arrow). The short broken segments correspond to the rest of the introns.

BoRps2: This gene is structurally very similar to its homolog, *AtRps2* (BACF20B18), with respect to orientation, size, and lack of introns. Furthermore, there is a single copy of this gene in both Arabidopsis and in at least two Brassica species, *B. nigra* and *B. oleracea* (SADOWSKI and QUIROS 1998; WROBLEWSKI *et al.* 2000). The amino acid identity of the ORFs of *AtRps2* and *BoRps2* is ~87%. A detailed comparative analysis in the Brassiceae of the different *Rps2* domains, such as the leucine zipper, nucleotide-binding site, and leucine-rich region, has been reported elsewhere (WROBLEWSKI *et al.* 2000). The upstream sequences carrying the promoters in both *AtRps2* and *BoRps2* could be aligned by two stretches of a few bases, showing the rest of the sequences' low similarity values. The only exception was a 59-nucleotide region upstream of the translation initiation signal, which has 78% nucleotide identity between the two homologs. Similarly, the trailer sequence of 33 bp downstream of the stop signal is also conserved (82%). Upstream of the conserved leader, the *B. oleracea* sequence contains putative promoter domains including CAAT and GC boxes and a TATA box (TATAT) at position 112. The promoter sequence was confirmed by sequencing this region in clones 1 and 2. These domains are also present in *AtRps2*, but at different positions. A putative TATA box (TATTAAA) is located in *AtRps2* at nucleotide 91 from the initiation signal.

BoCk1a: Sequencing of the two clones carrying *BoCk1a* disclosed that the 5′ end of this gene was truncated in both clones. We were able to sequence the missing portion of this gene from a BAC clone (as explained above) to complete all 14 exons and promoter (Table 1). *BoCk1a* and *AtCk1/4* have the highest identity of all homologs, in spite of belonging to different species (Table 2). The majority of the *Ck1* homologs have 14 exons, and the first 12 exons from the 5′ end of the gene show size conservation. *AtCk1/5* is the exception due to the fusion of exons 11 and 12 with the loss of a few bases, resulting in an exon of 150 bp. Exons 13 and 14 are variable in size, along with most of the introns in all homologs (Table 1). The promoter and transcribed but nontranslated regions of *AtCk1/4* (determined by the U12857 transcript) and *BoCk1a*, corresponding to 340 bases upstream of the ATG codon of *AtCk1/4*, had 54% shared identity. *AtCk1/4* displayed a $(GA)_8$ microsatellite sequence at 163 bases upstream of its ATG codon. This sequence was also present in *BoCk1a* but it consisted only of four GA copies.

BoN-myr: The sequences of *BoN-myr* and its two homologs, *AtN-myr/2* and *AtN-myr/5*, lack introns interrupting their coding regions. Alignment of the three coding sequences shows some differences between the three corresponding genes. The two Arabidopsis coding sequences are not identical: *AtN-myr/2* is larger than *AtN-myr/5* in its 5′ extremity. In their common part, the Arabidopsis homologs display 74% nucleotide identity. The *BoN-myr* gene differs from its Arabidopsis homologs

TABLE 1

Comparative exon (ex)/intron (in) sizes (from 5′ to 3′ end) in *Ck1* homologs for *B. oleracea* and *A. thaliana*

	BoCk1a	*BoCk1b*	*AtCk1/4*	*AtCk1/1*	*AtCk1/2*	*AtCk1/5*		*BoCk1a*	*BoCk1b*	*AtCk1/4*	*AtCk1/1*	*AtCk1/2*	*AtCk1/5*
Ex1	76	76	76	76	76	76	Ex8	64	64	64	64	64	64
In1	75	66	94	87	71	91	In8	82	74	89	86	79	85
Ex2	41	41	41	41	41	41	Ex9	85	85	85	85	85	85
In2	188	820	148	824	69	297	In9	97	80	88	76	109	81
Ex3	70	70	70	70	70	70	Ex10	127	127	127	127	127	127
In3	96	92	92	93	80	108	In10	73	115	73	108	95	79
Ex4	149	149	149	149	149	149	Ex11	83	80	83	83	83	150
In4	94	93	91	77	73	75	In11	78	81	76	91	95	
Ex5	96	96	96	96	96	96	Ex12	64	64	64	64	64	
In5	97	80	92	80	85	81	In12	76	119	91	111	95	85
Ex6	71	68	71	68	71	71	Ex13	128	137	116	137	89	98
In6	106	612	106	414	113	84	In13	110	175	112	181	86	609
Ex7	62	62	62	62	62	62	Ex14	267	276	249	276	225	219
In7	105	74	119	83	93	87							

by the presence of a 253-bp deletion in the central part of the gene, which results in a shorter predicted protein product (Figure 4). The conserved part of *BoN-myr* has amino acid sequence identities of 82.4 and 77.9% with the Arabidopsis N-myr/5 and N-myr/2 proteins, respectively. For the *AtN-myr/5* gene, several expressed sequence tags (ESTs) are available in the databases. One of them, T76600, allowed us to identify an untranslated 5′ exon of 144 bp followed by a 315-bp intron in the *AtN-myr/5* gene, upstream of the ATG translation initiation signal. The 3′ border of this intron is located 3 bp upstream of the ATG triplet. Such a structure can also be recognized in the *BoN-myr* gene by the following features: First, the 128 bp of the 5′ end of the *AtN-myr/5* exon can be aligned to the sequence upstream of the translation initiation signal in the *BoN-myr* gene with 71.8% identity. Second, the nucleotide sequences around the 5′ and 3′ borders of the *AtN-myr/5* intron are conserved in the corresponding region of the *BoN-myr* gene. These are not found in the *AtN-Myr/2* gene.

BoABI1: Only the 5′ end of this gene was available for comparison between *B. oleracea* and *A. thaliana*. This included the promoter region, the first exon, and part of the first intron. Both *A. thaliana* homologs, *AtABI1* and *AtABI2*, have four exons and three introns each, all of comparable sizes. The identity of the *BoABI1* 5′ portion was higher with *AtABI1* than with *AtABI2*. The first exon in *BoABI1* has 507 bases and is shorter than its Arabidopsis homologs (549 bp for *AtABI1* and 512 bp for *AtABI2*). At the amino acid level, homologies of the first exon of *BoABI1* to those of *AtABI1* and *AtABI2* are 82 and 72%, respectively. The first intron in the *A. thaliana* homologs is small (70 bp for *ABI1* and 82 bp for *ABI2*). In the *B. oleracea* homolog, this intron is much larger, with a partial sequence of 1044 bp. This intron lacks identity to its Arabidopsis counterparts.

Spacer *BoRps2-BoCk1a*: The intergenic spacers for these two genes in *A. thaliana* and *B. oleracea* were structurally different. In Arabidopsis the spacer between these genes (559 bp) is formed by the overlapping ends of the 3′ ends of the transcribed but nontranslated sequences (from stop to stop codons) of these two genes that are in opposite orientations (Figure 5A). There are only 12 nucleotides separating the end of the *Rps2* transcript from the *Ck1* stop signal. On the other hand, the 3′ ends of the *Rps2* and *Ck1* genes in *B. oleracea* are 2212 bases apart, including the *BoN-myr* gene. The spacer sizes for these genes are shown in Figure 2A. The spacer between the 5′ end of the *BoN-myr* gene and the 3′ end of the *BoCk1a* gene has a complex structure consisting of partially overlapping sequences of both the *N-myr* gene and the 3′ end of the *Rps2* gene. These *Rps2* interstitial sequences are homologous to the transcribed, nontranslated 3′ end of both *AtRps2* and cognate cDNA (U12860) and are in the expected orientation for *Rps2* (Figures 3A and 5B).

Spacer *BoABI1-BoRps2*: These two genes are in oppo-

TABLE 2

Identities of *Ck1* homologs in *A. thaliana* and *B. oleracea* at the amino acid level

Homologs	*AtCk1/4*	*AtCk1/1*	*AtCk1/2*	*AtCk1/5*	*BoCk1a*
AtCk1/1	63 (80)				
AtCk1/2	54 (73)	58 (72)			
AtCk1/5	78 (86)	67 (82)	48 (68)		
BoCk1a	90 (96)	67 (83)	53 (68)	82 (90)	
BoCk1b	69 (83)	91 (97)	58 (72)	66 (79)	66 (82)

% identical (% conserved).

site orientation in *B. oleracea* as well as in *A. thaliana*. The spacer between the translation initiation signals of both genes has 3446 bp in *B. oleracea*, whereas in the corresponding *A. thaliana* segment on chromosome 4 it has only 2767 bp (Figure 2A). This spacer includes the promoter region for both genes in both species. In the *BoABI1* putative promoter region, the 666 bases upstream of the ATG initiation signal, which include a $(GA)_{16}$ microsatellite, align well with the 623 bases located upstream of the *AtABI1* ATG triplet, with an overall 63% nucleotide identity. This microsatellite sequence was also present in *AtABI1* and *AtABI2* but in fewer copies. The corresponding region in *AtABI2* is less conserved since only a few segments can be aligned with both *BoABI1* and *AtABI1* putative promoters. At the opposite side of the *BoABI1-BoRps2* spacer, in the *BoRps2* promoter region, a stretch of 207 nucleotides located upstream of the putative TATA domain already mentioned can be aligned with two putative reverse transcriptase sequences contained in the *A. thaliana* accessions AC006300 and AF058914. No similarity with known sequences was found for the middle part of the *BoABI1-Rps2* spacer.

```
MGDENSLAGSLEEKADQVAEATPLVGDDASLETIVRRFQDSMSVEKT
+.+.++ .++.+.++++..++ +++ ++.+++++++++++++  ++
MADNNSPPGSVEQKADQIVEANPLVKDDTSLETIVRRFQDSMSEAKT

HKFWETQPVGQFKDIGDTSLPEGPIEAATPLSEVKQEPYNLPAAYEW
++++++++++++++++++++++++.+++++++++++++..+++
HKFWETQPVGQFKDIGDTSLPEGPIEPATPLSEVKQEPYNLPSVYEW

TTCDMKSDYVCSEVYNLLKNNYVEDDENMFRFNYSKEFLSWALRPPG
+++++.++ .++++++++++++++++++++++++++++ +++++++
TTCDMNSDDMCSEVYNLLKNNYVEDDENMFRFNYSKEFLRWALRPPG

YYQSWHIGVRAKVSK--------------------------------
++++++++++++ ++
YYQSWHIGVRAKTSKKLVAFISGVPARIRVRDEVVKMAEINFLCVHK

------------------------------------------------
KLRSKRLAPVMIKEVTRRVHLENIWQAAYTAGVILPTPITTCQYWHR

-----KLIDVGFSRLGARMTMSRTIKLYKLPDVPATPGFRKMEPCDV
     ++++++++++++++++++++++++++.+ +++++++++ ++
SLNPKKLIDVGFSRLGARMTMSRTIKLYKLPDAPITPGFRKMEPRDV

PAVTRLLRNYLSQFIVATDFDENDVEHWLLPREDVVDSYLVESPETH
++++++++++++++ ++++++++++++++++++++++++++++++
PAVTRLLRNYLSQFGVATDFDENDVEHWLLPREDVVDSYLVESPETH

DLTDFCSFYTLPSTILGNPNYSTLKAAYSYYNVATKTTFSSLMNDAL
+.++++++++++++++++.++++++++++++.+.+  ++++++
DVTDFCSFYTLPSTILGNPNYTTLKAAYSYYNVATQTSFLQLMNDAL

IVAKQKGFDVFNALDVMHNESFLKQLKFGPGDGQLHYYLYNYRLRSA
++.++++++++++++++++++.++++++++++++++++++.++
IVSKQKGFDVFNALDVMHNESFLKELKFGPGDGQLHYYLYNYRLKSA

LKPSELGLVLL
+++.+++++++
LKPAELGLVLL
```

FIGURE 4.—Protein alignment for *BoN-myr* (upper sequence) and *AtN-myr/5* (lower sequence) based on At EST T76600 and P1 clone MHM7. The missing portion in *BoN-myr* is indicated by dashes. In their common parts, both proteins are 90% homologous if only identical amino acids are considered (+). Identity is 96% taking into account conservative amino acids (.).

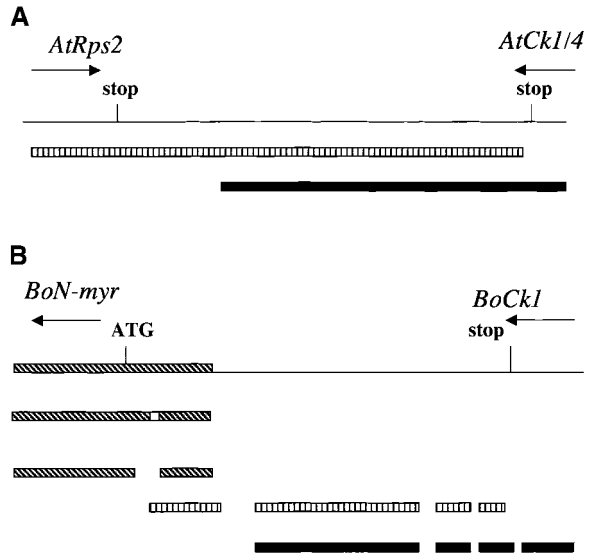

FIGURE 5.—Diagram depicting homologies in the spacers of two orthologous segments. Arrows represent direction of transcription for each gene. (A) *AtRps2-AtCK1/4* spacer from stop to stop codon of both genes. The vertically hatched box represents *Rps2* transcript U12860 and the solid bar represents the *CK1* transcript U12857. The transcripts of both genes overlap at their ends, and the *AtRps2* transcript ends 12 nucleotides before the *AtCK1/4* stop codon. (B) *BoN-myr-BoCK1* spacer. The diagonally hatched boxes depict homologous regions of *AtN-myr* sequences MHM17 (intron lacking homology to *BoN-myr* shown as open box) and EST T76600. The vertically hatched boxes show the region of homology to *AtRps2* transcript U12860, which overlaps with *AtCk1* sequences (F20B18 and U12857), shown as solid bars.

Other related segments: A fifth clone (clone 5, Figure 1) in the Brassica cosmid library hybridized only to the *Ck1* Arabidopsis probe and displayed a unique restriction profile different from those observed for clones 1 and 2. It was partially sequenced (GenBank accession no. AF180356), covering a total of 6639 bp of the *Hin*dIII fragment. This clone contains a *Ck1* homolog (named *BoCk1b*) that maps on *B. oleracea* chromosome 7 (J. Sadowski, D. Babula and M. Kaczmarek, unpublished results) as well as homologs for a hypothetical protein gene (named *BohypothA*) and transcription factor *eiF4A* gene (named *BoeiF4A*). The Brassica sequence AF180356 was found to have its counterpart in *A. thaliana* BAC F28P22 on chromosome 1 (AC010926). Although this Arabidopsis clone is nonannotated, its high sequence identity with its Brassica counterpart permitted identification of the homologous genes. The overall identity of the homologs in these two segments is shown in Figure 3B.

BoCk1b: Like *BoCk1a*, *BoCk1b* was truncated at the 5′ end, but only its first two exons and the promoter were missing in this case. We completed the two missing exons and promoter by sequencing a BAC clone as explained above. The amino acid identity between the *A. thaliana Ck1* homolog on chromosome 1 (named *AtCk1/1* hereafter) and *BoCk1b* is 91% (97% amino acid conservation), which is the highest among other homologs in both species (Table 2). The exons compared, 1–14, were identical in size in both homologs, including exons 6, 13, and 14, which were variable among other homologs. The overall nucleotide identity for the exons of *AtCk1/1* and *BoCk1b* was ~91%. Intron nucleotide identity for these two genes was ~70%, excluding introns 2 and 6, which are the largest exons and thus showed less similarity (Table 1). *BoCk1b* also displayed high identity to three *A. thaliana* ESTs that probably correspond to a single mRNA from *AtCk1/1* (Z25497, R90041, and T13780).

The region upstream of the ATG codon of *BoCk1b* of 520 bases could be aligned to its corresponding 560-base region of *AtCk1/1* with 48% identity. Several stretches of a few nucleotides shared the same sequence or one very similar, permitting alignment of these two regions corresponding to the promoters of the two genes.

BohypothA: This gene was found next to and in opposite orientation to *BoCk1b*. This is also the case for its *A. thaliana* homolog (hereafter named *AthypothA*) found in BAC F28P22 (AC010926), described as a gene coding for a "hypothetical protein," which is next to *AtCk1/1* and in opposite orientation. The amino acid identity between *AthypothA* and *BohypothA* is 68.5% (85% conservation), whereas the nucleotide identity from the ATG translation initiation signal to the stop signal is 75%. Both genes are small and intronless, differing in size by only 6 bp (Figure 2C). The spacer between the *Ck1* and *hypothA* genes in both species is also similar in size, 756 bases for Brassica and 724 bases for Arabidopsis. There are two stretches with higher sequence similarity: one of ~230 bases downstream of the *hypothA* translation stop signal (59.6% identity) and the other ~200 bases downstream of the *Ck1* translation stop signal (78% identity). The rest of the spacer showed little similarity in the two species.

In Arabidopsis we found two other chromosomes carrying homologs for the gene coding for the hypothetical protein *AthypothA*. The first one is on chromosome 2 (hereafter named *AthypothB*, AC005917) and the second one is on chromosome 3 (hereafter named *AthypothC*, AB019229). The ORFs of these homologs are 350 and 309 bp, respectively. Similarly to *AthypothA* and *BohypothA*, none of these homologs contained introns. Interestingly, *AthypothB* was next to a *Ck1* homolog (hereafter named *AtCk1/2*) on chromosome 2, mimicking the arrangement and orientation observed for these two genes on *A. thaliana* chromosome 1 and *B. oleracea* chromosome 7. Two more loci were observed, making a total of five genes coding a hypothetical protein of similar amino acid composition and size: a second locus on chromosome 2 (AC002535) and another one on chromosome 4 (AL049481). The identity correspondence among these proteins of the Arabidopsis hypothetical protein genes was lower in comparison to that observed between *AthypothA* and *BohypothA*.

BoeiF4A: BoeiF4A is next to the *BohypothA* gene; both are in the same orientation. This is also the case for the corresponding *A. thaliana* segment on chromosome 1 (BAC F28P22), which contains the homologs for these two genes (Figure 2C). The sequence of *BoeiF4A* and that of its Arabidopsis counterpart (hereafter named *AteiF4A/1a*) could only partially be compared because *BoeiF4A* was truncated at the 5′ end starting in the second intron, making a total length of 1063 bases available for analysis. *AteiF4A/1a* has four exons; therefore, the promoter and first two exons of the Brassica gene could not be inspected. The two homologs displayed high amino acid identity, as well as the mRNA for *AteiF4A/1a* (GenBank accession no. AJ010472; 96.9% identity, 99.7% conserved). However, the sequence corresponding to the third intron of the Arabidopsis gene was completely absent in the Brassica homolog. The coding sequences corresponding to exons 3 and 4 have 89.5% identity between both homologs. These sequences have the same size in both species, indicating exon size conservation. The spacer between the *eiF4A* and *hypothA* genes was 53 bases longer in Arabidopsis than in Brassica (Figure 2C), displaying an overall nucleotide identity of 57.2%. The promoter of these genes seems to be ~70 bases upstream of the ATG translation initiation signal, where TATA-like sequences could be observed. The sequences corresponding to the translated but nontranscribed sequences of both genes had higher sequence similarity than the rest of the spacer.

In *A. thaliana* there are at least two other homologs to the *eiF4A* gene. There is a second locus on chromo-

some 1 (hereafter named *AteiF4A/1b*; BAC clone AC005287) ~20 kb upstream of the first one (BAC F28P22). An mRNA sequence is available in GenBank for this gene. The third *eiF4A* homolog was located on chromosome 3 (hereafter named *AteiF4A/3*, P1 clone AB019229). There is also an mRNA accession in GenBank (AJ010472), described as a DEAD box RNA helicase, another name given to this gene. The ORFs of these genes are 1524 and 1494 bases, respectively. All three *A. thaliana eiF4A* homologs have four exons and three introns. None of them lack the third intron as does *BoeiF4A*. These are large genes with total regions covered by transcription of 2373 bp for *AteiF4A/3* and 2590 bp for *AteiF4A/1b*. The identity of the amino acid sequences between species and within Arabidopsis was very high, on the order of 95%. The highest identity, however, was observed between *BoeiF4A* and *AteiF4A/1a*. All three Arabidopsis chromosome segments carrying the *eiF4A* homologs also had homologs for a *B-Ca*$^{+}$ *trans ATPse* gene nearby (Figure 2B).

DISCUSSION

Sequencing of the *B. oleracea* segment spanning from *BoABI1* to *BoCk1a* allowed us to disclose changes in gene content with respect to Arabidopsis in this part of the genome.

On the basis of sequence similarity and gene content we were able to detect two sets of orthologous chromosome segments in Arabidopsis and Brassica. The first one on chromosome 4, spanning genes *ABI1* to *Ck1*, and the second one on chromosome 1, spanning genes *eiF4A* to *Ck1*. The orthology of the *ABI1-Ck1* segments in both species is supported by the following facts:

1. Highest sequence similarity between the chromosome 4 segments from both species than to any other Arabidopsis segments carrying homologs for *ABI1* and *Ck1*.
2. Single-copy status of *Rps2* in both species and high sequence similarity of the two homologs (Mindrinos *et al.* 1994; Wroblewski *et al.* 2000). The single-copy status of *Rps2* in *A. thaliana* is further supported by the absence of other ESTs or any other sequences displaying high similarity levels throughout the whole length of the gene.
3. The presence of three additional genes in common to both species downstream of their *Ck1* respective homologs. These genes have been detected by genetic mapping and pulsed-field gel electrophoresis in *B. nigra* (Sadowski and Quiros 1998) and *B. oleracea* (J. Sadowski, D. Babula and M. Kaczmarek, unpublished results). They code for a nucleosome assembly protein (*NAP1*), for a NPR1-like protein, and for an "uncharacterized" protein.

The alternative possibility is that the *ABI2-Ck1* segment on chromosome 5, which carries a *N-myr* gene next to *Ck1*, is orthologous to the *B. oleracea ABI1-Ck1* segment. This is unlikely because of the following facts: (1) the absence of *Rps2* in this segment, which is replaced by three other genes; (2) the absence of a *NAP1* sequence contiguous to *Ck1*, although there is a *NAP1* homolog ~40 kb from *Ck1* and in the same strand (in the other chromosomes these two genes are in opposite strands) on this chromosome; and (3) higher sequence identity for the *ABI1* and *Ck1* homologs in the chromosome 4 segments of both species.

The structural changes distinguishing the orthologous *ABI-Ck1* segments in *A. thaliana* and *B. oleracea* most likely have occurred after the separation of the lineages leading to the formation of the Arabidae and Brassiceae tribes. The question is, which arrangement is ancestral? We can only speculate that perhaps the *A. thaliana* arrangement may be recent. In such case we assume that a copy of an ancestral *N-myr* gene was present in the ancestor of *A. thaliana* chromosome 4 between the *AtRps2* and *AtCK1/4* genes, in the same orientation as we found them today in the *B. oleracea* orthologous segment. This possibility is supported by the virtual fusion of the ends of the *AtRps2* and *AtCk1/4* genes, whose transcripts overlap at their 3′ termini, which might have resulted from the excision of an *N-myr* gene. Conner *et al.* (1998) reported that the lack of the self-incompatibility locus in *A. thaliana* might be due to a similar event. Furthermore, the contiguous arrangement of *Ck1* and *N-myr* on *A. thaliana* chromosome 5 also supports the assumption that this is the ancestral arrangement in the Brassicaceae. An important consideration, however, is that the occurrence of the syntenic change resulting in the presence of *BoN-myr* between *BoRps2* and *BoCk1a* can be tested in other species of this family, which will tell us more about the time of occurrence of this event and its implication on the origin and evolution of these species. The absence of the *N-myr* sequence on the *A. thaliana* chromosome 4 segment carrying *Rps2* and the lack of analysis of all *N-myr* homologs in *B. oleracea* impede orthology assigment to *BoN-myr*. However, higher sequence identity of this gene to *AtN-myr/5* rather than to *AtN-myr/2* suggests a common origin for *BoN-myr* and *AtN-myr/5*. This is further supported by the similar structure of *BoN-myr* and *AtN-myr/5*. All N-myr proteins described to date present a structure comparable to that of the two Arabidopsis genes, *AtN-myr/2* and *AtN-myr/5*. The absence of an interstitial segment in *BoN-myr* does not shift its ORF; however, it puts in question the functionality of this gene. Additional homologs of *N-myr* in *B. oleracea* will have to be characterized to know more about the origin and relationships of these genes.

Orthology assignment for the *eiF4A-Ck1* segments from Arabidopsis chromosome 1 and *B. oleracea* chromosome 4 was straightforward on the basis of the high level of homology of the three pairs of genes compared. Furthermore, gene content is identical and even the spacers are not very different in size in the two segments.

Sequence identity conservation sheds light on the level of divergence among the genes compared. The

Ck1 sequences are the most informative for this type of inference because of their multiple copies in *A. thaliana* and *B. oleracea*. Several conclusions can be reached from this analysis:

1. Greater similarity exists between homologs of different species than within the same species; *e.g.*, *BoCk1a* and *AtCk1/4* and *BoCk1b* and *AtCk1/1* are two pairs of orthologs since they share higher identity than their respective homologs within the same species.
2. Divergence of these two pairs of orthologs from each other has taken place before tribal separation of the two species.
3. Intraspecific homologs share different levels of sequence identity. In Arabidopsis, the most divergent *Ck1* homolog is *AtCk1/2*, which might represent an ancient paralog. On the other hand, the relatively higher sequence identity *AtCk1/5* to *AtCk1/4* indicates that this might represent a paralog of more recent origin.
4. Changes in exon size due to the loss of an intron do not reflect the level of sequence divergence among *Ck1* or *eiF4A* homologs. For example, fusion of exons 11 and 12 due to the lack of intron 11 in *AtCk1/5* was unique and might represent a recent structural change to this homolog, since it was absent in all other *Ck1* homologs. On the other hand, the sizes of the last two exons located at the C-terminal part of the *Ck1* genes (exons 13 and 14) were variable, with the notable exception of *BoCk1b* and *AtCk1/1*, which had the highest level of sequence identity.

A likely explanation for this variation is that the first 11–12 exons of these genes encode for the kinase portion of the protein product, which is generally conserved among all kinases. The situation observed for the 5′ end portion of the *ABI* homologs available for analysis is apparently somewhat different. Although the information is quite limited, for these genes exon size might not be as conserved as it is for the *Ck1* and *eiF4A* homologs, since the first exon of *BoABI* was smaller than that of *AtABI1*.

The presence of the three pairs of homologs, *eiF4A* and *B-$^{+}$Ca trans ATPse*, close to each other on three different chromosome segments of Arabidopsis, is unlikely to occur by chance. These segments probably represent ancient duplications followed by rearrangements in the Arabidopsis genome. Six of the seven genes we compared have two to four copies, with the exception of *Rps2*, which was in single copy. The high level of gene replication observed is in agreement with the recent finding that the level of gene duplication in Arabidopsis is much higher than reported previously (Lin *et al.* 1999; Blanc *et al.* 2000; Grant *et al.* 2000). This demonstrates that the Arabidopsis genome is not a simple and primitive genome. Although sequencing information is currently quite limited in Brassica, most likely a similar or higher level of complexity will be disclosed in these species. In *B. oleracea* we were able to compare only two *Ck1* genes, *BoCk1a* and *BoCk1b*. In any case, it is clear that homologs for specific genes in both species can be readily distinguished by their sequences as well as their exonic/intronic structure. Now that the complete genome of Arabidopsis is sequenced (Arabidopsis Genome Initiative 2000), a full reassessment of these duplications in this species as well as in Brassica species will be possible.

Spacer size difference between Arabidopsis and Brassica has been an issue that various laboratories have tried to address, using mostly genetic mapping data (Lagercrantz and Lydiate 1996; Sadowski *et al.* 1996; Conner *et al.* 1998; Sadowski and Quiros 1998). In general, there is the impression that the difference in genome size between Arabidopsis and other cruciferae species was due in part to a smaller spacer size in the former (Sadowski *et al.* 1996). However, comparative sequence analysis shows that the situation is much more complex than previously thought.

Inspection of spacer sequences for the Brassica genes studied failed to disclose an extensive retrotransposon sequence as reported in maize (SanMiguel *et al.* 1997). Thus, for the chromosome segments analyzed, it was not possible to explain rearrangements in synteny solely by the action of these elements in Brassica. The only evidence for the presence of retrotransposons was in the spacer between *ABI1* and *Rps2* in Brassica. This is not surprising, considering that these elements are located in high density in the pericentromeric regions of Arabidopsis (Copenhaver *et al.* 1999). The spacers examined contained mostly the putative promoters and untranslated trailers of their respective genes. In general they disclosed poor sequence conservation among homologs, except for a few short nucleotide stretches.

Comparative sequencing allows detection of synteny breaks caused by chromosomal rearrangements that distinguish the genomes of Arabidopsis and Brassica. Genetic mapping based on DNA hybridization often fails to detect these changes since it affords only a rough approximation and must be followed by sequence analysis to gather precise information on the structure of complex genomes such as those of the crucifers. The latter approach will tell us more about the evolutionary paths followed by these species in the near future.

We are indebted to Vincent D'Antonio, Dinh Li, and Russell Wrobel for technical assistance and to Genyi Li, Sheila McCormick, and Roger Chetelat for reviewing the manuscript, and to Karen Olson for editing it. We are also indebted to Dean Lavelle at the UCD Plant Genetics Facility for sequencing our DNA samples. This work was supported in part by United States Department of Agriculture National Research Initiative grant no. 9600835 to C.F.Q. and by the Polish Committee for Scientific Research grant no. PO6A016 11 to J.S.

LITERATURE CITED

Altschul, S. F., W. Gish, W. Miller, E. Myers and D. Lipman, 1990 Basic local alignment search tool. J. Mol. Biol. **215:** 403–410.
Arabidopsis Genome Initiative, 2000 Analysis of the genome se-

quence of the flowering plant Arabidopsis thaliana. Nature **408:** 796–815.

Blanc, G., A. Barakat, R. Guyot, R. Cooke and M. Delseny, 2000 Extensive duplication and reshuffling in the Arabidopsis genome. Plant Cell **12:** 1093–1102.

Borodovsky, M., and J. McIninch, 1993 Genemark: parallel gene recognition for both DNA strands. Comput. Chem. **17:** 123–133.

Burge, C. B., and S. Karlin, 1998 Finding the genes in genomic DNA. Curr. Opin. Struct. Biol. **8:** 346–354.

Conner, J. A., P. Conner, P. E. Nasrallah and J. B. Nasrallah, 1998 Comparative mapping of the Brassica S locus region and its homeolog in Arabidopsis: implications for the evolution of mating systems in the Brassicaceae. Plant Cell **10:** 801–812.

Copenhaver, G. P., K. Nickel, T. Kuromori, M. I. Benito, S. Kaul *et al.*, 1999 Genetic definition and sequence analysis of *Arabidopsis* centromeres. Science **286:** 2468–2474.

Gale, M. D., and K. M. Devos, 1998 Plant comparative genetics after 10 years. Science **282:** 656–658.

Grant, D., P. Cregan and R. C. Shoemaker, 2000 Genome organization in dicots: genome duplication in Arabidopsis and synteny between soybean and Arabidopsis. Proc. Natl. Acad. Sci. USA **97:** 4168–4173.

Herbsgaard, S. M., P. G. Korning, N. Tolstrup, J. Engelbrecht, P. Rouze *et al.*, 1996 Splice site prediction in *Arabidopsis thaliana* pre-mRNA by combining local and global sequence information. Nucleic Acids Res. **24:** 3439–3452.

Higo, K., Y. Ugawa, M. Iwamoto and H. Higo, 1998 PLACE: a database of plant *cis*-acting regulatory DNA elements. Nucleic Acids Res. **26:** 358–359.

Kianian, S. F., and C. F. Quiros, 1992 Generation of a *Brassica oleracea* composite RFLP map: linkage arrangements among various populations and evolutionary implications. Theor. Appl. Genet. **84:** 544–554.

Kowalski, S. P., T-H. Lan, K. A. Feldmann and A. H. Paterson, 1994 Comparative mapping of *Arabidopsis thaliana* and *Brassica oleracea* chromosomes reveals islands of conserved organization. Genetics **138:** 499–510.

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., and D. Lydiate, 1996 Comparative genome mapping in *Brassica*. Genetics **144:** 1903–1910.

Lin, X., S. Kaul, S. Rounsley, T. Shea, M. I. Benito *et al.*, 1999 Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. Nature **402:** 761–767.

Marck, C., 1988 "DNA Strider": a "C" program for the fast analysis of DNA and protein sequences on the Apple Macintosh family of computers. Nucleic Acids Res. **16:** 1829–1836.

Mindrinos, M., F. Katagiri, G. Yu and F. M. Ausubel, 1994 The *A. thaliana* disease resistance gene *RPS2* encodes a protein containing a nucleotide-binding site and leucine-rich repeats. Cell **78:** 1089–1099.

Paterson, A., T-H. Lan, K. Reischmann, C. Chang, Y. R. Lin *et al.*, 1996 Toward a unified genetic map of higher plants, transcending the monocot-dicot divergence. Nat. Genet. **14:** 380–382.

Price, R. A., J. D. Palmer and I. A. Al-Shehbaz, 1994 Systematic relationships of *Arabidopsis*: a molecular and morphological perspective, pp. 7–19 in *Arabidopsis*, edited by E. M. Meyerowitz and C. M. Somerville. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Sadowski, J., and C. F. Quiros, 1998 Organization of an *Arabidopsis thaliana* gene cluster on chromosome 4 including the *RPS2* gene, in the *Brassica nigra* genome. Theor. Appl. Genet. **41:** 226–235.

Sadowski, J., J. Hu, M. Delseny and C. F. Quiros, 1994 Genetical and physical mapping of an *Arabidopsis* gene complex in *Brassica* genomes. Cruciferae Newslett. **16:** 47–48.

Sadowski, J., P. Gaubier, M. Delseny and C. F. Quiros, 1996 Genetic and physical mapping in *Brassica* diploid species of a gene cluster defined in *Arabidopsis thaliana*. Mol. Gen. Genet. **251:** 298–306.

Sambrook, J., E. F. Fritsch and T. Maniatis, 1989 *Molecular Cloning: A Laboratory Manual*, Ed. 2. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

SanMiguel, P., A. Tikhonov, Y.-K. Jin, N. Motchoulskaia, D. Sakharov *et al.*, 1996 Nested retrotransposons in the intergenic regions of the maize genome. Science **274:** 765–768.

Wroblewski, T., S. Coulibaly, J. Sadowski and C. F. Quiros, 2000 Variation and phylogenetic utility of the Arabidopsis thaliana Rps2 homolog in various species of the tribe Brassiceae. Mol. Phylogenet. Evol. **16:** 440–448.

Yang, Y.-W., K. N. Lai, P. Y. Tai and W.-H. Li, 1999 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.

Zhang, M. Q., 1997 Identification of protein coding regions in the human genome based on quadratic discriminant analysis. Proc. Natl. Acad. Sci. USA **94:** 565–568.

Communicating editor: B. S. Gill